Ingenieurfakultät Bau Geo Umwelt
Methodik der Fernerkundung

Automated and Precise 3D Building Reconstruction using UAVs

Tobias Koch

Vollständiger Abdruck der von der Ingenieurfakultät Bau Geo Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktor-Ingenieurs (Dr.-Ing.)

genehmigten Dissertation.

Vorsitzender:	Prof. Dr.-Ing. habil. Xiaoxiang Zhu
Prüfer der Dissertation:	1. Prof. Dr.-Ing. habil. Richard Bamler
	2. Ass. Prof. Dr. techn. Friedrich Fraundorfer
	3. Prof. Dr.-Ing. Markus Gerke

Die Dissertation wurde am 30.10.2019 bei der Technischen Universität München eingereicht und durch die Ingenieurfakultät Bau Geo Umwelt am 27.02.2020 angenommen.

AUTOMATED AND PRECISE 3D BUILDING RECONSTRUCTION USING UAVS

TOBIAS KOCH

Dissertation

Chair of Remote Sensing Technology
Department of Civil, Geo and Environmental Engineering
Technical University of Munich

January 11, 2021

Tobias Koch: *Automated and Precise 3D Building Reconstruction using UAVs* , Dissertation , © January 11, 2021

ABSTRACT

In light of the tremendous advances in the fields of unmanned aerial vehicles (UAVs) and imaging sensors in recent years, UAV-photogrammetry has become an essential part of remote sensing methodology. Being more than an alternative to conventional image acquisition platforms, UAV-photogrammetry has revealed novel possibilities and explored a variety of application fields, including the generation of high-quality 3D building models which are of growing importance in the area of 3D city modeling and civil engineering. Nevertheless, practical utilization of UAVs for the task of 3D modeling is still accompanied by various cumbersome activities, such as manual flight planning, deployment of ground control points (GCPs) and manual registration of disconnected 3D models. To this end, this thesis aims to address key challenges in the process of using UAVs for photogrammetric applications and proposes several methods for advancing the state-of-the-art in different stages of UAV-based photogrammetry. Focusing on 3D modeling of buildings, this thesis contributes methods for an automation of the reconstruction process ranging from (i) an accurate image-based multi-modal geo-referencing of acquired images, (ii) an automatic and semantic-aware 3D UAV image acquisition flight planning, (iii) an automatic alignment between individual 3D reconstructions of interior and exterior building models and (iv) a comprehensive investigation of current deep learning-based methods for the task of single-image depth estimation (SIDE), which could contribute to certain areas of image-based 3D building reconstruction. Based on the results of several real-world experiments, the proposed image matching method achieves pixel-level registration accuracies between UAV and multi-modal remote sensing imagery despite significant geometric, radiometric and temporal differences. The model-based 3D path planning method allows for acquiring close-range multi-view stereo-capable image sequences in tightly built-up environments that cover the entire building in a demanded resolution. By incorporating semantic cues into the path generation process, the resulting trajectories are by far more desirable in terms of flight safety by respecting pre-defined restricted and hazardous airspaces such as adjacent buildings or roads. The alignment of individual image-based indoor and outdoor building models is addressed by matching insufficiently overlapping geometric structures, which are shared in both models using 3D line segments as geometric features. A wide variety of experiments on different buildings have verified an accurate registration in centimeter-level accuracy. A comprehensive assessment of current SIDE methods with novel evaluation metrics on a high-quality RGB-D dataset reveals their current suitability for potential practical application fields and emphasizes remaining challenges in this research field. Backed by thorough experimental evaluations confirming the validity of the proposed methods, this thesis marks a step towards an automated, fast, accurate and safe use of UAV photogrammetry.

ZUSAMMENFASSUNG

Angesichts der enormen Fortschritte auf den Gebieten der unbemannten Luftfahrzeuge (UAVs) und bildgebenden Sensoren ist die UAV-Photogrammetrie zu einem wesentlichen Bestandteil der Fernerkundungsmethodik geworden. Dabei ist sie weit mehr als nur eine Alternative zu herkömmlichen Bildaufnahmeplattformen, sondern stellt vor allem ihr Potential für neue Anwendungsmöglichkeiten unter Beweis, darunter die Erstellung hochwertiger 3D-Gebäudemodelle, die im Bereich der 3D-Stadtmodellierung und des Bauwesens von zunehmender Bedeutung sind. Dennoch wird die praktische Nutzung von UAVs für die Aufgabe der 3D-Modellierung weiterhin von verschiedenen aufwändigen Aktivitäten begleitet, wie z.B. einer manuellen Flugplanung, dem Einsatz von Bodenpasspunkten und einer manuellen Registrierung von nicht verbundenen 3D-Modellen. Zu diesem Zweck widmet sich diese Arbeit den wichtigsten Herausforderungen der UAV-Photogrammetrie und präsentiert Methoden zur Weiterentwicklung des Stands der Technik in verschiedenen Teilbereichen. Mit dem Schwerpunkt auf 3D Gebäudemodellierung trägt diese Arbeit mehrere Verfahren für eine Automatisierung des Rekonstruktionsprozesses bei, die von (i) einer genauen bildbasierten Georeferenzierung, (ii) einer semantisch-bewussten 3D UAV-Flugplanung, (iii) einer Registierung von 3D Innen- und Außenmodellen und (iv) einer umfassenden Untersuchung aktueller Methoden des tiefen Lernens für die Aufgabe der Einzelbildtiefenschätzung reichen. Basierend auf den Ergebnissen mehrerer Experimente erreicht das vorgeschlagene Bildregistierungsverfahren trotz signifikanter geometrischer, radiometrischer und zeitlicher Unterschiede Pixelgenauigkeiten zwischen UAV- und multimodalen Fernerkundungsbildern. Die modellgetriebene 3D Flugplanungsmethode ermöglicht die Akquise mehrbildmessungsfähigen Nahbereichs-Bildsequenzen in dicht bebauten Umgebungen, die ein gesamtes Gebäude in einer geforderten Auflösung abdecken. Durch die Integration semantischer Merkmale in den Pfadgenerierungsprozess sind die resultierenden Trajektorien in Bezug auf die Flugsicherheit erstrebenswert, da sie eingeschränkte und gefährliche Lufträume, wie angrenzende Gebäude oder Straßen in den Planungsprozess integrieren. Die Verknüpfung einzelner unzureichend überlappender bildbasierter Gebäudeinnen- und außenmodelle wird durch die Registrierung geometrischer Strukturen angegangen, die in beiden Modellen unter Verwendung von 3D-Liniensegmenten als geometrische Merkmale geteilt werden. Eine Vielzahl von Experimenten an verschiedenen Gebäuden hat eine genaue Registrierung in Zentimetergenauigkeit bestätigt. Eine umfassende Evaluierung aktueller Methoden zur Einzelbildtiefenschätzung mit neuartigen Fehlermetriken an einem hochwertigen RGB-D Datensatz bewertet die aktuelle Eignung dieser Methoden für potenzielle praktische Anwendungsfelder und verdeutlicht die noch bestehenden Herausforderungen in diesem Forschungsgebiet. Unterstützt durch aussagekräftige experimentelle Evaluierungen, die die Leistungsfähigkeit der vorgeschlagenen Methoden bestätigen, markiert diese Arbeit einen Schritt in Richtung einer automatisierten, schnellen, genauen und sicheren Verwendung der UAV-Photogrammetrie.

CONTENTS

ABBREVIATIONS

ALS	airborne laser scanning
AT	aerial triangulation
BIM	building information model
cGAN	conditional generative adversarial network
CNN	convolutional neural network
CRF	conditional random field
DoF	degrees of freedom
DSLR	digital single-lens reflex
DSM	digital surface model
FCN	fully convolutional network
GCP	ground control point
GIS	geographic information system
GNSS	global navigation satellite system
GSD	ground sampling distance
ICP	iterative closest point
IMU	inertial measurement unit
INS	inertial navigation system
IoU	intersection over union
LiDAR	light detection and ranging
LoD	level of detail
MAV	micro aerial vehicle
MCMC	Monte Carlo Markov chain
MDE	monocular depth estimation
MIP	mixed-integer programming
MMS	mobile mapping system
MRF	Markov random field
MVS	multi-view stereo
NBV	next-best-view
NCC	normalized cross-correlation
NDVI	normalized difference vegetation index
OGC	open geospatial consortium
OSM	open street map
RANSAC	random sampling consensus
RGB-D	RGB-depth
RMSE	root mean square error

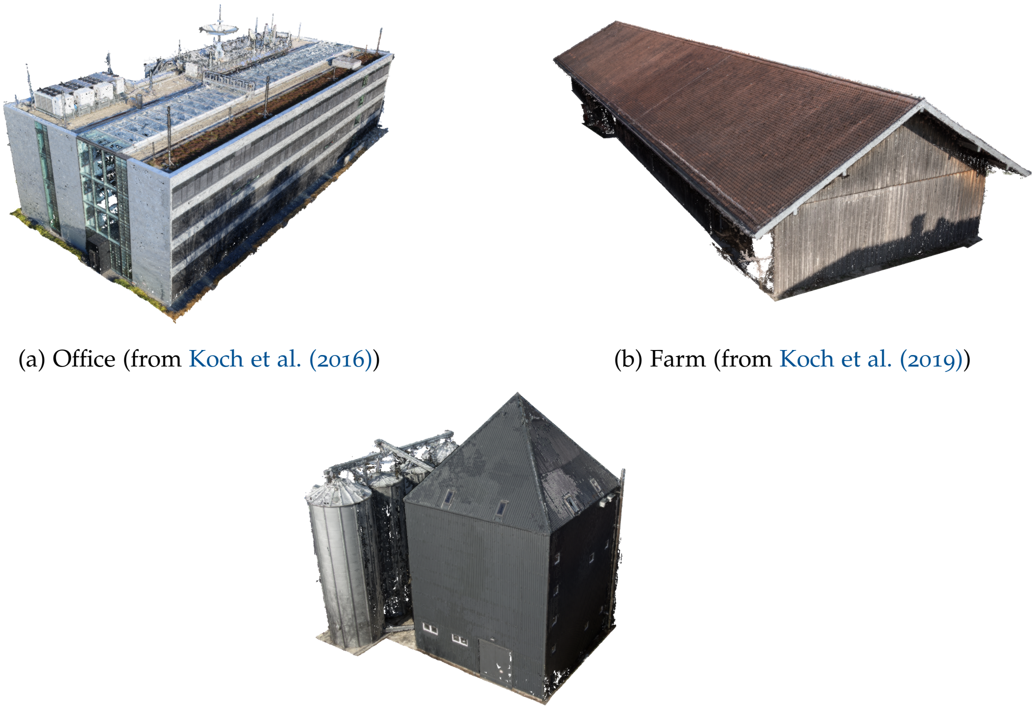
RPAS	remotely piloted aerial system
RPV	remotely piloted vehicle
RTK	real-time kinematic
SfM	structure from motion
SfS	shape from shading
SGM	semi-global matching
SIDE	single-image depth estimation
SLAM	simultaneous localization and mapping
SLIC	simple linear iterative clustering
TLS	terrestrial laser scanning
TSP	traveling salesman problem
UAS	unmanned aerial system
UAV	unmanned aerial vehicle
VHR	very high resolution
VRS	virtual reference system

1 INTRODUCTION

1.1 MOTIVATION

The generation of accurate and high-resolution as-built 3D building models is decisive in the field of 3D city planning and management by integrating heterogeneous geo-information and managing complex urban processes, such as 3D cadastre, property management, geo-design and intelligent transportation systems (Biljecki et al., 2015). However, highly dynamic and complex processes of urban development demand for efficient, cost-effective, and fast acquisition and processing methods for providing precise high-quality 2D and 3D geospatial data. Latest advances in microcontroller, hardware, and sensors have led to the emergence of an expanding market of small-scale unmanned aerial vehicles (UAVs) equipped with high-quality sensors, which have already proven to become an essential part of photogrammetry and remote sensing (Cummings et al., 2017; Yao et al., 2019). Equipped with cameras, these high-quality platforms eventually bridge the gap between close-range terrestrial and airborne photogrammetry and open various new application fields, while being a low-cost alternative to classic manned aerial photogrammetry. Compared to satellite-borne or airborne remote sensing, UAVs stand out for to their considerably higher spatial resolution, fast execution, and flexibility, turning them into a popular instrument for a variety of remote sensing applications, such as environmental monitoring, precision farming, cultural heritage documentation and civil engineering (Pajares, 2015). The ability to freely navigate to any accessible airspace and the complete control over perspective through modern gimbal systems enable UAVs to recover details that would remain unseen from aerial views. This property is particularly advantageous for 3D building reconstruction, since UAVs, in contrast to airborne systems, are able to capture both aerial-like nadir images, as well as oblique and even horizontal images of façade elements. Through the tremendous development of image-based 3D reconstruction in the fields of photogrammetry and computer vision, current structure from motion (SfM) and multi-view stereo (MVS) methods have proven their capability in UAV-based building reconstruction tasks yielding high-quality, dense and clutter-free 3D point clouds from image sets without rigid baselines in a comparable accuracy and density towards light detection and ranging (LiDAR)-based systems (Vacca et al., 2017). Samples of 3D building models generated from UAV image sequences in the course of this thesis are shown in Figure 1.1, while numerous works have confirmed the potential of UAV-photogrammetry for as-built building reconstruction tasks (Malihi et al., 2018; Wu et al., 2018).

However, limitations and challenges remain in the retrieval of geospatial products from UAV imagery. The attainable georeferencing accuracy of UAV images utilizing low cost and lightweight localization sensors, consisting of an inertial navigation system (INS) and global navigation satellite system (GNSS), is generally too low for precisely geo-localizing the resulting building models in a superordinate reference system, and thus integrating them with other geo-data. To avoid elaborative and costly deployment of ground control points (GCPs), indirect georeferencing methods



(a) Office (from Koch et al. (2016))

(b) Farm (from Koch et al. (2019))

(c) Silo (from Koch et al. (2019))

Figure 1.1: Samples of 3D building models generated from UAV images in the course of this thesis. The generated reconstructions stand out due to their high resolution, dense measurements, a high degree of detail, low noise and high completeness, particularly for building façades

can exploit the availability of numerous georeferenced image data from other modalities and data sources. By precisely localizing pixel correspondences in cross-modal image pairs, it is possible to achieve georeferencing accuracies of UAV target images in the range of the reference image. Nevertheless, the task of registering multi-modal image data is affected by significant radiometric, geometric, and temporal differences among the image pairs, demanding special attention in the development of a robust image matching approach. In contrast to large-scale mapping tasks, such as the generation of orthomosaics or digital surface models (DSMs), small-scale 3D building reconstruction, on the other hand, requires more demanding data acquisition techniques involving close-range images from different viewpoints and altitudes. Thus, flight planning substantially influences the quality of the resulting 3D model and — particularly in urban environments — is usually conducted manually or on the basis of potentially outdated image material. Besides legal regulations defining clear rules for the execution of UAVs, the safety of the vehicle, surrounding objects, and humans have to be guaranteed before and during each flight. Until now, this safety has to be ensured by the pilot. However, integrating these safety aspects into an automated flight planning scheme would lead to a more efficient and safer UAV campaign. Exploiting semantic cues of acquired UAV images can contribute to a better scene understanding and is therefore utilized for an automatic generation of safe 3D UAV paths, avoiding restricted and hazardous airspaces, while simultaneously aiming at acquiring a sequence of close-range images of a single object, suitable for creating complete and high-quality 3D models in an intended resolution. Due to the facilitated reconstruction of building models with UAVs, increased demand

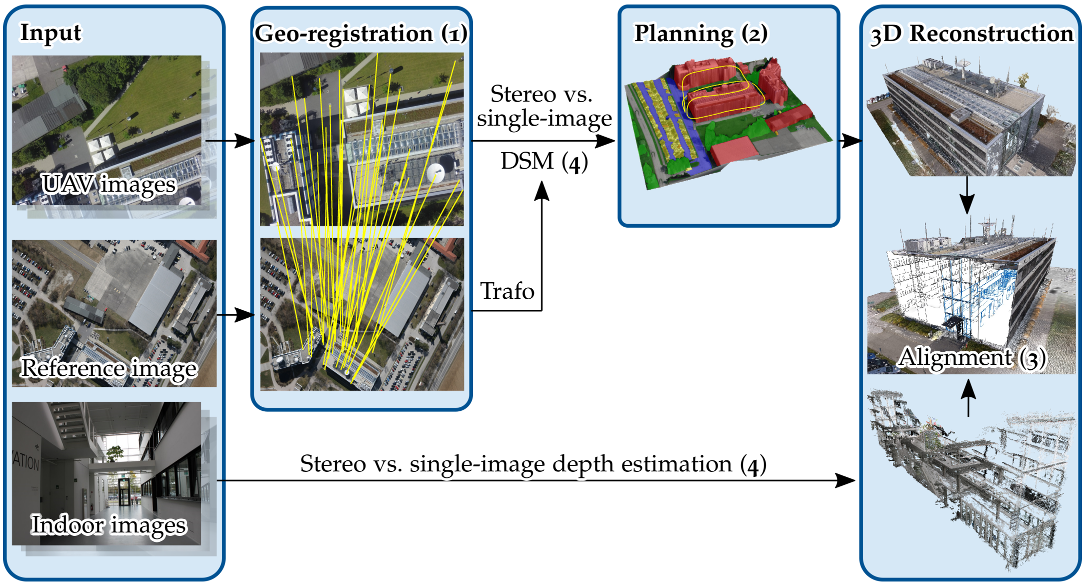


Figure 1.2: Combination of this thesis's contributions as a workflow to generate 3D building reconstructions using UAV-photogrammetry and image sequences of corresponding interior parts. Parenthesized numbers refer to the objective definitions in Section 1.2

for enriching those models with their interior parts is emerging. Joint indoor and outdoor building models enable seamless navigation and location-based services from street level to specific building locations. However, the reconstruction of interior models usually differs from the exterior model generation in terms of temporal acquisition and sensor selection. Lack of visual correspondences between shared building elements further complicates the task of an accurate alignment between both models. Although the interest in this task has already been reflected in various projects, primarily manual and cumbersome techniques have been used to address the alignment of indoor and outdoor building models. These include the deployment of clearly visible markers in the environment or a manual 3D alignment of the resulting 3D point clouds or 3D models in a post-processing step (Strecha et al., 2014). Instead of relying on unstable appearance-based correspondences, rigid geometric building structures (partially) visible in both models, such as window and door frames, serve as valuable features for an accurate and automatic registration.

1.2 OBJECTIVES

This dissertation aims to develop suitable methodologies for extending the capabilities of camera-based UAVs in terms of an automated generation of accurate and detailed 3D building models. The main objectives of this thesis are summarized as follows:

- **Objective 1: Georeferencing by co-registration of aerial and UAV imagery**

Accurate georeferencing of UAV imagery is crucial for subsequent tasks, such as DSM generation and 3D modeling. The low accuracy of onboard geo-localization sensors, however, hinders direct georeferencing, whereas alternative methods, such as registration of UAV images towards already georeferenced image data, can vastly enhance the georeferencing accuracy beyond sub-decimeter level. The challenges of matching multi-modal remote sensing images include large geometric, radiometric and temporal differences, which have to be considered in the developed registration approach to enable accurate pixel-wise correspondences.

- **Objective 2: Semantically-aware safe 3D UAV path planning**

The generation of UAV-based 3D building models requires multiple overlapping close-up images fulfilling complex requirements in the multi-view geometry. Automated flight planning based on a coarse terrain model would enhance the time-consuming and error-prone manual or semi-automated planning stage, particularly for tightly built-up environments. The identification and selection of suitable image acquisition viewpoints rely on incorporating multi-view requirements and photogrammetric properties in the path planning methodology. Besides, the semantic information embedded in UAV imagery can contribute to a more thorough scene understanding and be exploited to generate safe UAV paths that consider prohibited, restricted, and hazardous airspaces.

- **Objective 3: Alignment of indoor and outdoor building models**

The generation of holistic 3D building models composed of interior and exterior parts requires an accurate alignment between these individual 3D models. An automated registration approach has to address the complex challenges of a lack of visual correspondences and a limited degree of geometric overlap. A 3D line-based scene representation can contribute to the identification and registration of shared geometric structures in indoor and outdoor models, such as openings of windows and doors.

- **Objective 4: Evaluation of single-image depth estimation methods**

The emerging success of recent deep learning-based methods for predicting dense depth maps from single image views, referred to as single-image depth estimation (SIDE), will likely affect several UAV-photogrammetry and 3D building reconstruction areas. These might include generating coarse terrain models from single views, relaxing demands for image acquisition viewpoints, and modeling poorly textured and narrow spaced indoor environments. Despite the tremendous upsurge of this research field, a sophisticated evaluation protocol addressing relevant and interpretable geometric properties of a predicted depth map is still lacking. Current evaluation strategies and datasets can hardly provide in-depth analyses of developed methodologies and hamper further advances in this area.

The following contributions concerning the specified objectives can be combined to automate a UAV-photogrammetry campaign aimed at the reconstruction of complete 3D building models, as visualized in Figure 1.2, While an automated and precise georeferencing (1) and trajectory planning (2) contribute to the generation of exterior models, the investigation of current SIDE methods (4) aims at verifying their maturity for replacing established stereo vision-related tasks such as interior reconstruction and DSM generation. Finally, an automated alignment of indoor and outdoor models merges individual building parts (3).

1.3 PUBLICATIONS

This cumulative dissertation is based on the following peer-reviewed journal papers:

1. Zhuo, X., Koch, T., Kurz, F., Fraundorfer, F. and Reinartz, P. (2017). Automatic UAV Image Geo-Registration by Matching UAV Images to Georeferenced Image Data. Remote Sensing 9 (4), p. 376
2. Koch, T., Körner, M. and Fraundorfer, F. (2019). Automatic and Semantic-aware 3D UAV Flight Planning for Image-based 3D Reconstruction. Remote Sensing 11 (13), p. 1550
3. Koch, T., Liebel, L., Körner, M. and Fraundorfer, F. (2020). Comparison of Monocular Depth Estimation Methods using Geometrically Relevant Metrics on the IBims-1 Dataset. Journal of Computer Vision and Image Understanding, Volume 191, 102877

and one peer-reviewed conference paper:

1. Koch, T., Körner, M. and Fraundorfer, F. (2016). Automatic Alignment of Indoor and Outdoor Building Models using 3D Line Segments. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR-WS)*, pp. 10–18

Pre-print versions of the published papers are provided in Appendices A to D.

1.4 THESIS OUTLINE

The dissertation is organized into five chapters. The motivations and objectives of the thesis have already been addressed in this chapter. In Chapter 2, a brief introduction on the applicability of UAVs in remote sensing demonstrates their successful usage alongside traditional remote sensing platforms. In addition, a comprehensive study regarding the generation of 3D building models from various platforms and sensors is presented. Chapter 3 describes the current state-of-the-art and limitations in using UAV imagery for 3D reconstruction applications, including the tasks of georeferencing, depth estimation, flight planning and building model registration. Summaries of the contributions developed in the course of this thesis are provided in Chapter 4, including the tasks of accurate image georeferencing, semantic-aware 3D path planning for MVS image acquisition, the registration of individual indoor and outdoor building models as well as the evaluation of existing SIDE methods. A discussion on the proposed works and their applicability in related fields, as well as an outline on possible future work, concludes this thesis in Chapter 5.

2 THE ROLE OF UAVS IN REMOTE SENSING AND BUILDING RECONSTRUCTION

2.1 UAVS IN COMPARISON TO CONVENTIONAL PLATFORMS

Unmanned and remotely controlled airborne platforms, which are popularly known as drones, have numerous designations, such as unmanned aerial vehicle (UAV), unmanned aerial system (UAS), micro aerial vehicle (MAV), remotely piloted vehicle (RPV), and remotely piloted aerial system (RPAS). These include remotely controlled, semi-autonomous, or fully autonomous systems. Although no unified standard has been yet established for the classification of civil and commercial UAVs, the aspects of weight, payload, and frame structure are often used to categorize the enormous variety of currently existing systems. However, these guidelines may vary from one state to another. According to current German law (October 2019), UAVs are classified into 0.25–2 kg, 2–5 kg, 5–25 kg, and more than 25 kg, whereas more stringent regulatory conditions must be complied with as their weight increases. Thus, solely a mandatory identification of the device applies to the first group, whereas the use of UAVs between 5–25 kg is only permitted with individual ascent permissions. Systems of more than 25 kg are generally prohibited by law in the civilian sector. According to the frame type, UAVs can be further classified into fixed-wing, multi-rotor, or hybrid systems. While fixed-wing systems stand out for long operation times and high airspeed allowing to cover large-scale environments in flight, the flexibility of multi-rotor UAVs and their hovering capability enable precise navigation even in narrow and tightly built environments. Hybrid systems should combine the advantages of both systems, but are still under development (Saeed et al., 2018).

Alongside the vast progress in UAV systems, microcontrollers, and navigation sensors, remarkable advances have been achieved in the development of high-quality small-scaled sensors that are lightweight enough to be carried even by mini-UAVs, such as systems below 5 kg. Equipped with cameras, RGB-depth (RGB-D) cameras, thermal or infrared cameras, multi-spectral cameras, or even light detection and ranging (LiDAR) sensors, UAVs have become valuable remote sensing platforms and geo-data suppliers.

As a result of tremendous economic and social developments, Earth is continually undergoing major changes in rural and urban areas that demand for flexible, fast, cost-effective, and automated data acquisition and processing strategies for earth observation on a small and large scale. Today's predominant possibilities of remote sensing-based earth observation and geospatial information acquisition include satellites, manned aircraft and UAVs. Despite its young age, UAVs have already become an indispensable part of remote sensing to address these challenges and has gained ground in various applications in both research and practice. The following chapter presents a detailed comparison between these remote sensing platforms in terms of different aspects of data quality and operability for remote sensing applications. The comparison exclusively refers to camera-supported platforms. A

Table 2.1: Comparison between different platforms used for photogrammetric applications. The table merges parts of the works of Zhuo et al. (2017) and Gerke (2018)

	Satellite	**Manned Aircraft**	**UAV**
Coverage	Global	km²	m² - km²
GSD	dm - m	cm - dm	mm - cm
Capture Geometry	nadir	nadir, oblique	full flexibility
Onboard GNSS/IMU	high quality dm-level accuracy	high quality cm-level accuracy	low - moderate quality m-level accuracy
Price	very high	high	low - moderate
Operating cost	low	high	low
Flexibility	fixed orbit weather-dependent autonomous	less mobile weather-dependent pilot needed	mobile cloudy, drizzly weather remotely controlled
Applicable in hazardous areas	yes	partial	yes
Optical sensors	RGB, multispectral, hyperspectral	RGB, multispectral, hyperspectral, thermal	RGB, multispectral, hyperspectral, thermal

concise summary and visualization of the key aspects are listed in Table 2.1 and depicted in Figure 2.1.

Coverage refers to the spatial extent of the observed area and is mainly determined by the flight altitude. Regarding global monitoring and mapping the earth's surface, it is beyond any question that in addition to radar-based satellites, current optical satellites such as the WorldView series, GeoEye, Pléiades, SPOT, KOMPSat, or Planet Labs provide earth observation data on an unprecedented scale. Fixed orbits at high altitudes up to 700 km facilitate swath widths as high as 20 km, enabling the Earth's surface to be largely covered (up to 65° latitude) with only a few days of revisiting times. Manned aircraft can cover large regional territories up to several hundreds of kilometers on a single flight due to their relatively high flight altitude of several hundreds of meters up to a few kilometers. Because of its low flight altitude of only a few hundreds of meters (and even legal restrictions to 100 m altitude in many countries), as well as short flight endurances, UAVs are limited to cover comparable small regional areas. Most current multi-rotor UAVs can not exceed a coverage of more than 0.50 km² at 100 m altitude with 70 % overlap of adjacent images, whereas fixed-wing UAVs can survey up to 10 km² with a single flight.

Spatial resolution is usually reported as a measure of the smallest object that can be resolved by the sensor and depends on the flight altitude and the focal length and resolution of the equipped sensor. It can be expressed as the ground distance of one pixel in the image, referred to as ground sampling distance (GSD). The high altitude of satellites constitutes a major drawback in terms of the spatial resolution of the recorded data. Although imaging sensors on satellites are constantly improving, the spatial resolution even of very high resolution (VHR) satellite imagery is currently

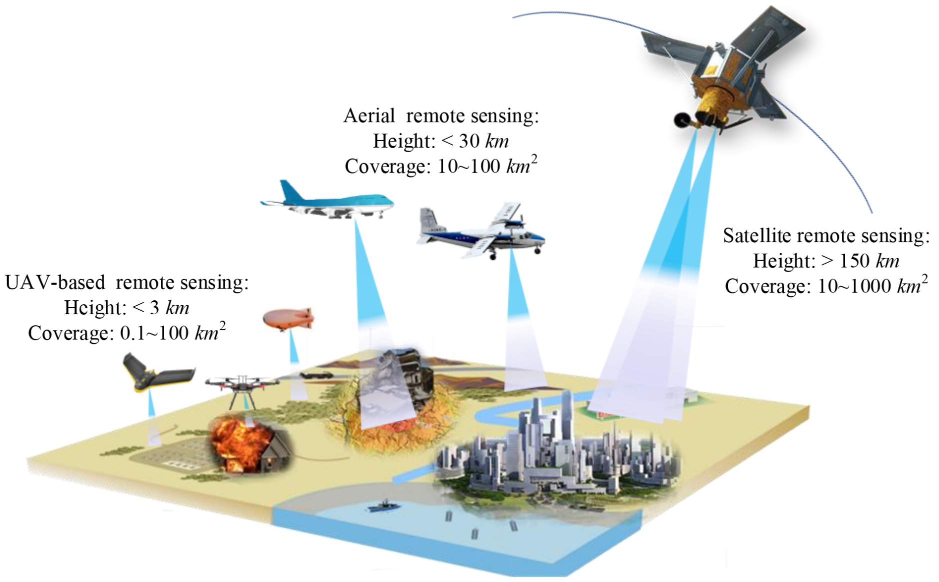
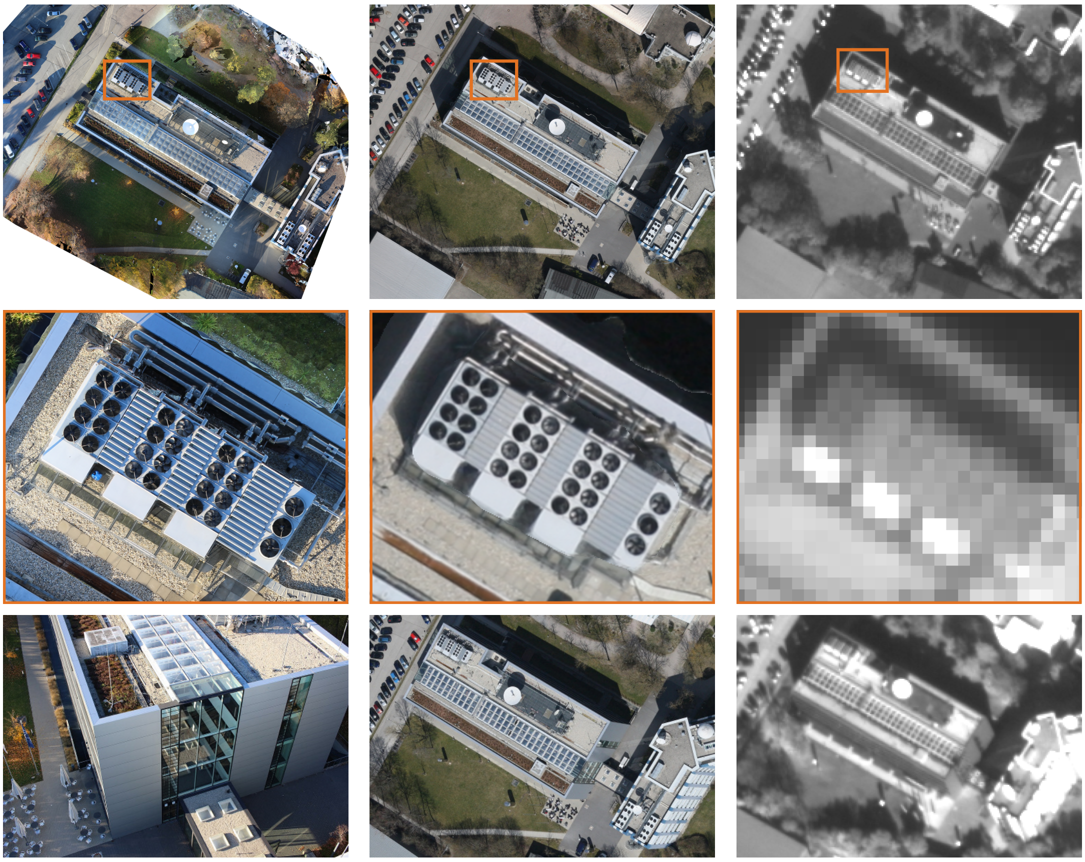


Figure 2.1: Comparison of different remote sensing platforms. Image source: Xiang et al. (2018)

confined to at best 30 cm GSD for WorldView-3 and WorldView-4 images and 50 cm for WorldView-1 and WorldView-2, Pléiades, and KOMPSat-3. This resolution may already be sufficient for many applications but does not entirely cover fields requiring higher precision, such as exact building modeling. Manned aircraft equipped with large format digital imaging sensors, such as Vexcel UltraCam (Wiechert et al., 2011), Leica DMC III (Mueller and Neumann, 2016), or Z/I DMC II (Neumann, 2011) achieve much finer GSDs up to 5–10 cm, depending on the flight altitude. Over many decades, manned aircraft set the bar for high-precision aerial images. However, with the advent of UAVs and high-resolution small-format cameras, spatial resolutions in centimeter and even millimeter range have become feasible given the unsurpassed flexibility of flight altitudes between a few meters up to hundreds of meters. Figure 2.2 compares orthomosaics generated from UAV and aerial images towards a WorldView-2 satellite image of the same scene. The high resolution and low acquisition altitude of UAV images clearly reveal a substantially enhanced level of detail in the depicted objects, which even surpasses that of airborne images.

Temporal resolution in remote sensing expresses the amount of time needed to repeatedly acquire data for the exact location. Many applications demand for in-situ measurements, such as customized data acquisition campaigns or post-disaster management. Other domains require high temporal resolutions and frequent data acquisition, such as precision farming or monitoring tasks. Although revisit cycles of satellites have already been decreased to a single day, this may not be sufficient for monitoring continuous terrain changes or necessary in-situ measurements. Due to atmospheric distortions in high altitudes and weather conditions, it is furthermore not guaranteed that each cycle provides usable imagery. Manned aircraft are capable of collecting data without the limitation of revisit periods. However, they suffer from complex logistics of flight preparation and the availability of pilots. These high logistic efforts and the requirement of nearby runways slow down the response time for urgent tasks. Frequent data acquisition campaigns in the same area are generally



(a) UAV

(b) Aerial

(c) Satellite

Figure 2.2: Comparison of remote sensing imagery acquired from a UAV (a), manned aircraft (b) and VHR satellite (c). First row shows an orthomosaic derived from UAV and aerial image sequences and a cropped WorldView-2 image to the same image content. By visualizing an enlarged image part, the higher degree of local details obtained from UAV images is revealed (middle). The GSD for UAV, aerial and satellite images is 1 cm, 5 cm, and 30 cm, respectively. The last row shows oblique images of the same scene. The used data is part of the TUM-DLR multi-modal earth observation benchmark (Koch et al., 2016b) which was established in the course of this work

feasible, however, extensive logistical and monetary costs are usually too high for practical realization. UAVs, on the other hand, are ideally suited for both in-situ measurements and arbitrary acquisition frequencies, since they are flexible, easy to operate, able to start, fly, and land in hazardous areas and can already be set up for autonomous flights with pre-designed trajectories for monitoring applications.

Data quality is of high importance in remote sensing and embraces image quality, correctness, georeferencing accuracy and availability. The high altitude of satellites influences the image quality since the images are affected by atmospheric distortions. Bad weather conditions and clouds might lead to unsuitable and unusable image data. Based on meteorological data for the region of Central Europe, Van der Wal et al. (2013) emphasized that merely 20 % of obtained satellite imagery from daily revisiting satellites are of adequate image quality. Lower flight altitudes of manned aircraft mitigate these effects allowing for more reliable data acquisition even for adverse weather conditions. Modern UAVs are also capable of flying under drizzly weather conditions. At the same time, the close distance between sensor and surface still allows to capture relatively sharp images assuming the equipment of high-quality large-format cameras with fast lenses that enable quick shutter speeds. In general, the quality of the equipped imaging sensor profoundly affects the image quality, and thus the resulting geo-product. Satellites and manned aircraft do usually not have strict payload issues and are able to carry heavy sensors. As a result, they have access to the most precise and accurate remote sensing sensors currently available. The dependence on lightweight sensors to be carried by a UAV constitutes a major drawback. However, the quality of small-scale imaging sensors has been vastly improved in recent years providing valuable earth observation data in very high resolution (Colomina and Molina, 2014). One of the biggest challenges in remote sensing is to accurately integrate the acquired image data into a consistent reference frame. This georeferencing task requires high-quality positioning systems mounted on the platform that assigns each image with precise external orientations. In addition to extremely accurate measurements usually obtained from a global navigation satellite system (GNSS) and inertial navigation system (INS), a precise synchronization between localization sensors and the imaging sensor is essential. Such high-precision localization systems are currently still enormous in terms of size, weight, and price and, therefore, only deployed on satellites and manned aircraft. However, they allow for georeferencing accuracies in decimeter range, while lightweight and low-quality localization sensors equipped on UAVs can only reach accuracies up to a few meters. A comparison of different low-cost and high-quality localization sensors is provided in the work of Colomina and Molina (2014).

Flexibility plays an essential role in the use of remote sensing methods for specific applications. Although modern remote sensing satellite missions consist of multiple satellites for increasing the temporal resolution, they are fixed on their orbits and limited in terms of the acquisition geometry capturing primarily nadir-directed images. At most, sensors can be manually panned to some degree. Manned aircraft possess a higher degree of flexibility as their trajectory can, but also has to be planned individually for each mission, which, however, is time-consuming and requires complex logistics. The safety of pilots limits the flexibility in hazardous areas, while imposed restrictions on the airspace often constrain their usage in urban and highly populated areas. In contrast to simple grid-like flight patterns conducted by airplanes, helicopters enable more flexible maneuverability and customized flight paths, while requiring smaller launching and landing areas. Besides commonly used nadir-directed cameras, modern photogrammetric imaging systems, such as the DLR-3K system (Kurz et al., 2011) utilizes multiple calibrated and tilted cameras mounted on a platform allowing for synchronized nadir and oblique image sequences from the same position. UAVs can be deployed on-demand, and flight parameters can be adjusted in accordance with the desired image resolution and perspective. The last row in Figure 2.2 shows samples of oblique images obtained from the different platforms. The flexibility of gimbal adjustments of UAVs is particularly beneficial in capturing building façades, which are often highly distorted or even remain unseen in aerial and satellite imagery. The small size and easy control of UAVs facilitate launching and landing even in uneven and steep landscapes while their full flexibility and maneuverability allow for fast motion, hovering, quick turnarounds, and the ability to keep current positions. The independence from pilots permits to fly even in hazardous and narrow spaced regions. However, due to the big growth in the

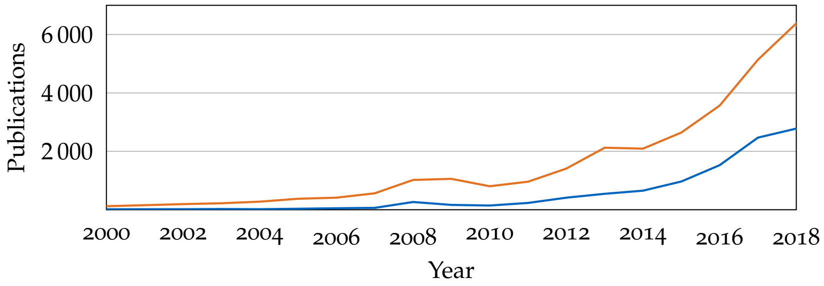


Figure 2.3: Upsurge in published papers combining the fields of UAVs, photogrammety and remote sensing. Results derived from a dimensions.ai literature search[3] on April, 25, 2019, looking for the terms *UAV* and *Photogrammetry* (—) and *UAV* and *Remote Sensing* (—)

commercial market, governmental regulations have become stricter in recent years, complicating the usage of UAVs in populated and protected areas. An overview of current regulations for UAV operations on a global scale is given by Stöcker et al. (2017)[1].

Costs are a decision criterion in the selection of remote sensing-based geo-data. The immense expenses of planning, constructing and maintaining satellites, as well as processing the image data reason for the relatively high cost of purchasing VHR imagery up to 50€ per km^2 for a high-resolution stereo image pair with 50 cm GSD[2]. Manned aircraft also suffer from high costs due to the expenses of the vehicle, pilots and logistics. UAVs have witnessed a rapid decrease in purchase costs to a few thousands of Euro for mid-quality devices and a few tens of thousands for high-quality platforms. At the same time, the cost of high-quality optical imaging sensors has also got reduced to an affordable range. Besides the purchase cost, there is no further cost necessary, which decreases the total acquisition costs already after a few flights.

2.2 APPLICATION FIELDS OF UAVS IN REMOTE SENSING

The utilization of UAVs in photogrammetry and remote sensing has recently gained increasing importance, which is also evident in the number of published scientific articles, as shown in Figure 2.3. The following section provides a brief overview of current application fields of UAVs in remote sensing. An excerpt from the presented works is listed in Table 2.2.

2.2.1 *Environmental Monitoring*

Environmental monitoring describes the observation of scientifically relevant aspects of the environment and the documentation of ecological parameters. Besides the diagnosis of climate and human impacts on natural and agricultural systems, the investigation of hydrological processes and the prevention of natural disasters are elementary tasks within this discipline. Most monitoring systems are based on a

[1] The report refers to international regulations in 2017 and may no longer be accurate for some countries

[2] The stated costs are based on exemplary purchase prices from renowned geodata providers on January 11, 2021 (`https://apollomapping.com`, `http://www.landinfo.com/`)

[3] `https://app.dimensions.ai/`

Table 2.2: Application fields of UAVs in remote sensing

Application Field	**Task**	**Literature**
Environmental Monitoring	Precision wildlife monitoring	Hodgson et al. (2016)
	Monitoring of land cover changes	Ahmed et al. (2017)
	River monitoring	Detert and Weitbrecht (2015)
Disaster Management	Assistance after avalanche catastrophe	Bejiga et al. (2017)
	Wildfire detection and prediction	Merino et al. (2012)
	Assistance after earthquakes	Qi et al. (2016)
Agriculture and Forestry	Precision farming	Gómez-Candón et al. (2014)
	Crop monitoring	Maes and Steppe (2018)
	3D canopy height estimation	Saari et al. (2011)
Cultural Heritage Preservation	Excavation mapping	Sauerbier and Eisenbeiss (2010)
	Modeling of archaeological sites	Fernández-Hernandez et al. (2015)
	Historical city reconstruction	Balsa-Barreiro and Fritsch (2018)
Civil Engineering	Monitoring of transportation systems	Brooks et al. (2015)
	Monitoring of construction sites	Tuttas et al. (2017)
	Crack detection	Gopalakrishnan et al. (2018)
	Building reconstruction	Vacca et al. (2017)

variety of terrestrial, manned airborne and satellite observations, however high costs, as well as low spatial and temporal resolutions, are bottlenecks of both global and local monitoring tasks (Manfreda et al., 2018). Moreover, some habitats do not allow for on-ground surveys as they might damage the natural ecosystem. Therefore, UAVs have the potential to bridge the gap between field observations and traditional airborne and satellite-borne remote sensing and, thus to improve spatial, spectral, and temporal data retrieval. Researchers have already demonstrated the potential of using UAVs for numerous monitoring tasks, such as for plant population (IV et al., 2006), phenology (Klosterman and Richardson, 2017), pest infestation (Lehmann et al., 2015), land cover change (Ahmed et al., 2017), and biomass estimation (Dittmann et al., 2017). UAVs allow monitoring of river system dynamics with a level of detail that is several orders of magnitude higher compared to alternative remote sensing platforms, leading to a deeper understanding of hydrological processes (Detert and Weitbrecht, 2015). A recent survey on the use of UAVs for environmental monitoring outlined the benefits and increasing usage in various application fields but also demonstrated remaining tasks and challenges, such as inconsistent GSDs for elevated areas due to missing 3D flight planning techniques that consider the underlying surface structure, heterogeneous regulations across different countries and the challenge of accurate image geo-registration without the deployment of ground control points (GCPs) (Manfreda et al., 2018).

2.2.2 *Disaster Monitoring and Search and Rescue*

The rapid utilizability of UAVs in the event of a disaster is beneficial in assisting emergency responders by providing real-time aerial imagery for supporting rescue planning and decision making. The rapid retrieval of high-quality 2D and 3D spatial data for relatively large areas reduces the operational time and number of required rescuers and minimizes the risks for search and rescue missions. An exemplary search and rescue campaign after a simulated strong earthquake showed a greatly improved search efficiency and rescue strategy by automatic detection of collapsed buildings in UAV images (Qi et al., 2016). To protect rescuers during their search for missing people and victims in the ruins of collapsed buildings, Puerta and Fraundorfer (2016) developed a semi-automatic environmental-aware UAV navigation strategy. Images of a manual or pre-designed overview flight are used to generate a 3D occupancy map defining free and occupied airspaces, which is frequently updated in real-time during the flight. The accessible airspace with respect to the map can then be used for an automatic collision-free navigation of the UAV even in complex and narrow 3D scenarios. A live stream of the UAV camera and an integrated 2D human recognition system enables a safe and fast search and rescue operation for environments that are too dangerous for human access due to the risk of collapsing buildings. The real-time capability of onboard image processing and data transmission to a ground station can save valuable time in searching for missing persons (Sun et al., 2016). For instance, victim localization after a catastrophe requires immediate action and online feedback for the fastest possible response for human rescue missions. Sensors, onboard computational power, and robust data transmission are sufficient to employ even state-of-the-art computational expensive convolutional neural networks (CNNs) with real-time feedback for the rescue team after an avalanche occurred (Bejiga et al., 2017). An automatic framework for wildfire detection and the prediction of the evolution of forest fires has experimentally proven the applicability of UAVs for fire fighting activities (Merino et al., 2012). UAVs were also utilized for early damage assessment and post-event reconstruction planning after an earthquake damaged thousands of buildings in Italy in 2009 (Baiocchi et al., 2013). Damaged and collapsed buildings were detected in real-time in a video stream which helped to receive a timely estimate about the extent of the disaster. A damage assessment after the 2015 Gorkha earthquake in Kathmandu could be achieved by utilizing UAV imagery for creating a 3D reconstruction of the destroyed environment (Yamazaki et al., 2015).

2.2.3 *Agriculture and Forestry*

The high temporal and spatial resolution of UAV imagery has led to a new range of time-critical agriculture-related applications, such as determining harvest schedules (Khanal et al., 2017) and precision farming (Gómez-Candón et al., 2014). Lightweight multi-spectral and hyper-spectral sensors mounted on UAV have proven to estimate useful parameters used in agriculture monitoring, such as the generation of normalized difference vegetation index (NDVI) maps, which feature comparable accuracy towards satellite-based observations, but with immediate feedback for farmers and in a higher spatial resolution (Manfreda et al., 2018). Further variables derived from UAV-based monitoring systems include, for instance, crop water stress index (Park et al., 2015), photosynthetic activity (Zarco-Tejada et al., 2013b) or carotenoid content (Zarco-Tejada et al., 2013a). These variables are used for various agricultural applications, such as pest control, crop monitoring, field surveys, sowing and spraying (Maes and Steppe, 2018). Precision agriculture tasks require highly accurate orthomosaics which can be derived with the help of GCPs (Gómez-Candón et al., 2014). Becirevic et al. (2019) have shown that crop heights can be estimated from multi-temporal UAV imagery in a comparable accuracy than manual ruler-based height measurements. A reference digital surface model (DSM) was computed before the growing season, while differences towards DSMs from other time steps yield accurate crop field heights in high spatial resolution. UAVs have also proven to assist in numerous forestry applications, such as forestry activity monitoring, species identification, and tree height estimation (Adão et al., 2017; Getzin et al., 2012; Saari et al., 2011). Natesan et al. (2019) proposed a CNN for tree species classification from nadir directed UAV images from 150 m altitude. Even with the current development of nano-satellites, UAVs offer unique features, such as the combination of 3D canopy height measurements, orthomosaics and multi-angular data (Maes and Steppe, 2018). However, challenges still arise in terms of fully automated pipelines for such applications, including the tasks of automatic flight preparation, flight planning, accurate and effortless georeferencing, as well as different and frequently changing flight regulations across different countries (Maes and Steppe, 2018).

2.2.4 *Cultural Heritage Documentation*

The recording and 3D modeling of complex archaeological sites is usually associated with high monetary, logistical and temporal costs, since, typically, laser scanners, tachometers, tapes, rulers, and numerous people are required for the documentation task, which often demands daily updates. In the 2000s, laser scanners were the most common sensors, however, recent times have witnessed an increasing interest in using UAVs for creating DSMs, orthomosaics, and 3D models. The dynamic processes of archaeological excavations feature immense terrain changes which require easy, fast, and accurate data retrieval methodologies for timely updates of the current excavation process and campaign management. Pre-designed acquisition flights can be easily executed in a high frequency and the capability of acquiring nadir and oblique images of hardly observable parts reveals the potential of generating orthomosaics, elevation models and full 3D models of excavated objects or entire historical cities (Balsa-Barreiro and Fritsch, 2018; Fernández-Hernandez et al., 2015; Sauerbier and Eisenbeiss, 2010; Themistocleous et al., 2015). A comparison between UAV and terrestrial LiDAR DSMs has proven the suitability of camera-equipped UAVs with only slight differences between both DSMs of few centimeters, while multiple acquisition views from UAVs resulted in less occlusions in the DSMs than LiDAR-based mapping (Eisenbeiss and Zhang, 2006). Mostegel et al. (2017) presented a multi-scale surface reconstruction pipeline for generating a detailed joint 3D surface model of a prehistorical rock art site from multi-modal and multi-scale imagery acquired from a manned hang glider, a fixed-wing UAV, an autonomous octocopter UAV and a terrestrial stereo setup. The individual reconstructions were georeferenced with the use of GCPs and merged towards a consistently connected 3D mesh with a spatial resolution that ranges from 1 m for the surrounding environment up to 50 µm for the engraved rocks. An extensive survey on the potential of UAVs for archaeological applications was presented by Campana (2017).

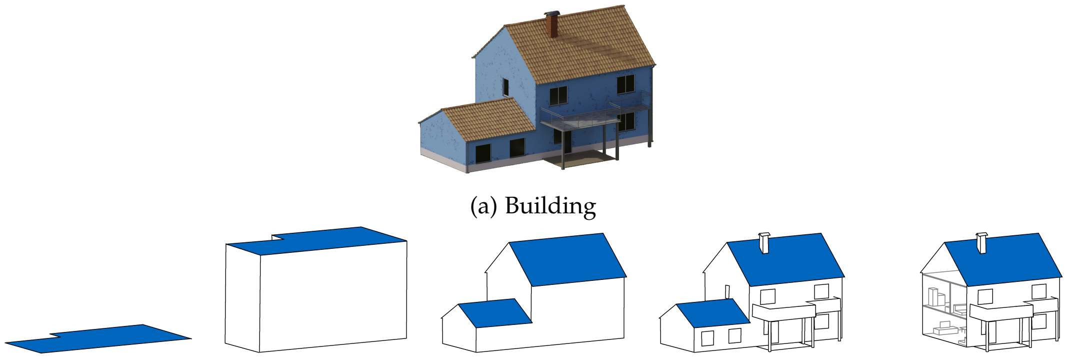
2.2.5 *Civil Engineering*

A tremendous amount of effort related to UAV development was made in the realm of civil and construction engineering. The attraction of integrating UAVs in many civil infrastructure applications is primarily based on accelerating accessibility to remote and dangerous sites. It plays an important role in decision making and can be used comprehensively for urban planning analysis, tourism, and other fields. UAVs — mostly equipped with cameras or LiDAR sensors — are used for urban planning, monitoring of linear structures, such as pipelines (Rathinam et al., 2008), in bridge inspections (Ellenberg et al., 2014), as assistance in construction sites (Tuttas et al., 2017) and for traffic monitoring in transportation systems (Brooks et al., 2015). Gopalakrishnan et al. (2018) proposed a deep learning-based approach for crack detection in close-up images of diverse civil infrastructure systems, such as storage silos or local roadways. Zhuo et al. (2018) utilized oblique UAV images and semantic segmentation for refining open street map (OSM) building footprints, showing the potential of updating currently available topographic maps. Based on an elaborate project, Zekkos et al. (2018) shared some lessons-learned of UAV-photogrammetry after conducting numerous campaigns in 26 different sites across the world. They reported promising and satisfying results that were comparable towards LiDAR but also indicated current limitations, such as the requirement of complicated flight plans by experienced pilots, the modeling of poorly textured objects, and the considerations of dynamic governmental regulations of UAV operations.

As will be discussed in more details in the following chapter, UAVs offer very effective tools in terms of (3D) urban planning, since they can provide an indispensable basis of useful geo-data for management and planning, ranging from accurate medium-scale 3D maps, orthomosaics and DSMs up to highly-accurate 3D models of single objects in centimeter-level resolutions. However, the use of UAVs is not regulated by similar stringent safety standards as manned aviation, and — although the development of UAV control is already at an advanced stage — unexpected and uncontrolled crashes can still occur (Clothier and Walker, 2015). Such failures, caused by wind, undesirable human interaction, malfunctions, or inaccurate navigation due to poor satellite constellations or fully GNSS-denied areas can cause crash-landings, and thus damage to the operation of the targeted facility or even to humans. A risk assessment of a potential crash needs to be considered during the campaign planning, which could vastly differ for employing UAVs in urban and industrial environments than for flying in large, rural and mostly uninhabited spaces. Although UAV breakdowns and crashes are unpredictable, precautions should be taken during the flight planning stage, *e.g.*, by avoiding hazardous flights above populated and dangerous areas. Ensuring safety remains a manual task for pilots and path planners and usually requires expert knowledge and experienced pilots.

2.3 COMPARISON OF PLATFORMS, SENSORS AND METHODS FOR 3D BUILDING RECONSTRUCTION

Since the mid-1990s, the demand for 3D city models is continuously growing and has meanwhile become commonplace. Remote sensing provides an indispensable foundation for the generation of 3D buildings and 3D city models in small and large scales. With the official open geospatial consortium (OGC) standard for modeling



(a) Building

(b) LoD-0

(c) LoD-1

(d) LoD-2

(e) LoD-3

(f) LoD-4

Figure 2.4: Definition of different abstraction levels for building representation in accordance with CityGML standard (Kolbe et al., 2009)

and exchanging virtual 3D city and landscape models, CityGML proposes the demands for easy, clear, and explicit handling of topographical and semantical information (Kolbe et al., 2009). Reconstruction, modeling, and representation of as-built buildings are of particular interest in the field of building information model (BIM) and — depending on the defined quality requirement for a civil project — 3D buildings can be represented in different resolutions and abstraction levels, which is commonly known as level of detail (LoD) in computer graphics. LoDs in BIMs are different geometrical, graphical, and semantical representations of built assets beginning with generic 2D models at the lowest LoD (LoD-0) to various amounts of graphic and non-graphic information attached to 3D modeled objects (LoD-3), up to the supplementation with interior features (LoD-4). Figure 2.4 visualizes different abstraction levels applied for 3D building models.

The foundation of creating such 3D building models is based on the acquisition of 3D information from either direct (*e.g.*, LiDAR) or indirect measurements (*e.g.*, photogrammetry). Prerequisites on the acquired or processed 3D data for subsequent building modeling is a precise georeferencing, high resolution, low noise, and complete recovering of the objects. If these requirements are met, techniques exist for generating 3D building models in agreement with CityGML standards. Most data acquisition systems collect dense 3D information that can be further processed into watertight and polyhedral 3D models (Balsa-Barreiro and Fritsch, 2018; Duan and Lafarge, 2016; Kim and Shan, 2011; Malihi et al., 2018; Poli et al., 2015; Pu and Vosselman, 2009; Sampath and Shan, 2010; Toschi et al., 2017; Verma et al., 2006; Wen et al., 2019; Wu et al., 2017; Yang et al., 2016). The first part of the following chapter describes the state-of-the-art in generating 3D building models up to LoD-3 from different remote sensing data, while the second part deals with LoD-4 models and their relation to remote sensing.

2.3.1 *Reconstruction of Building Exterior from Remote Sensing Data*

This section provides an exhaustive review on the state-of-the-art for as-built reconstruction of building exteriors. Since this is a long-standing task with increasing interest in various fields, numerous different ways have been developed during the last decades, focusing on an automated procedure for both large-scale and small-scale modeling. To this end, building reconstruction has been addressed with various sensors from different platforms. With the widely-accepted standard of CityGML, clear rules have been established regarding model preparation, accuracy, information richness, representation, and visualization. Recent works of as-built building reconstruction try to formulate their approaches following the rules of CityGML. According to this standard, buildings are represented via different LoDs, which are visualized in Figure 2.4. While LoD-0 solely depicts an accurate building footprint in a global reference frame, LoD-1 additionally adds the building's height resulting in a simplified box-shaped representation of the building. LoD-2 models are obtained by augmenting these models with real roof shapes. These purely geometric representations of the building's shape are currently widely spread among many geographic information systems (GISs). However, they do not exhibit façade and detailed roof elements, nor overhangs, or balconies. Adding these local details to the building models yield comprehensive and complex LoD-3 models for which, however, accurate and extensive geo-information is required. As a final step, these pure exterior representations of the building can be supplemented and completed with an interior model resulting in LoD-4 models. As depicted in Figure 2.5, diverse platform and sensor combinations can address the task of 3D building reconstruction, including satellite and airborne sensing approaches using active (*e.g.*, LiDAR) and passive (*e.g.*, cameras) sensors, as well as UAV-borne imaging and ground-based sensing methods from static or moving platforms (*e.g.*, mobile mapping systems (MMSs)). Apart from different data acquisition strategies, a variety of data processing methodologies have been developed for 3D building reconstruction up to LoD-3, which are described in the following section. Table 2.3 lists the main features of the presented techniques with regard to the possibility and quality of 3D building reconstruction.

Satellite-based remote sensing, despite its long history since the launch of Landsat-1 as the first multi-spectral remote sensing satellite in 1972, has primarily focused on natural environments, since early sensors did not achieve the required spatial resolution for urban analysis and modeling. Only with the latest technical development of VHR optical satellite-based sensors, spatial resolutions have reached sub-meter ground resolutions allowing to provide crucial information for urban planing on building level (Weng et al., 2018). Nowadays, a multitude of optical satellites provide large-scale data of adequate quality for urban modeling purposes, while the ongoing development of new satellites will further increase their quantity and continuously improve the achieved spatial resolutions (Zhu et al., 2018b). Due to the nearly global coverage in a high temporal resolution, satellite-borne single or stereo images allow for automated extraction of 3D city models utilizing photogrammetric or machine learning-based methods. An assessment of the geometric and radiometric quality of stereo imagery from VHR sensors (GeoEye-1, WorldView-2, and Pléiades-1A) for the task of 3D information extraction was conducted by Poli et al. (2015). The high altitude of these satellites affects the radiometric image quality by atmospheric distortions, and the localization accuracy and inaccurate camera intrinsics influence the geometric quality of satellite images. This work reported acceptable radiometric distortions for 3D modeling, while comparable geometric accuracies of derived DSMs have been observed for the investigated satellites. Accuracies of height estimations range between 6–8 m root mean square error (RMSE), which were assessed with the help of airborne LiDAR ground truth data. Many errors occurred in the case of

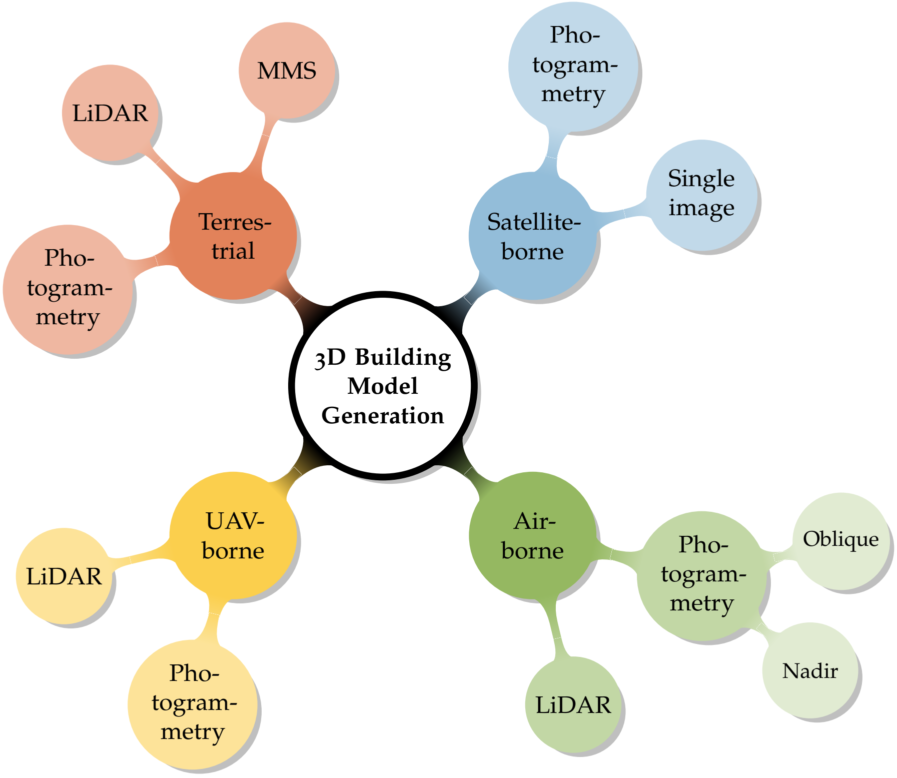


Figure 2.5: Data acquisition for the task of 3D building reconstruction addressed by different remote sensing platforms and sensors

occlusions (*e.g.*, narrow streets in urban areas), shadows (*e.g.*, close to tall buildings), and homogeneous textures (*e.g.*, special roof covers). The potential of the generated DSMs for automatic extraction of LoD-1 and LoD-2 building models was studied and compared towards the usage of aerial images and LiDAR data. The extraction of LoD-1 building models from the obtained DSMs of different sensors and the help of topographic maps has shown comparable and reliable results. In contrast, the generation of LoD-2 models lacked in accuracy for the satellite-borne DSM, due to a higher level of noise, resulting in means of the residual values in height up to 2 m and only 40 % of correctly modeled buildings. Wang and Frahm (2017) proposed a multi-view stereo (MVS) matching approach for satellite images considering technical challenges at different stages of satellite-MVS, such as radiometric changes, inaccurate sensor calibration, and excessive correspondence search space. Reliable 2D feature matches are used to compensate for extrinsic calibration errors, and an edge-aware interpolation of the sparse feature matches generates a set of dense feature matches in a fast and contour-preserving manner. Experiments on WorldView-3 images with 30 cm GSD and LiDAR ground truth has shown an improvement in the completeness of the derived 3D point cloud up to 80% with a registration accuracy of 0.25 m horizontally and 2.57 m vertically. A comparison towards other dense matching algorithms, such as SiftFlow (Liu et al., 2011), S2P (De Franchis et al., 2014), PMBP (Wang et al., 2016a) and SGM (Hirschmuller, 2005) demonstrated the superior performance of the proposed method. The enhanced 3D point cloud could further improve processing steps such as the generation of DSMs and, thus, building extraction and reconstruction. Another approach for refining satellite-borne stereo image DSMs for the task of LoD-2 building model generation was proposed by Bittner et al. (2019) by formulating

Table 2.3: Different ways for as-built building reconstruction, categorized into different platforms, observability of building parts, reconstruction accuracy, resolution and attainable LoD level. Samples of current related literature are given in the last column

Platform	Approach	Observability	Accuracy	Resolution	LoD	Literature
Satellite	Single image	Roof	200–300 cm	30–100 cm	1-2	Wang and Frahm (2017)
	Stereo images	Roof	200–300 cm	30–100 cm	1-2	Duan and Lafarge (2016)
Airborne	LiDAR	Roof	10–30 cm	1–10 cm	2	Elberink and Vosselman (2011); Song et al. (2015)
	Photogrammetry	Roof (& façade)	40–80 cm	5–30 cm	2(-3)	Frommholz et al. (2015); Haala et al. (2015)
Terrestrial	LiDAR	Façade	5–10 cm	1–5 cm	(3)	Pu and Vosselman (2009)
	Photogrammetry	Façade	5–10 cm	1–5 cm	(3)	Schindler and Bauer (2003)
	MMS	Façade	10–50 cm	5–50 cm	(3)	Xiao et al. (2009)
UAV	LiDAR	Roof	10–100 cm	5–10 cm	2	Chiang et al. (2017)
	Photogrammetry	Roof & façade	1–5 cm	1–10 cm	2-3	Aicardi et al. (2016); Murtiyoso et al. (2017); Vacca et al. (2017)

a multi-task learning problem consisting of semantic segmentation and building roof model generation. They utilized a conditional generative adversarial network (cGAN) with an objective function based on least-square residuals and an auxiliary term based on normal vectors for further roof surface refinement. Experiments with WorldView-1 images on a test site revealed geometric accuracies of 3.1 m RMSE, 64 % intersection over union (IoU) and 55 % recall. Duan and Lafarge (2016) have shown that jointly utilizing geometric and semantic cues brings robustness to occlusions and low image quality. Object shapes can be well preserved by including a region-based stereo matching strategy which resulted in faithful LoD-1 representations of buildings from 50 cm resolution satellite stereo imagery of QuickBird-2, WorldView-2 and Pléiades. However, the low resolution of the images impedes the generation of LoD-2 models, and already small buildings (*e.g.*, houses in residential areas) and poorly textured and reflective objects are challenging for deriving accurate LoD-1 models. A statistical approach was proposed by Partovi et al. (2015) as a hybrid method comprised of a top-down and bottom-up strategy. Building roof components are extracted and classified as pitched and flat roofs in a bottom-up approach and afterward fitted towards a satellite-borne DSM in a top-down approach via a Monte Carlo Markov chain (MCMC) with simulated annealing to optimize roof parameters iteratively. Experiments have demonstrated the potential of the method on WorldView-2 imagery, but have also revealed problems for complex building shapes and robust classification of roof types. A deep learning-based approach for performing parametric reconstruction models on single-view satellite images was presented by Wang and Frahm (2017). WorldView-3 satellite images, LiDAR data and geo-registered GIS vector maps served as training data for a single-shot detection

CNN to localize building instances and fit parametric models simultaneously from a single satellite image, formulated as a 3D object detection task. The combination of building detection and reconstruction eliminates the need for topographic maps for building footprint extraction, and the ability to reconstruct building models from single-view images is a promising direction for autonomous, fast and robust urban planning on a large scale. However, the approach could only extract LoD-1 models with a mean average precision of 50% and a RMSE in the height estimation of more than 2.7 m.

With the recent development of both satellite sensors and algorithms in the fields of photogrammetry and computer vision, global and large-scale 3D city models can be derived to a highly automated degree, which is already used in urban planning and city monitoring. However, resulting building models can not yet exceed LoD-2, as they only capture nadir views, and even a robust generation of faithful LoD-2 models is challenging for small and complex building shapes. Even with the availability of VHR satellite imagery, the spatial resolution is too low for recovering local details of the buildings. At the same time, images often feature a high degree of noise, which also depends on the weather conditions during data acquisition. Future satellites, such as Cartosat-3, announce lower GSDs of 0.25 m, suggesting further improvements in the development of large scale reconstructions from satellite imagery. The accurate positioning of the satellites allows for direct georeferencing within a few meters, which is sufficient for large-scale investigations but is not accurate enough for detailed urban planning. The utilization of GCPs is still required for more precise georeferencing of satellite-based geo-products.

Airborne systems have become the most prevalent technique for deriving medium-scale urban 3D models in an automated manner. Due to their relatively low altitude, manned aircraft and helicopters can recover a much higher degree of detail than satellites, although manual and costly flights have to be executed. The large payload allows for carrying high-resolution observation sensors and highly accurate localization sensors used for precise direct georeferencing of the derived data. Usually, mapping sensors utilize LiDAR and large format digital cameras, both featuring individual characteristics. While airborne laser scanning (ALS) directly provides highly accurate and dense 3D point clouds, photogrammetrically derived point clouds have to be triangulated from multiple overlapping images of the scene. Subsequently, building reconstruction methods from ALS point clouds can rely on accurate geometric 3D data and mostly focus on the extraction and modeling of building models. Photogrammetric techniques, on the other hand, must pay attention to the generation of dense and noise-free 3D point clouds. The subsequent building modeling task often relies on similar techniques to those used for ALS point clouds. The following sections briefly describe current advances in both research directions.

Airborne laser scanning (ALS) enables the acquisition of accurate and dense point clouds in unsurpassed quality in terms of homogeneous distribution and high precision of derived 3D points of few centimeters. However, as summarized in Elberink and Vosselman (2011), systematic and stochastic errors may occur in the obtained measurements especially in urban areas, as well as inconsistent and relatively low point cloud densities, and data gaps due to occlusions by adjacent objects (*e.g.*, clutter of trees, absorption of the laser pulse by water features and reflections from windows on the roof). Nevertheless, compared to airborne photogrammetric approaches, particularly the uniform distribution, high accuracy, and low degree of clutter of ALS point clouds provide an optimal basis to model complex building structures. As a necessary step towards the reconstruction of 3D building models, the dense point cloud has to be classified so that building points can be distinguished from other points. Methods for this task usually rely on stochastic graphical models (*e.g.*, MCMC (Yang et al., 2013), conditional random field (CRF) (Lafarge and Mallet, 2012)), machine learning techniques (Özdemir and Remondino, 2019), or on the availability of accurately georeferenced building footprints (Kada and McKinley, 2009). Once buildings are extracted, the subsequent task of explicit building modeling is accomplished either by model-driven or data-driven approaches. Former make use of a pre-defined geometric library of various roof and building type shapes and select a suitable parametric instance of this library for each building point and refine the corresponding model parameters by minimizing deviations of the measured LiDAR points towards the 3D shape (Arefi et al., 2008; Huang et al., 2013; Kada and McKinley, 2009). Although model-driven methods are robust towards noise and yield geometric correct building models, they are limited in the flexibility of modeling complex building shapes, since they rely on pre-defined shapes from a library. Data-driven methods decompose the building into individual segments for which parametric planar or curved shapes are fitted towards the point cloud, which are finally connected considering topological relations between the components. The segmentation of individual building structures is typically addressed by plane detection and extraction using random sampling consensus (RANSAC) (Verma et al., 2006) or Hough transform (Vosselman and Dijkman, 2001), by performing a clustering analysis (Vosselman, 1999), using region growing (Xu et al., 2017b) or via contour extraction (Song et al., 2015). Topological relationships between the segmented components are established via graph-based methods (Verma et al., 2006), minimization of intersecting vertices (Matei et al., 2008) or boundary regularization (Sampath and Shan, 2010). Model-free approaches benefit from the relaxation on pre-defined building type libraries and are therefore more flexible for modeling complex building shapes. However, since they purely rely on the acquired 3D points, they are highly sensitive towards noise, outliers and gaps causing geometrically incorrect building models.

Despite these problems, current model-driven and data-driven approaches are already mature in processing large-scale areas in an automated fashion achieving accurate building models up to LoD-2. Due to the pure geometric nature of ALS and the absence of 3D points on building façades caused by the acquisition geometry, the resulting 3D building models lack in geometric and textural details which are required for the generation of LoD-3 models. To overcome these limitations, other ways of reconstructing 3D buildings need to be considered.

Photogrammetric approaches, in comparison to ALS, can not directly provide 3D information, due to its passive nature. Therefore, great attention must be paid towards flight planning for acquiring highly overlapping and matchable images. After data acquisition, the task of building modeling consists of first detecting and extracting buildings by using, for instance, GIS maps or semantic segmentation and second by obtaining 3D information from the images and final modeling of the extracted buildings. First aerial systems only captured overlapping nadir-directed images for photogrammetric methods on generating 3D models. Due to the large degree of noise and low image quality in the early stage, the generation of LoD-2 models was a challenging task. Early aerial photogrammetric methods focused on the robust detection of corresponding corner points (Fischer et al., 1997) and building outlines (Suveg and Vosselman, 2000; 2002) in stereo images with the aid of GIS maps while fitting predefined building shapes via hypothesis generation and validation. These model-based methods could create purely geometric and simplified LoD-2 building models. However, modern manned aircraft are integrated with multiple synchronized high-resolution cameras featuring both nadir and slightly oblique images. The development of MVS methodologies facilitated the generation of accurate and dense DSMs, which has also affected the task of urban modeling. These systems have not only increased the robustness of generating 3D building models but also allowed to capture façade elements which can be integrated into the reconstruction process for generating textured 3D building models up to LoD-3 (Dahlke et al., 2015; Frommholz et al., 2015; Haala and Rothermel, 2015; Haala and Kada, 2010; Moe et al., 2016; Remondino and Gerke, 2015; Toschi et al., 2017; Zhu et al., 2018a). Based on dense 3D point clouds or surface meshes, the generation of 3D building models corresponds to that of ALS outputs, revealing reconstruction approaches with parametric shapes, based on semantic segmentation or via DSM simplification. Nevertheless, differences exist in the quality of the obtained 3D data and new challenges arise. Photogrammetric 3D point clouds usually suffer from a a considerable degree of noise and inconsistent point density, which dramatically reduce in the presence of poorly textured object surfaces. Moreover, difficulties of dense matching of oblique imagery arise from large perspective distortions, large scale variations due to a higher depth of field, illumination changes and multiple occlusions. Modern computer vision methods, however, have proven to provide reliable 3D information enabling detailed cadastral applications (Ostrowski, 2016). Unlike the generation of 2.5D raster DSMs, which is sufficient for the creation of LoD-2 building models, LoD-3 models require full 3D information. This aggravates the geometric elevation processing, since filtering and meshing have to be performed in the full 3D space rather than in 2D space, which often results in smudged and inaccurate depth transitions around object boundaries. To compensate for these effects, Holzmann et al. (2016) presented a method for regularizing noisy 3D building models by integrating detected lines in the images into a cell labeling optimization. For the extraction of buildings in the images and their separation from the surroundings, modern CNN-based semantic segmentation approaches can be applied, which have proven their applicability in the domain of aerial imagery (Marmanis et al., 2016). Additionally, oblique images enable texturing roof and façade elements of the extracted building models yielding more realistic and visually appealing 3D building models.

A semi-automatic method for 3D city model generation in accordance with CityGML standards was proposed by Buyukdemircioglu et al. (2018) using large-format digital aerial images and vector basemaps. The method generates textured LoD-2 building models with a top-down approach for roof modeling with a predefined library. Experiments were conducted in the city of Chesme Town in Turkey, consisting of 43000 buildings. The automated method could successfully reconstruct 73 % of the buildings on a 10 cm DSM, however, the reconstruction still included erroneous building models due to a limited roof library and occlusions from adjacent trees and other objects. Blaha et al. (2016) integrated semantic segmentation into a 3D reconstruction framework aiming at densely reconstructing both 3D shape and segmentation for semantic object classes. By jointly reasoning about shape and class, class-specific shape priors (*e.g.*, walls and roofs) led to improved reconstruction results reported by overall accuracies between 80–90 %, whereby a hierarchical refinement alleviated an increase of the runtime and reduction of memory size helping to scale the method for large areas. Another work carried out by Chen et al. (2018) has shown, that even structure from motion (SfM) and semantic segmentation can benefit from each other by utilizing semantic information to boost the accuracy of feature point matching by assigning each feature point a corresponding label and include these labels to an equality constrained bundle adjustment. The results yielded improvements in processing speed and a semantically-enriched 3D reconstruction model.

Terrestrial systems utilize close-range images or laser scans and record information of building façades that hardly reach the buildings' roof. These approaches make use of static data acquisition from terrestrial laser scanning (TLS) or utilize street-level images captured in a dynamic automated fashion, *e.g.*, via MMS. Therefore, they perfectly augment aerial reconstructions which suffer from weakly reconstructed building façades (Haala and Kada, 2010). However, although MMS can generate highly detailed façade models, no information on the buildings backsides can be obtained, and even for static terrestrial acquisitions, numerous overlapping images have to be carried out from different locations. A subsequent registration is necessary to retrieve complete reconstruction models. Pu and Vosselman (2009) proposed an approach for the automatic reconstruction of building models from terrestrial LiDAR data. Knowledge-based feature constraints are defined to extract walls, doors, roofs, windows, protrusions and intrusions from raw laser point clouds while introducing assumptions for occluded parts. The resulting polyhedron models consist of detailed façade elements, however, the methodology is construed for simple building shapes. A photogrammetric model-based approach for automatically recovering detailed building models from images was presented by Schindler and Bauer (2003). After reconstructing a dense point cloud, a coarse building model consisting of principal planes is recovered via linear regression. Smaller features, such as indentations and protrusions, are detected, and for each element, the most suitable template from a model library is selected and refined in cooperation with the images. As a result, geometrically parametric building models could be derived. However, the amount of feature sets that can be modeled is limited, and occlusions, as well as extreme lighting conditions like weak contrast, reflections, and shadows, affected the reconstruction quality. Lee and Nevatia (2004) presented a method for reconstructing 3D windows frames from a single calibrated rectified ground view image. 2D windows are extracted by exploiting regularity and symmetry, while the classification of various window classes helps for deriving the 3D depth of the windows. A method for generating street-side 3D photo-realistic models from image sequences of an MMS was presented in the work of Xiao et al. (2009). A multi-view semantic segmentation method segments each image at pixel level into meaningful areas, such as building, sky, ground, vegetation, and car, while buildings are separated into independent blocks, for which a facçade model is constructed through regularizing noisy and missing 3D data by an inverse patch-based orthographic composition and structure analysis method. Another MMS method made use of panoramic image sequences and developed an MVS method enforcing piece-wise planarity constraints (Micusik and Kosecka, 2009). They presented a novel depth fusion method by exploiting the constraints of urban environments while combining the advantages of volumetric and viewpoint-based fusion methods. The partial piecewise-planar models of the scene were fused into one textured triangle surface mesh. The resulting models could not recover local depth transitions of the façades and often lacked in inaccurate local planar patches. Moreover, trees were typically reconstructed as planar surfaces, and moving cars occluded substantial parts of the road, causing gaps in the final model. By setting up topological relationships between extracted main surface patches, an improvement in recovering the building's shape structure can be achieved, which was demonstrated in the work of Tian et al. (2010). Both 3D points and 3D edges are combined to extract surface path outlines that are connected to neighboring patches. However, the results have shown that the accuracy of surface patches, as well as the building outline detection still needs to be improved, and detailed depth transitions could not be recovered.

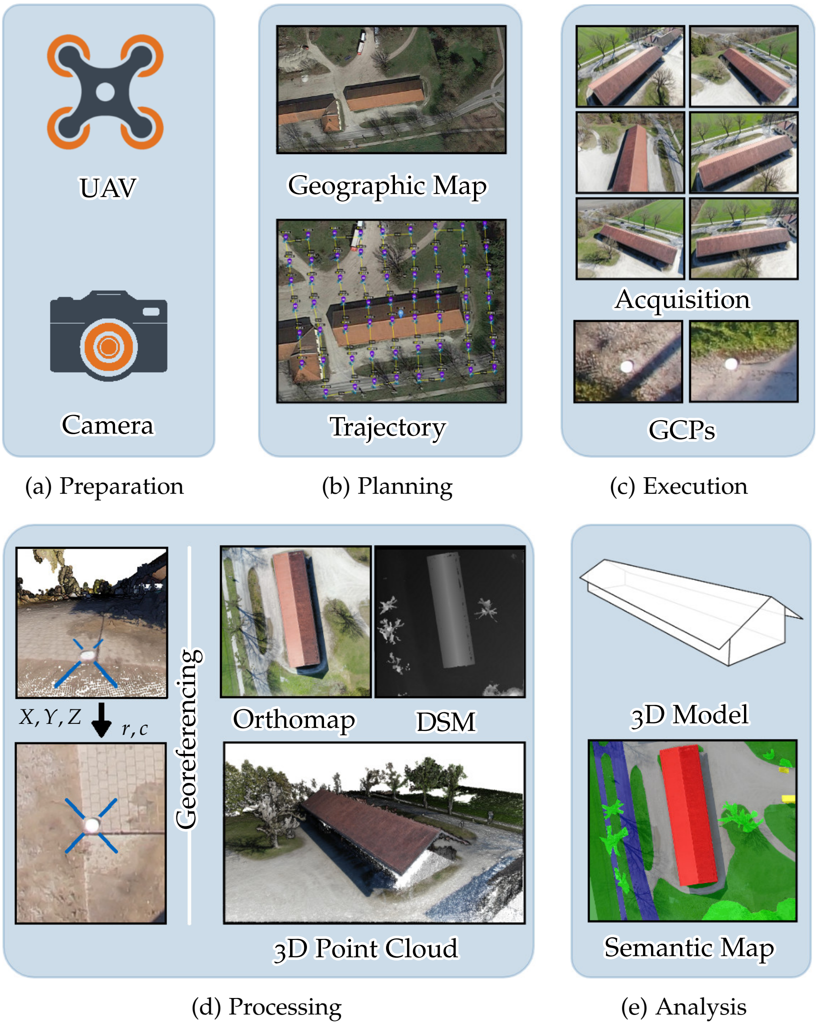
UAV-BORNE photogrammetry finally bridges the gap between airborne and terrestrial data acquisition. Especially, multi-rotor drones stand out for their flexibility in hovering and navigating to any free airspace acquiring images with arbitrary perspectives from various positions. Although modern light-weight LiDAR sensors can be mounted on UAVs, cameras have proven to be able for generating 3D information in a comparable quality towards LiDAR-based 3D point clouds in terms of density and accuracy (Aicardi et al., 2016b; Caroti et al., 2015; Themistocleous et al., 2016; Vacca et al., 2017; Wefelscheid et al., 2011). This development was strongly influenced by the recent advances of 2D and 3D computer vision algorithms which were integrated into entire 3D reconstruction pipelines, such as Bundler (Bundler), VisualSfM (Wu), Colmap (Schönberger and Frahm, 2016), Aigsoft (Agisoft), Pix4D (Pix4Da) or Micmac (Rupnik et al., 2017). These frameworks focused on the generation of 3D point clouds and meshes from unordered image sequences with relatively large perspective changes and already led to numerous researches on investigating the possibility of using UAV imagery for 3D mapping and 3D building reconstruction (Aicardi et al., 2016b; Caroti et al., 2015; Jarzabek-Rychard and Karpina, 2016; Malihi et al., 2018; Murtiyoso et al., 2017; Murtiyoso and Grussenmeyer, 2017; Themistocleous et al., 2016; Vacca et al., 2017; Wefelscheid et al., 2011; Wu et al., 2018). An analysis of using nadir UAV images for automatically generating LoD-2 building models has revealed a comparable quality and accuracy towards ALS-based methods (Jarzabek-Rychard and Karpina, 2016). The possibility of acquiring and utilizing both nadir and oblique images in the reconstruction process does not only increase the completeness of the 3D model but also improves the geometric accuracy of the entire model (Vacca et al., 2017). A pipeline as a whole for generating LoD-3 building models and building footprints from dense point clouds derived from UAV imagery was presented in the work of Malihi et al. (2018). The method addresses clutter and unwanted sections and decomposes the point cloud into several smaller parts, which are divided into potential planar segments of façades, roofs, or grounds. These geometric primitives are subsequently fused to generate building models. Experiments yielded accuracies up to 0.25 m for roof parts and 0.2 m for façade elements. In an experimental study on reconstructing complex historical buildings, Murtiyoso et al. (2017) studied the acquisition and processing protocols for UAV photogrammetry in terms of camera calibration, flight planning, and data management. They have shown that different available 3D reconstruction pipelines could generate promising 3D models. Still, they outlined that calibration parameters of small-scale camera sensors tend to be unstable and highlighted the necessity of prior camera calibration and self-calibration in the bundle adjustment. Moreover, dense matching of poorly textured façade elements often resulted in gaps and clutter in the reconstructed model. These and other works have shown that capturing both nadir and oblique close-up images from UAVs enables the generation of high-resolution, textured, and complete LoD-3 building models. However, the success of utilizing UAVs for detailed image-based 3D reconstructions depends on several essential aspects. A general workflow for UAV-photogrammetry, as shown in Figure 2.6, highlights the major steps that need to be considered in a photogrammetric UAV campaign.

A *preparation* stage, as shown in Figure 2.6a) involves the selection of a compatible vehicle and sensor among the versatility of currently available camera-equipped UAV systems. This choice encompasses considerations on vehicle size, flight time, intended GSD, and the expected environment of the campaign area. While tightly built-up urban environments potentially tend to offer only limited and narrow free airspace, small-scale vehicles with wide-angle lenses are preferred, enabling sufficient coverage and overlap even from views with short distances towards the targeted building. Conversely, wide and open spaces permit enlarged accessible airspace and thus safer trajectories in higher flight altitudes, which may allow the use of cameras with increased focal lengths offering improved image quality.

The further *planning* step consists of the determination of a suitable UAV path and image acquisition viewpoints allowing the generation of satisfying 3D models (*cf.* Figure 2.6b). In terms of generating sophisticated LoD-3 building models, commonly used simple grid-like or circular patterns are usually not sufficient since surrounding obstacles and overhangs might occlude parts of buildings. Therefore, a precise trajectory must be planned, featuring views from various altitudes and directions that cover the entire building. Generating optimal and safe UAV paths requires accurate and up-to-date topographic maps of the whole environment as a basis for the planning step, which still is often a manual and time-consuming task even for experienced pilots. Without precise elevation data or by the use of outdated maps, flight planning is exposed to a vast risk of collision.

On-site *execution*, as depicted in Figure 2.6c, embraces the acquisition of images according to the flight plan and groundwork for a subsequent georeferencing task. Most available UAVs are capable of navigating along a pre-designed trajectory and capture images from determined viewpoints without human interaction. As already introduced in Section 2.1, UAVs are generally equipped with low-quality localization sensors impeding direct georeferencing of the acquired imagery. In order to obtain accurately georeferenced photogrammetric products in the range of few centimeters, the deployment of GCPs is indispensable, which, however, is a manual and elaborated process including planning, ground-based surveying, and localization of the GCPs in the images. This task even deteriorates in case of inaccessible or dangerous areas.

After a successful operation, the acquired UAV images are *processed* for obtaining georeferenced photogrammetric geo-products, such as orthomosaics, DSMs, and dense 3D point clouds (*cf.* Figure 2.6d). Pixel coordinates of the GCPs in the images are assigned with precise terrestrial 3D measurements and subsequently exploited to transform the local 3D model into a subordinate reference system. Current state-of-the-art SfM and MVS methodologies are capable of integrating GCP coordinates and generating highly accurate photogrammetric geo-products.



(a) Preparation (b) Planning (c) Execution

(d) Processing (e) Analysis

Figure 2.6: General framework of UAV photogrammetry. Based on the selection of a compatible vehicle and camera sensor (a), suitable image acquisition viewpoints are determined with respect to an available topographic map of the campaign site (b). On-site image acquisition, as well as deployment and measurement of GCPs (c). 3D reconstruction and geo-product generation using 3D reconstruction pipelines and precise georeferencing through integration of GCPs coordinates into pose estimation (d). Analysis and further processing of the derived geo-products (*e.g.*, semantic 2D and 3D maps, 3D modeling, etc.) (e)

Subsequent *analysis* methods highly depend on the application and include 3D modeling and interpretation methods, such as the generation of 3D building models, semantic maps and class extractions, as exemplarily shown in Figure 2.6e.

In addition to the exclusive use of UAVs for 3D building reconstruction, several works attempted to further combine and integrate terrestrial measurements into the modeling process. Caroti et al. (2015) demonstrated a potential integration of UAV-photogrammetry and TLS for reconstructing a church, including its interior. UAV and LiDAR point clouds were registered with the help of GCPs with an accuracy of 1 cm RMSE. Another multi-modal reconstruction approach integrated LiDAR and oblique UAV point clouds for the generation of LoD-3 building models (Wen et al., 2019) by extracting plane features from LiDAR point clouds with accurate boundary constraints obtained from the oblique images. Another work that integrated oblique UAV and terrestrial images has shown an optimization of 3D modeling in urban areas which does not only improve the performance of integrated 3D modeling but also may solve the problem of GNSS tracking loss for MMS in urban areas (Wu et al., 2018). The results of the proposed image matching procedure exhibited accurate and highly-detailed 3D building models integrated into a superordinate reference system.

Although the presented works exposed the ability of using UAV images for generating textured and highly-accurate and detailed LoD-3 building models, the utilization of UAVs for building reconstruction has its own challenges. First, the small size and low payload impede the integration of high-quality positioning systems that could be used for direct georeferencing. Although a bundle adjustment can integrate synchronized low-quality GNSS coordinates for all acquired images, allowing for georeferencing accuracy of several meters, applications, which require more precise georeferenced photogrammetric products, such as DSMs, orthomosaics, or full 3D models, have to rely on the use of accurately surveyed GCPs. Second, the quality of generated 3D models is profoundly affected by the quality of the preceding UAV flight plan. Especially for the combination of both nadir and oblique views capturing roof and façades, large parallax angles might result in a failure of image matching methods. A major drawback of current UAVs is their relatively short operational time, requiring thoughtful flight plans for acquiring sophisticated images for the demanded task. Lastly, due to the rising popularity of using UAVs, extensive regulations on their use have been and still are being adopted. This complicates the use of UAVs, especially in tightly built-up urban scenarios, and flight plans can become challenging in terms of safety regularization on the accessible airspace.

2.3.2 *LoD-4 Building Models and Remote Sensing Data*

3D models of indoor environments are increasingly gaining importance due to the wide range of applications they can be subjected to: from automated floor plan generation, visualization, and 3D redesign to monitoring, simulation, and navigation. With the combination of accurate exterior building models, such as LoD-3, an extension to LoD-4 allows to assist seamless navigation and location-based services from street level to specific building locations. Public buildings, such as airports or shopping malls, could highly benefit from joint indoor and outdoor models for efficient navigation and to plan appropriate evacuation routes (Nagel et al., 2010). Although remote sensing-based building reconstruction can generate accurate and detailed LoD-3 models, as demonstrated in Section 2.3.1, the further step of enriching the model with its interior is beyond the scope of remote sensing. Other ground-based platforms are required for generating indoor models; however, remote sensing is vital for the integration into an outdoor model. Since indoor models are likely generated independently from the outdoor model, the registration of both models is crucial for accurate alignment and, therefore, the generation of LoD-4 models. Nevertheless, temporal and modal differences and the lack of visual overlap between both models pose significant challenges for an accurate registration, which has scarcely been the subject of current research.

The generation of visually appealing indoor models is an independent, extensive, and long-term field of research, particularly characterized by narrow and complex spaces, heavy occlusions from obstacles, and homogeneous surfaces. The following section presents some of the latest developments in this field, which refer to the use of TLS, RGB-D, and RGB images. A comparison of various off-the-shelf indoor scanning systems was elaborated by Lehtola et al. (2017).

Terrestrial Laser Scanning produces exceptionally dense and accurate point clouds independent of the surface texture due to its active nature. However, specular surfaces, such as windows, mirrors, and varnished objects, cause a high level of clutter, requiring sophisticated filtering techniques. Decisive disadvantages of TLS are high purchase costs and large sensor sizes, as well as the requirement of numerous and time-consuming scans from multiple positions. Various methods focused on reducing necessary scans by continuously updating a captured 3D map and estimating the next best scan positions that facilitate a successful registration of individual scans but avoiding redundancies in the derived data (Frías et al., 2019; Kriegel et al., 2015). After data acquisition, the unstructured 3D points are filtered, clustered into planar surface segments, merged to individual rooms, and finally completed to floor plans and complete 3D models of the entire building. To identify topological relations in the raw 3D point cloud, variations of RANSAC enable the detection of planar clusters (Ochmann et al., 2016; Previtali et al., 2014). Vectorization and piecewise-linear partitioning generate candidates for walls from detected vertical planes and constraints related to typical building geometries, such as the prevalence of straight lines and orthogonal intersections are included in the merging process of individual planar surfaces (Macher et al., 2017). Besides concentrating on estimating geometric room layouts, volumetric hybrid methods generate distinct meshes for representing the permanent structure of the building while persevering fine details of interior objects, such as furniture, in a separate model (Turner and Zakhor, 2015).

RGB-D cameras, such as Microsoft Kinect[4], Intel RealSense[5], Google Tango[4], or Occipital Structure Sensor[6] are low-priced, compact, and lightweight depth sensors and constitute a popular alternative towards TLS. Technologies for deriving depth values are based on structured light or time-of-flight, yielding color and depth images in reasonable resolutions with real-time frame rates. Although they are limited in range (up to 10 m) and field of view, which complicates the reconstruction of large-scale rooms and are sensitive towards clutter, the high frame rate allows for online RGB and depth map acquisition. Methods concentrate on the registration of

[4] The development and manufacturing of the device has been ceased
[5] https://www.intelrealsense.com/
[6] https://structure.io/

succeeding data frames (Kerl et al., 2013), efficient noise removal (Yan et al., 2018), and incremental map updates (Liu et al., 2017). As a result of the limited field of view, comparable long RGB-D image sequences have to be acquired in particular for large rooms, which can lead to erroneous 3D maps due to accumulated drifts in frame-to-frame registrations. Sophisticated loop detection schemes can help to reduce this effect (Whelan et al., 2015). The obtained depth maps from registered image sequences are fused into 3D point clouds and further processed to triangulated polyhedral models, while Manhattan-world assumptions of indoor scenes assist in generating floor plans or semantic 3D indoor models (Chen et al., 2015; Choi et al., 2015b; Ikehata et al., 2015).

Image-based indoor modeling from RGB cameras encounters multiple challenges. First, due to the requirement of overlapping images with large parallax angles for deriving accurate depth maps, interiors mostly do not exhibit the necessary space for suitable acquisition viewpoints. Consequently, narrow parallax angles and ego-motion acquisitions lead to high uncertainty in the depth maps and, thus, triangulated 3D points. Secondly, the triangulation of corresponding 2D pixels requires distinct texture, which is often not guaranteed through predominantly homogeneous surfaces in interiors. Therefore, massive gaps can characterize the resulting 3D models, even when using state-of-the-art MVS methods. A technique for deriving dense depth maps from stereo images of poorly textured indoor scenarios was proposed by Furukawa et al. (2009) by assigning a set of candidate planes obtained from a raw dense point cloud to each pixel, posed as a Markov random field (MRF) and solved via graph-cuts. However, the quality of the derived watertight 3D model highly depends on the preceding dense point cloud, which itself can often not be computed if the camera pose estimation, such as SfM, fails. Though, it has been shown that indoor environments often feature numerous distinct edges, which led researchers to address camera pose estimation and 3D modeling by detecting lines instead of points (Holzmann et al., 2016a). Another branch of research focuses on estimating planar primitives or entire room layouts from single views using machine learning-based methods and Manhattan world assumptions (Fouhey et al., 2013; Hedau et al., 2009). Beyond that, recent advances in deep learning have led to investigations in estimating dense and detailed depth maps from single views (Eigen and Fergus, 2015; Eigen et al., 2014; Laina et al., 2016; Liu et al., 2018; Liu et al., 2016). These methods could vastly improve indoor models derived from image-based methods since they yield pixel-wise depth maps even in the presence of poorly textured surfaces. Therefore, walls, ceilings, and floors could be entirely and accurately reconstructed without performing time-consuming and erroneous MVS. Although this field has recently aroused tremendous interest computer vision, the research is at an early stage, and the applicability of these methods in the field of indoor reconstruction still needs to be assessed.

2.4 SUMMARY

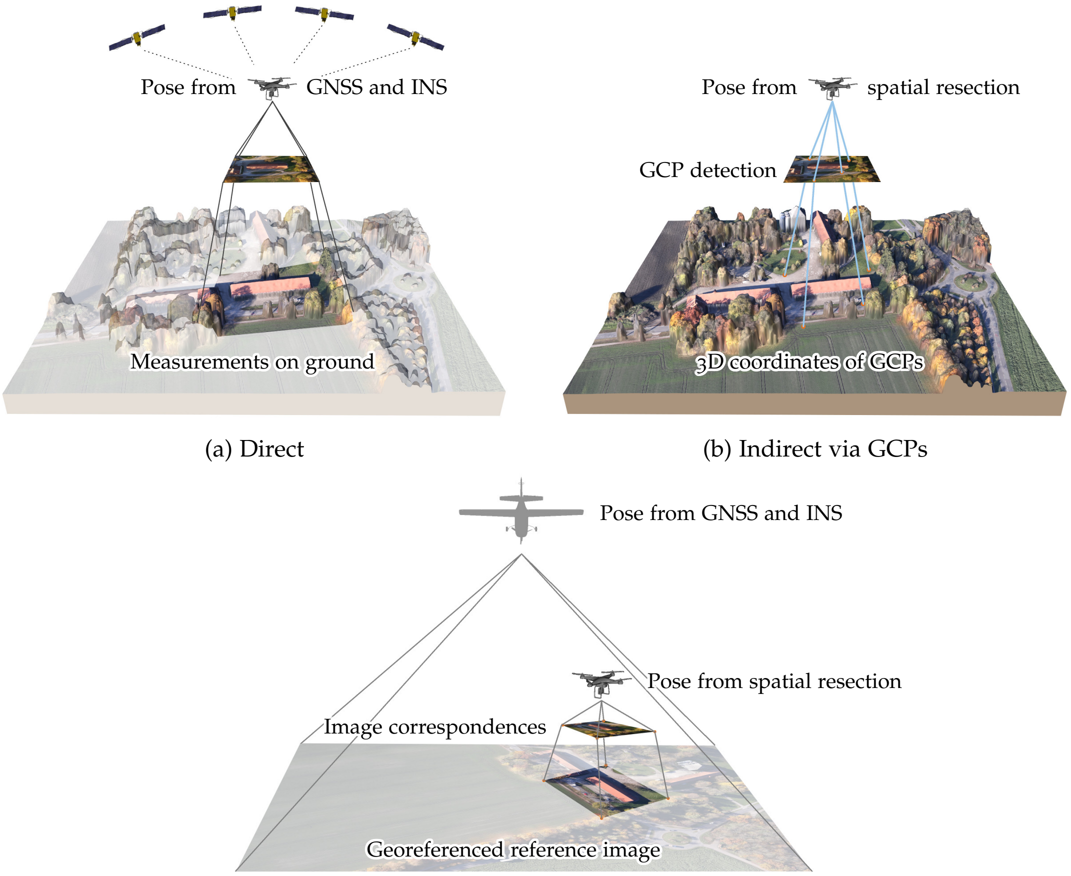
With the progress of UAV systems, camera sensors, and 3D vision algorithms, UAVs have become an essential component in remote sensing, witnessing increasing popularity for an expanding field of applications. While previously experts and expensive manned aircraft have exclusively accomplished high precision aerial photogrammetry, the maturity of UAV developments and cost efficiency have facilitated automation of airborne surveying even for non-experts. This development has contributed to the increasing use of UAVs for various small and medium-scale applications, which would have been too costly with established remote sensing methods. Facilitating access to remote sensing for non-specialists offers a cost-effective alternative to manned aerial photogrammetry and expands the portfolio of remote sensing by providing new possibilities and application fields.

3 STATE-OF-THE-ART IN THE USE OF UAV IMAGERY FOR BUILDING RECONSTRUCTION

Based on the workflow in Figure 2.6, a photogrammetric unmanned aerial vehicle (UAV) campaign consists of the individual steps of *preparation*, *planning*, *execution*, *processing*, and *analysis*. This chapter describes the state-of-the-art in UAV-photogrammetry related to some of the tasks addressed in this thesis, including *planning*, *processing*, and *analysis*. Besides accurate georeferencing described inSection 3.1, adequate flight planning (*cf.* Section 3.2) is crucial for the quality of the generated geo-products. Since these steps are indispensable for any remote sensing-related application requiring UAV images, they are not exclusively relevant for the task of building reconstruction. With the facilitation of generating level of detail (LoD)-3 building models using modern computer vision methods based on a carefully designed flight plan, georeferencing of the images furthermore enables accurate positioning of the generated model in a spatial reference system and the integration with other spatial geo-data. With the further step of generating LoD-4 models, separately produced indoor models must be accurately aligned with the exterior model (*cf.* Section 3.3), requiring eligible *analysis* methodologies. With the recent success of using deep learning-based methods to estimate dense depth maps from single views, as will be described in Section 3.4, a simplification in the derivation of 3D structures independent from multiple multi-view stereo (MVS)-related views, could facilitate some of the tasks associated with this work that require 3D information. Besides easing UAV flight planning based on a prior 3D proxy model generated from a single view, complete indoor models could be generated solely from RGB images which do not have to meet the strict requirements of classical MVS image geometry. A summary of the development, as well as current methodologies for these topics, are presented.

3.1 GEOREFERENCING OF UAV IMAGERY

Besides the derivation of georeferenced photogrammetric products from UAV imagery, accurate image geo-registration serves as a prerequisite for joint information extraction for multi-scale earth observation, enabling a seamless fusion of satellite, aerial and UAV-borne remote sensing data for multi-scale observation and monitoring of the environment. In terms of UAV image registration, different methodologies are discussed, beginning from the capability of direct georeferencing using onboard navigation sensors (*cf.* Figure 3.1a). Concerning higher accuracy, practical considerations suggest indirect methods, such as aerial triangulation (AT), the utilization of ground control points (GCPs) (*cf.* Figure 3.1b), or image-based registration with already georeferenced image data (*cf.* Figure 3.1c).



(a) Direct

(b) Indirect via GCPs

(c) Indirect via image matching

Figure 3.1: Different methods for georeferencing of UAV imagery. Direct georeferencing via global navigation satellite system (GNSS) and inertial navigation system (INS) (a), indirect georeferencing via GCPs (b) and via image matching with georeferenced reference images (c)

3.1.1 *Direct Georeferencing*

In the field of aerial photogrammetry, manned aircraft have access to high-end GNSS and INS localization sensors, allowing direct georeferencing of the images in the range of several centimeters without the need of external GCPs or photogrammetric adjustments in a post-processing step (Kurz et al., 2014). Due to payload limitations, many commercial UAVs are usually equipped with lightweight and low-quality localization sensors that affect the achievable localization accuracies. Accuracies of better than 5 m horizontal and 10 m vertical (Chiang et al., 2012; Padró et al., 2019; Verhoeven et al., 2013) can hardly be achieved, which might be sufficient for image archiving and subsequent retrieval, but may not be sufficient for photogrammetric applications or combination with other spatial data. An investigation regarding the ability of direct georeferencing with UAV systems shows that the geolocalization accuracy of current UAV systems is still too low to perform direct photogrammetry applications at a very large scale (Chiabrando et al., 2013). To compensate for the particularly critical vertical component of GNSS measurements, Turner et al. (2012) showed that including barometer altitude observations may improve vertical geo-registration accuracy. Although recent real-time kinematic (RTK)-enabled UAVs have improved the georeferencing accuracy towards decimeter range by connecting the onboard GNSS receiver to a base station or a virtual reference system (VRS), leading to comparable accuracies towards airborne and satellite-borne systems, the additional costs of acquiring RTK-based GNSS sensors might be too high for the reachable absolute accuracy (Forlani et al., 2018; Padró et al., 2019). However, if the position of the UAV is held for several minutes, accuracies increase towards centimeter-level (Turner et al., 2013). In case of monitoring or surveillance applications that capture images or video streams from the same location, such systems are convenient. However, photogrammetric applications require almost constant movement of the UAV, while short operation times of UAVs prevent from numerous stationary measurements.

3.1.2 *Indirect Georeferencing*

Aerotriangulation (AT) offers another approach to compensate for the inadequate accuracy of onboard UAV localization sensors. The imprecise GNSS and inertial measurement unit (IMU) measurements for each acquired image are introduced as initial approximates for the exterior orientation parameters and optimized in the bundle adjustment of the entire overlapping image block (Nex and Remondino, 2014). This integration helps to improve the georeferencing accuracy despite of the low accuracy of IMU measurements and is relatively robust against outliers and missing measurements caused by sensor outages. Although the estimation of orientation parameters highly benefits from this integration, the global accuracy of the localization parameters can not exceed the GNSS measurements' average accuracy.

Ground Control Points (GCPs) are indispensable when aiming at reliability and accuracy and are therefore the most established technique for achieving an unprecedented georeferencing accuracy which is even recommended when utilizing high-end devices on aerial and satellite imagery due to the existence of systematic errors (Cramer, 2001; Poli et al., 2015). This technique has been widely applied in numerous UAV-based researches approving georeferencing accuracies of few centimeters, which even improves the utilization of expensive RTK-based systems (Gerke and Przybilla, 2016). Various works investigated the effect of the configuration of GCPs for indirect georeferencing (Agüera-Vega et al., 2017; Ai et al., 2015; Rumpler et al., 2014; Turner et al., 2012). They agreed on the importance of an even distribution within the entire survey area, however, an increase in the number of used GCPs will eventually not further improve the georeferencing accuracy. Although indirect georeferencing with GCPs can obtain the highest accuracy, the deployment is often expensive, requires fieldwork operations, and is unpractical or even infeasible in hazardous or inaccessible areas. Based on experiences from UAV field campaigns, Nex and Remondino (2014) claimed that GCP field measurements absorb at least 15 % of the toal campaign duration.

Image Registration with Reference Image Data presents a promising alternative for geo-registration of UAV imagery due to the growing accessibility of accurately georeferenced high resolution aerial and satellite imagery. Image registration is the task of defining image transformations between multiple overlapping images captured either from various platforms, times, viewpoints, or a combination thereof. In order to estimate geometrical transformations (*e.g.*, affine, similarity, homography, projective), corresponding parts between a target and reference image have to be assigned in terms of image regions or single points. Speaking of UAV image georegistration, already georeferenced image data, such as aerial or satellite images, serve as reference images, whereas spatially imprecise or even entirely unknown UAV target images have to be aligned towards these reference images. This matching task, often formulated as wide baseline stereo matching, is characterized by large differences in image scale, baseline, orientation, and temporal changes. The attainable georeferencing accuracy depends first on the absolute georeferencing accuracy of the reference images and second on the registration accuracy of the image matching. Existing methods for this task can be generally divided into two categories, which are known as area-based and feature-based matching methods.

Area-based image matching methods compute a similarity measure between image patches in two frames by comparing intensity values using a specific matching metric, such as normalized cross-correlation (NCC), mutual information, or Fourier cross power spectrum. The comparison is usually conducted by a sliding window approach with a pre-defined window size of the image patch, resulting in high computational cost for high-resolution images. After pixel-wise comparison of the target image with patches of the reference image yielding individual matching costs for each specific location in the reference image, the final correspondence is found as the matching pair with minimum cost. For speeding up the correspondence estimation, the matching can be conducted in a hierarchical scheme for various image scales, starting from down-sampled images and refining the matching for a higher resolution in the spatial neighborhood of the local minimum of the prior matching result. Area-based matching methods are applicable for short baselines and small ratios between baseline and scene depth, which usually leads to small radiometric and geometric differences between the images. In such cases, the transformation between the images is usually only a small translation without scale and rotation differences. These requirements are usually not given for UAV image georeferencing. UAV and multi-modal reference images often exhibit large baselines, projective distortions, rotations, tilted views, radiometric changes between the used sensors and topographic changes of the scene. Since area-based matching methods compare intensity values between rectangular image patches of equal sizes in a pixel-wise manner, the image pairs have to be pre-aligned for eliminating rotation and scale differences. Although the matching can be extended for rotation and scaling by augmenting the search window, the complexity of the matching increases linearly for each discrete augmentation. Moreover, area-based methods do not consider any structural analysis and therefore are sensitive to intensity changes introduced by noise, varying illumination, and radiometric differences. In contrast, image patches representing homogeneous areas without any distinct details will likely result in ambiguous matches with high similarity to numerous similar smooth patches in the reference image. Besides the pre-alignment of the matching image pairs and similar radiometric properties, the size of the search window is crucial for the matching performance and hard to select appropriately. While small window sizes result in more accurate matching results, the spatial extent of the captured image content is limited, leading to ambiguous matching costs and failure in poorly textured areas.

Although enlarging the windows size yields increased image content and more unique image structures, projective distortions and inaccurate matching results can lead to poor registration accuracy.

While feature-based matching algorithms have proven to be more accurate and robust than area-based methods, there are several works based on area-based matching methods for the task of UAV image georeferencing. Lin et al. (2007) applied mutual information as similarity metric for estimating a homography between a UAV and reference Google Earth image depicting the same area. Due to the requirement of a planar scene and eliminating scale and rotation differences, these strong constraints limit the utilization of this method to very specific applications, and the choice of a suitable window size highly depends on the topography and can not be set intuitively. Based on two experiments, this registration technique achieved accuracies of 3–6 px towards the reference satellite images, representing a ground resolution of several meters. Conte and Doherty (2009) proposed a similar approach utilizing NCC as similarity metric, motivating their approach for visual localization of UAVs in case of GNSS outages. To compensate for geometric distortions, Fan et al. (2010) proposed a deformable template matching approach combining edge and entropy features. They used this method to register nadir-view UAV imagery and high-resolution satellite imagery with known scale differences. Experiments were only conducted on one real-world scene, and the registration accuracy was only visually presented without a quantitative assessment. Although the obtained registration appeared accurate, the presented correlation map has barely shown a distinct maximum, indicating that ambiguous peaks in other scenarios could lead to misalignments. Karel et al. (2014) utilized NCC-based matching of perspectively transformed image patches for an automatic georeferencing of archaeological UAV images. Although a pre-alignment was conducted in terms of the same image scale and orientation towards the reference image, the authors reported spurious matches in homogeneous areas, requiring a subsequent filtering strategy within the bundle adjustment by detecting false matching hypotheses with the aid of a digital surface model (DSM).

Summarizing, area-based matching possibly achieve a rough registration, however, the achievable accuracy is far from the desired pixel or even sub-pixel accuracy. Numerous works have confirmed the necessity of strong prior knowledge for a pre-alignment of the images. However, the larger the time interval between the acquired images, the more likely geometric and radiometric differences will occur that can hardly be resolved with area-based methods. Additionally, area-based methods are vulnerable to low textures, illumination changes, and repetitive image structures, limiting the general applicability in a wide range of different scenarios. In this sense, robust image matching against large scale and viewpoint changes, as well as robustness against radiometric and local geometric differences is the key to successful UAV image georeferencing. Feature-based methods are widely considered capable of overcoming these challenges and have been investigated for this task as well.

Feature-based image matching methods rather detect and describe highly discriminative local features instead of pixel-wise intensity comparisons of image patches across the entire image. They locate specific interest points (feature extraction) and then furnish them with quantitative information (feature descriptors) for re-identification in target images via feature matching. The local features should be characterized by their locality, making them robust to occlusion and clutter, distinctiveness for re-identification in other images, efficiency, and generality to exploit different types of feature points in different situations. The overall goal of feature-based methods is to achieve geometric and photometric invariance as far as possible. The former focuses on the recognition of identical object points in reference images that have undergone geometric transformations, such as translation, rotation, scale, or either projective transformations. Latter should ensure the identification even for different illumination settings, such as shadows, brightness, and exposure. Among numerous feature-based methods, SIFT (Lowe, 2004), SURF (Bay et al., 2008), ORB (Rublee et al., 2011), KAZE (Alcantarilla et al., 2012) and its variants have become the most widely used hand-crafted feature-based methods.

Scale-invariant feature transform (SIFT) (Lowe, 2004) stands out for its robust scale, orientation, and illumination invariant property, making it one of the most widely used feature-based methods. Based on a scale-space, consisting of convolutions with Gaussian kernels of different sizes for different octaves, keypoints candidates are localized by detecting extrema of the Difference of Gaussians (DoG), which are further refined by eliminating low contrast points and performing a sub-pixel refinement. Subsequently, keypoint orientations are assigned to all localized keypoints based on local image gradients. Lastly, a descriptor generator computes the local image descriptor for each keypoint based on the distribution of image gradient magnitudes and orientations in the neighborhood of the keypoint as a histogram over local-oriented gradients and stores the bins into a compact 128 element vector. As a variant of SIFT, a full affine invariant matching framework A-SIFT (Yu and Morel, 2011) was proposed to handle large differences in viewpoints by simulating a series of transformed images to cover the entire affine space. In the case of matching images with large differences in viewpoints, A-SIFT offers a more robust performance than SIFT, which was also confirmed in the evaluation presented in Apollonio et al. (2014).

Inspired by SIFT's performance, speeded up robust features (SURF) (Bay et al., 2008) have been established as a faster variant of SIFT with comparable matching performance. It approximates the DoG with box filters, fastening up the convolution step with the help of integral images. The detector uses an integer approximation of the determinant of the Hessian blob detector and uses wavelet responses in both horizontal and vertical directions by applying adequate Gaussian weights for orientation assignment. These wavelet responses are also used for the descriptor, by first dividing the neighborhood around each keypoint into subregions, and second, taking wavelet responses as representations for each subregion. The sign of the Laplacian distinguishes bright blobs on dark backgrounds from the reverse case, which allows for faster matching of solely equally signed features.

Oriented FAST and rotated BRIEF (ORB) (Rublee et al., 2011) is a fusion of the FAST keypoint detector (Rosten and Drummond, 2006) and the BRIEF descriptor (Calonder et al., 2012) with some modifications. ORB aims to provide another fast and efficient alternative to SIFT in terms of matching performance but allowing for real-time applications even on low-power devices. After keypoint detection using FAST, Harris corner measure is applied to find top n points in a multi-scale image pyramid. In order to overcome the absence of keypoint orientations from FAST, intensity weighted centroids of each patch are estimated, and the directions from the center of the patch towards the centroid define the keypoint orientations. Moreover, moments are introduced to improve the rotation invariance. A rotation matrix enhances the rotation invariance of the BRIEF descriptor, by steering the descriptors according to the patches' orientations.

KAZE features (Alcantarilla et al., 2012) substitute the building or approximation of Gaussian scale-spaces used in most other approaches by building non-linear scale-spaces using non-linear diffusion filtering alongside additive operator splitting. Herewith, they tackle the problem of vanishing local image features from the convolution with Gaussian kernels, resulting in reduced localization accuracy and distinctiveness. By detecting keypoints in a non-linear scale-space, the proposed algorithm reduces noise but retains the object boundaries. The authors claimed that KAZE features succeed especially in presence of deformable objects while being slightly more computationally demanding than SURF. As an adaption of KAZE, accelerated-KAZE (A-KAZE) reduces the computational complexity by introducing Fast Explicit Diffusion for feature extraction yielding comparable matching performance than KAZE while dramatically reducing the computational time in the range of ORB features.

Although many variants and alternatives have been developed, numerous investigations demonstrated that, despite the comparable high computational complexity, SIFT is still more robust to viewpoint changes and common image disturbances and yields the most accurate matching performance (Bekele et al., 2013; Calonder et al., 2012; Dwarakanath et al., 2012; Heinly et al., 2012; Juan and Gwun, 2009; Tareen and Saleem, 2018).

Most recent times have witnessed a shift from using hand-crafted feature-based approaches towards leveraging deep learning-based architectures, such as learned invariant feature transform (LIFT) (Yi et al., 2016), LF-Net (Ono et al., 2018) or the approach of Altwaijry et al. (2016). These works have shown that feature extraction can be learned as part of an end-to-end pipeline with neural networks to detect and describe meaningful keypoints. However, Schönberger et al. (2017) carried out an experimental evaluation of learned and advanced hand-crafted feature descriptors and demonstrated that hand-crafted features still perform on par or better than learned features in the practical context of image-based reconstruction. Learned feature extractors still show a high variance across different datasets and applications, evidencing that more training data is still needed to develop generalized feature extractors.

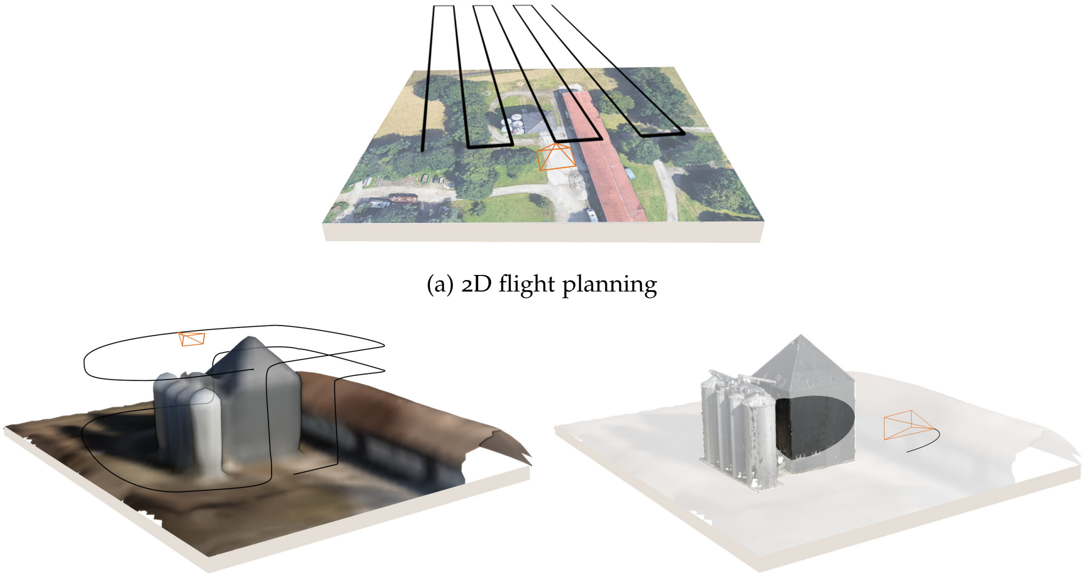
Various attempts have been made for georeferencing UAV images with feature-based methods. A recent review about vision-based UAV image localization by Xu et al. (2018) pointed out current methods and remaining challenges of this task. One of the most severe difficulties is the presence of diverse view angles of acquired images due to the platform's high dynamics during acceleration, deceleration, direction changes, flips, shakes and gimbal steering. Additionally, it is very challenging to find a sophisticated descriptor that is adaptive to various surroundings, ranging from urban environments with dense and repetitive textures up to sparse rural areas. Moreover, scale-invariant feature descriptors might fail for varying object sizes in UAV images according to the object size, flight altitude, and focal length. The following introduces the most recent feature-based methods for the task of UAV georeferencing. Aicardi et al. (2016) adopted an approach for co-registering multi-temporal UAV image datasets. Georeferenced images are chosen from a reference epoch, whereas images from stable areas act as anchor images. Images from subsequent epochs are matched towards the anchor images with SIFT features, while exterior orientation parameters from the anchor images are integrated into the global bundle adjustment with the new images. This procedure yielded registration accuracies of nearly 1 px between different epochs, even in the presence of noticeable changes in the scene. However, it only estimated the relative transformation between the epochs, while the epoch's absolute transformation still relied on the georeferencing accuracy of the reference epoch, achieved, for instance, by the deployment of GCPs. Onyango et al. (2017) made use of A-KAZE features to register oblique UAV images to oblique aerial images. They pointed out the necessity of eliminating scale differences between the image pairs by adjusting the octaves in the feature extraction step towards the same ground sampling distance (GSD). Multiple homography hypotheses are computed from the putative matches in an iterative manner and a projective transformation is estimated from the homography with the maximum number of inliers. They evaluated their method on urban images from the ISPRS Dortmund benchmark (Nex et al., 2015) reporting residuals of respective fundamental matrices in the range of 2.5–5.0 px, which corresponds to 3–5 cm GSD of the UAV. Although several images achieved good matching results, a couple of failure cases were presented. An evaluation with other datasets and a comparison towards other feature-based image matching methods were not reported. Yuan et al. (2019) used deep features for registering UAV and satellite images from GoogleMaps. A corresponding satellite image patch of the same area is automatically derived from the coarse localization of the UAV. Absolute registration accuracy in the range of 60–250 cm can be achieved with the proposed method and a comparison towards SIFT revealed an improved matching accuracy, indicating a more discriminative property of deep features. Tsai and Lin (2017) proposed an image registration scheme of UAV images and historical aerial images with a novel accelerated-BRISK algorithm. Unlike other approaches, the matching is performed for orthomosaics generated from UAV image sequences. An adaption of the BRISK feature extractor includes a sorting ring for analyzing spatial relationships between the descriptor pairs found in the two input images to remove false matches. The method outperformed SIFT and reported an acceleration towards SIFT and BRISK by 19 times and 5 times, respectively. The achieved positioning accuracy ranges from 20 cm to 1 m for reference images with a GSD of 25 cm. However, UAV images of the test site were captured in a large altitude with a GSD of 13 cm, resulting in relatively low scale differences between UAV and reference images. A performance analysis of the method for larger scale differences was not demonstrated. Nassar et al. (2018) integrated semantic cues for improving image-based UAV geo-localization. UAV image sequences are registered to each other while the registration towards satellite reference maps is enhanced by a novel semantic shape matching that performs registration by matching semantically segmented object shapes such as roads or buildings. Results demonstrated that using shape and contextual information provides improved geo-localization than relying solely on local features. However, experiments were only carried out for two datasets from the work of Nex et al. (2015), and the results have shown that the method lacks accuracy for scenes with dense building blocks due to the tendency of huge blob representation after semantic segmentation complicating the matching step. Reported accuracies range between 3–5 m, allowing sufficient online geo-localization capabilities for navigation without GNSS, but lacks an accurate georeferencing of photogrammetric geo-products. Challenges of matching UAV images and aerial reference images were also reported in Karel et al. (2014), claiming that SIFT did not succeed in correctly matching the cross-modal images from an archaeological scene due to large differences in shape, texture, and illumination. Instead, an area-based matching scheme with NCC as cost function was utilized; however, a reduction of the search space could be conducted by exploiting prior knowledge from the image acquisition parameters.

Another branch of UAV geo-localization approaches focuses on matching UAV imagery with available geo-tagged terrestrial images, which results in extremely large viewpoints changes. Addressing this task by utilizing A-SIFT feature matching yielded meter-level registration accuracies, however, a preliminary selection of street views images and low altitude oblique UAV images facing the building façades had to be conducted in advance (Majdik et al., 2015). To overcome large viewpoint changes, Shan et al. (2014) synthesized aerial views from pre-aligned Google Street View images using depth maps and corresponding camera poses. By warping ground-level images into target aerial views, a registration with UAV images utilizing SIFT matching was facilitated. However, the availability of a ground-level MVS reconstruction, a consistent appearance in both aerial and ground imagery, and the absence of severe occlusions are necessary for the proposed approach. A new feature representation for learning a rough geo-localization of terrestrial images with the help of aerial images was proposed by the work of Lin et al. (2015). Given a ground-level query image and a reference database of aerial imagery, a heat map of its potential geo-localization can be obtained by the proposed Where-CNN. Although this method could also be used vice versa for the geo-localization of UAV images with the help of geo-tagged terrestrial images, manual interventions were still needed to estimate the scale differences between aerial and terrestrial queries. Furthermore, an absolute orientation of the query image could not be resolved by the method. Zamir and Shah (2014) developed a multiple nearest neighbor feature matching method using Generalized Minimum Clique Graphs (GMCP). SIFT features are extracted, and multiple neighbors from the reference dataset are retrieved. The consistency among global features is enforced by selecting a correct single nearest neighbor using GMCP. An evaluation was conducted by matching unconstrained user images from photo-sharing platforms against a multitude of street view images. Localization errors within several meters were reported, while an accurate absolute geo-registration was not proven and investigated in the paper.

In summary, feature-based methods generally yield better results in georeferencing UAV imagery, although this task still remains an open problem. None of the presented methods was able to meet all decisive requirements, particularly accuracy, robustness, and applicability in varying environments. Although work has already been invested in pre-processing and approximating multi-modal imagery, as well as in investigating different feature-based methods, a standard matching procedure has mostly been applied for the registration process. Detailed investigations regarding the failure of feature-based methods and the consideration of the differences between multi-modal remote sensing images were lacking. A careful integration of these aspects could help to unleash the full potential of feature descriptors, and thus improve the image geo-registration.

3.2 UAV PATH PLANNING FOR PHOTOGRAMMETRIC IMAGE ACQUISITION

The generation of photogrammetric geo-products, such as orthomosaics, DSMs, and entire 3D models from UAV imagery, requires an accurately designed flight plan for the image acquisition process, highly influencing the quality of the resulting products. Concerning 2.5D mapping tasks of comparatively large-scale areas, the flight plan-



(a) 2D flight planning

(b) Model-based 3D flight planning

(c) Model-free 3D flight planning

Figure 3.2: Schematics of different UAV flight planning techniques. Simple geometric 2D flight patterns based on a topographic map allow for fast trajectory planning but likely fail to recover vertical surfaces (a). Automated model-based 3D flight planning methods (b) rely on a proxy 3D map of the environment and estimate an optimal trajectory for coverage maximization in an offline manner. Online-capable model-free 3D flight techniques (c) do not rely on a proxy 3D model but estimate the next-best-views based on a continuously updated 3D map by merging incoming measurements

ning can be reduced to simple geometric flight patterns with nadir-directed views from safe flight altitudes (*cf.* Figure 3.2a). Basic geometric dependencies between camera intrinsics and flight altitude allow for customized image acquisition for an intended spatial resolution of the derived photogrammetric products. The generation of high-resolution and seamless 3D models, however, requires close-up and oblique views, which is particularly demanding for designing a proper flight plan in tightly built and inhabited environments. For this purpose, optimization methods on automatic flight planning have been developed for UAV-based 3D mapping tasks. These methods are either based on a very coarse proxy model of the environment (*cf.* Figure 3.2b) or operate in completely unknown environments (*cf.* Figure 3.2c). The following sections introduce current flight planning methods, from established planning tools for 2.5D mapping tasks in Section 3.2.1 to automated model-free (*cf.* Section 3.2.2) and model-based (*cf.* Section 3.2.3) 3D flight planning methods. An overview of different automated path planning methodologies is given in Table 3.1.

3.2.1 *Basics and Practical Realization*

To a very large extent, the success of a UAV campaign depends on the preceding flight planning. Most available UAV systems come with waypoint navigation technologies, allowing to fly autonomously along pre-designed trajectories defined by a set of waypoints specifying the location and orientation of the vehicle and equipped gimbal. Current UAV "mission planners", such as PrecisionHawk (Precisionhawk), Pix4D

Table 3.1: Overview of automatic approaches for UAV path planning using different sensors and scene representations

Approach	Sensor	Scene Representation	Literature
Model-free	Laser-Scanner	3D occupancy grid	Nuske et al. (2015); Yoder and Scherer (2016)
		Mesh	Kriegel et al. (2015)
	RGB-D Camera	3D occupancy grid	Heng et al. (2011); Hepp et al. (2018); Meng et al. (2017); Michael et al. (2012); Sturm et al. (2013)
	Camera	3D occupancy grid	Mendez et al. (2017); Palazzolo and Stachniss (2018); Stumberg et al. (2016)
		Density representation	Border et al. (2018)
Model-based	Camera	3D occupancy grid	Alsadik et al. (2013); Hepp et al. (2018); Roberts et al. (2017); Smith et al. (2018)
		Mesh	Bircher et al. (2016); Hoppe et al. (2012); Jing et al. (2016); Peng and Isler (2019)

Capture (Pix4Db), DJI Flight Planner (DJI) or ArduPilot (ArduPilot) facilitate UAV flight plans as simple geometric patterns, such as regular grids or circular flights with respect to a desired GSD. By defining the mission area, relevant geometric parameters, camera intrinsics, and the intended spatial resolution, waypoint positions for the UAV to be flown automatically, are provided by such tools. These waypoints include exterior orientation parameters of suitable image acquisition positions that ensure the derivation of geo-products in the desired GSD, which is inverse proportional to the flight altitude. In order to derive photogrammetric products, images have to be acquired with a sufficient degree of overlap. According to many investigations, overlaps of at least 70–80 % should be maintained and even increased for challenging areas, such as poorly textured surfaces or uneven terrains (Nex and Remondino, 2014). Detailed explanations on the geometric relationships for photogrammetric flight planning are provided in Albertz and Wiggenhagen (2009).

Assuming a flat terrain and a nadir-directed camera, the required altitude h (in m)

$$h = \frac{\mathrm{GSD} \cdot \mathrm{f} \cdot \mathrm{I_w}}{\mathrm{s_w}} \tag{3.1}$$

to achieve the intended GSD (in $\frac{\mathrm{m}}{\mathrm{px}}$) is defined by the focal length f (in mm) and the proportion of the image width $\mathrm{I_w}$ (in px) and sensor width $\mathrm{s_w}$ (in mm) of the calibrated camera. Clearly, the implicit assumption of a flat terrain does not hold for many situations featuring large elevation changes in the captured environment. Knowledge of the surface topography, particularly an accurate and up-to-date DSM, is mandatory for adjusting the flight altitude to achieve a consistent GSD for the entire area. However, different flight altitudes also affect the horizontal and vertical ground coverage $d_{\mathrm{h,v}}$ of the images

$$d_{\mathrm{h}} = \mathrm{GSD} \cdot \mathrm{I_w} \tag{3.2}$$

$$d_{\text{v}} = \text{GSD} \cdot \text{I}_{\text{h}}, \tag{3.3}$$

and thus the overlap of successive views. The baselines $b_{\text{forward,side}}$ between spatially adjacent images in accordance with the intended overlaps, comprised of end lap o_{end} and side lap o_{side}, are given by

$$b_{\text{forward}} = d_{\text{h}} \left(1 - \frac{\text{o}_{\text{end}}}{100}\right) \tag{3.4}$$

$$b_{\text{side}} = d_{\text{v}} \left(1 - \frac{\text{o}_{\text{side}}}{100}\right) \tag{3.5}$$

In case of a tilted camera with off-nadir angle θ, the GSD must be corrected by $\cos(\theta)^{-1}$. After computing the relevant geometric parameters, a regular grid of viewpoints can be arranged in accordance with the required flight altitude and baselines between adjacent viewpoints in both dimensions. Major problems of off-the-shelf planners comprise the exclusion of the underlying 3D geometry such as an accurate DSM, the availability and the reliance on recent topographic maps or satellite images, and in particular, the ability to calculate solely simple 2D trajectories. In the absence of accurate and up-to-date DSMs, complex trajectories encompassing close-up views from various altitudes may prove fatal due to a collision with the surrounding environment. A similarly pronounced risk is associated with the use of outdated or poorly georeferenced planning maps. Even in spacious rural areas, newly built high ranging power lines can pose a danger of collision.

The aforementioned off-the-shelf planners are adequate 2.5D mapping tools for spacious and flat terrains without obstacles (Nex and Remondino, 2014), but encounter their limits for precise 3D modeling in uneven, densely built or heavily vegetated environments. Since no 3D model of the environment is taken into consideration, the obtained trajectories either do not cover every part of the object of interest due to visual occlusions or may even cause an accident with an adjacent obstacle. For that purpose, automated 3D flight planning methodologies have been developed to overcome the outlined problems.

3.2.2 *Automated Model-free Methods*

Automated model-free methods do not rely on any prior knowledge of the scene and solve an *exploration* task by iteratively selecting the most favorable view to refine the explored model based on a current view with new measurements. Therefore, they can be carried out immediately on-site without prior planning since they generate a current model of the observed environment in an online manner. This incremental scene modeling and viewpoint planning is commonly known as next-best-view (NBV) planning, which is already a long-standing part of research in Robotics. The methods alternately fuse incoming measurements from a new viewpoint into the reconstruction of the scene and estimate novel viewpoints to incrementally increase the information about the object or the surrounding environment. Utilized sensors for these measurements include laser scanners (Kriegel et al., 2015; Nuske et al., 2015; Yoder and Scherer, 2016), RGB-depth (RGB-D) sensors (Fan et al., 2016; Heng et al., 2011; Hepp et al., 2018a; Meng et al., 2017; Michael et al., 2012; Sturm et al., 2013) and cameras (Border et al., 2018; Kumar Ramakrishnan and Grauman, 2018; Mendez et al., 2017; Palazzolo and Stachniss, 2018; Stumberg et al., 2016). In the work of Nuske et al. (2015), an autonomously navigating UAV explored arbitrarily shaped river courses while avoiding overhanging trees. 3D information from a rotating 3D laser scanner is continuously merged to update a volumetric representation of the entire environment. For each position, the frontier to the unseen river course is detected, while a collision-free traversal path towards a suitable position close to this frontier is computed. In order to avoid building up a computational expensive dense 3D map of the environment, Shen et al. (2012) proposed an efficient particle-based frontier method that represents known and unknown space through sparse samples. Vision-based navigation and exploration from monocular cameras require a semi-dense reconstruction of the environment in real-time. Unlike the usage of RGB-D cameras or laser scanners, which have access to direct 3D measurements from the current position simplifying the update process of the model, vision-based approaches are usually hard to implement, since the generation of depth maps requires multiple views and significant onboard processing power or at least a wireless connection to the ground-station for data transmission in order to merge incoming measurements with the current model. Additionally, selecting the next best views in accordance with MVS requirements on the fly — in particular maintaining sufficient baselines and parallax angles of adjacent views — is a challenging task since the actual mapped free airspace might be very limited. Stumberg et al. (2016) applied LSD-SLAM (Engel et al., 2014) for both motion tracking of the camera and model generation. Based on these estimates, a 3D occupancy grid is built which is further used for planning obstacle-free exploration maneuvers to unobserved regions. Experiments have shown the capability of safely exploring unknown indoor scenes, the applicability in outdoor scenarios, however, was not presented. Due to short operating times of UAVs and the unconstrained goal of exploring unknown environments, the task of a safe return of the UAV is highly relevant in practice typically but usually not contemplated by such methods. Nevertheless, Palazzolo and Stachniss (2018) presented an approach that takes into account the cost of reaching a new viewpoint in terms of distance and predictability of the flight path for a human observer and, finally selects a path that reduces the risk of crashes when the expected battery life comes to an end, while still maximizing the information gain during the return flight. The computation of the information gain is based on an uncertainty reduction through expected changes in the entropy from multiple measurements. Since exploration approaches usually rely on image streams with high-frame rates for self-localization and mapping, an enormous amount of images have to be managed when a high-resolution 3D model of the scene should be obtained in a post-processing step. For that purpose, Mendez et al. (2017) proposed the Scenic Route Planner, encompassing a collaborative behavior that allows the camera to switch between acting as independent structure from motion (SfM) agents or as a variable baseline stereo pair for MVS. While SfM images are required in a high frame rate for self-localization — especially in GNSS-denied environments — the MVS agent suggests a highly reduced number of viewpoints that are used for a high-quality dense reconstruction of the scene. Experiments have revealed promising 3D models, obtained from an autonomously navigating robot with a vastly reduced amount of images. However, experiments were only conducted in indoor environments, and the incorporation of 6-degrees of freedom (DoF) camera poses was ignored, since only the yaw angle was considered in the camera orientation estimation. Border et al. (2018) presented the Surface Edge Explorer, which is completely scene-model free in contrast to many other NBV planning methods. While others mostly rely on volumetric (*e.g.*, voxel grids) or surface (*e.g.*, triangulated meshes) representations, they proposed a density representation to detect and explore observed surface boundaries. Therefore, the complexity only scales with the number of measurements and not the scene scale, making it suitable even for large-scale reconstructions. Additionally, due to the absence of a scene-model, no non-intuitive parameters have to be tuned, as this is the case with surface representations. Experiments have shown more accurate and complete 3D models than other approaches, but with a reduced amount of viewpoints. However, only small-scaled object models were used in a controlled environment which exhibited free accessibility in the entire surrounding area. The applicability in realistic, tightly built-up, and large-scale environments with UAVs was not demonstrated.

Kumar Ramakrishnan and Grauman (2018) introduced sidekick policy learning as a preparatory learning phase that attempts simplified versions of the eventual exploration task, then guides the agent via reward shaping or initial policy supervision. However, the approach does not consider the surrounding environment, so that potential novel viewpoint predictions might not be located in free and accessible airspace.

Summarizing, model-free NBV approaches have the advantage of being capable of exploring novel environments without prior knowledge of the scene, allowing efficient on-site operation by skipping a preceding planning stage. However, since they require to iteratively fuse new measurements in real-time to update the 3D map creation, strong hardware requirements have to be met. Especially for the generation of large-scale 3D maps, a continuously expanding 3D model increases the complexity of merging new measurements with the model. Although many of these methods face the exploration of complex indoor scenes by extending the observed frontiers, they are, in principle, also applicable to outdoor scenes using UAVs as mapping agents. However, the approaches concentrate on gaining information about the certainty of occupied and free airspaces for estimating new viewpoints which maximizes the observability of unseen space, rather than estimating meaningful viewpoints for maximizing the reconstructability of the observed scene, which, in the sense of photogrammetry, requires complex viewing configurations to derive precise depth maps. For the generation of 3D building models in realistic environments, the task of view planning has to incorporate the surrounding environment, which leads to two challenges. First, the generation of 3D building models requires the agent to focus on a targeted object instead of exploring the entire scene, though, the sensor placement of new viewpoints has to guarantee accessible airspaces by recovering the surrounding as well. Therefore, the objectives are two-fold: exploring the environment in the close neighborhood of the target building, and second, estimating viewpoints pointing towards this building. Considering the resulting need for collision avoidance, either additional sensor equipment is needed to capture the adjacent neighborhood of the targeted object, or complex maneuvers are required to capture the entire surrounding. The latter challenge implies that the agent should never lose track of the targeted building, even in case of immense appearance changes when observing the object from different perspectives. Although current instance segmentation approaches have seen substantial improvements due to the rise of deep learning-based methods (Garcia-Garcia et al., 2017), it is not guaranteed that the agent always focuses on the indented building. Alternatively, identifying the intended object can be conducted in 3D space by transferring the semantic label maps into the fused 3D map. The consideration of a complex environment has not been addressed in methodologies that focused on exploration tasks for detailed 3D reconstruction since all of these methods relied on an entirely free accessible airspace around the object. Moreover, the obtained trajectories are not optimal in the sense of obtaining short paths that guarantee complete coverage of the object until the end of the flight, since the complete extent of the targeted object is unknown in advance. Apart from a high-resolution 3D reconstruction, the obtained 3D model's completeness is crucial for photogrammetric reconstruction tasks. Concerning flight safety, trajectories obtained by such online methods tend to be rather unpredictable, complicating the required observability of the UAV for the pilot, while hazardous flights above other buildings, facilities, streets, and populated areas are unlikely to be avoided. In order to realize such restrictions, flight planning has to consider the semantics of the entire environment, however, since model-free methods do not have access to the entire scene layout in advance, constraints on the accessibility of the airspace are difficult to implement in real-time.

3.2.3 *Automated Model-based Methods*

In contrast to model-free *exploration* methods that focus on autonomy and real-time capability in unknown environments, model-based path planning algorithms make use of an available proxy model of the environment and focus on estimating a subsequent optimal path to maximize the coverage and accuracy of the object in a global manner (Hepp et al., 2018b; Hoppe et al., 2012; Jing et al., 2016; Peng and Isler, 2019; Roberts et al., 2017; Smith et al., 2018). By including a proxy model, many of the challenges that occurred in model-free-based methods, which were pointed out in the previous section, are obsolete. The accessibility of information about free and occupied airspaces in the planning phase enables optimization tasks to focus on the objective of a suitable viewpoint selection for complete and high-quality reconstructions and on the estimation of optimal and short paths to combine these viewpoints. On the other hand — contrary to active modeling — these *explore-and-exploit* methods do not receive any feedback from the acquired images during the *exploitation* flight, and thus feedback about the obtained reconstruction quality. This leads to the importance of carefully designing useful heuristics being used for the generation of the refinement path. The global optimization of coverage and accuracy, however, usually leads to larger completeness and smoother trajectories compared to model-free methods. Since the computation of the flight paths is done in an offline manner, it can be conducted on a separate hardware device which eliminates the need for powerful onboard computations on the UAV. This prevents such approaches from being fully autonomous, however, the execution of the optimized *exploitation* paths can be easily conducted on any type of UAV by simply navigating along a waypoint file obtained from the optimization result. The necessity of a proxy model of the environment might be a critical factor for many tasks, however, in the sense of UAV-photogrammetry, an approximation of the scene is either available through a current DSM or can be easily and quickly obtained from a preceding safe overview flight using off-the-shelf planners, such as those introduced in Section 3.2.1. The acquired overview images can be processed with current and fast 3D reconstruction pipelines yielding sufficient coarse proxy models. Alternatively, modern computer vision methods, such as those that will be introduced in Section 3.4, allow for deriving a dense depth map from a single-view, which could even further ease the process of generating an approximate 3D representation of the environment. The representation of the prior model is either based on an existing map with height information (Jing et al., 2016), expressed by a set of discrete 3D points in a voxel space (Alsadik et al., 2013; Hepp et al., 2018b; Roberts et al., 2017; Smith et al., 2018) or by volumetric surfaces (Bircher et al., 2016; Hoppe et al., 2012; Jing et al., 2016; Peng and Isler, 2019).

In order to define appropriate views for the optimized trajectory, a multitude of camera viewpoint hypotheses are either regularly sampled in the free 3D airspace (Roberts et al., 2017; Smith et al., 2018) resulting in 3D camera graphs, or are sparsely sampled in a 2D view manifold (Peng and Isler, 2019) or skeleton sets (Snavely et al., 2008) around the object. The subsequent optimization task selects a subset of these viewpoint hypotheses suitable for generating detailed 3D reconstructions and finds a feasible and short path through the camera graph. Alternatively, the locations of the regularly sampled viewpoint candidates can be continuously refined during the optimization (Hepp et al., 2018b).

As a means of assessing the suitability of camera viewpoints for the reconstruction, heuristics are usually defined considering the necessities for a successful SfM and MVS workflow. Although Fraser (1984) already pointed out in an early stage that the task of suitable camera viewpoint configurations for image-based 3D reconstructions is an ill-suited problem due to its high degree of non-linearity and multi-modality, approximations of multi-view requirements and strong constraints were developed to find adequate and practically feasible solutions in the optimization. An exhaustive amount of work addressed the problem of selecting the best views from a large amount of different views hypotheses (Furukawa and Hernández, 2015; Furukawa et al., 2010; Goesele et al., 2007; Rumpler et al., 2011; Snavely et al., 2006). These works highlighted the key parameters that influence the reconstruction quality, such as parallax angles and baselines between views and their observation angles and distances towards the object's surface. They proposed meaningful heuristics to model the reconstruction quality from different camera constellations. Details about fundamental mathematics covering the subject of close-range photogrammetry and MVS are found in renowned literature (Förstner and Wrobel, 2016; Hartley and Zisserman, 2003; Luhmann et al., 2013).

The development of heuristics should imitate the stereo reconstruction process for arbitrary camera configurations and a given object surface by predicting the object surface's reconstruction quality from this camera view configuration. By defining a suitable objective function, the selection of additional viewpoints is guided by a reconstruction score to ensure a complete 3D model. Different hand-crafted heuristics were developed encompassing multi-view requirements (Alsadik et al., 2013; Hoppe et al., 2012; Smith et al., 2018), ground resolution (Bircher et al., 2016; Hoppe et al., 2012), 3D uncertainty (Dunn and Frahm, 2009; Mostegel et al., 2016), or coverage of the object (Hepp et al., 2018b; Roberts et al., 2017; Smith et al., 2018). In Alsadik et al. (2013), multi-view requirements were addressed by optimizing whole image blocks, which are further guided by synthetically rendered images from the proxy model to ensure the matchability in the image graph. Hoppe et al. (2012) included the multi-view requirements already in the candidate view generation. For each triangle in the proxy model, fronto-parallel views and short distances are sampled following MVS demands. This iterative process selects views that observe most triangles at a novel angle and considers the matchability towards the surface. However, the matchability between adjacent views is not directly assessed in the viewpoint generation, which might lead to deficient overlap. A similar approach was proposed by Bircher et al. (2016) by selecting one admissible viewpoint for each triangle in the target object. An iterative resampling scheme is employed to compute viewpoints that allow low-cost connections, while the best tour among theses viewpoints is finally obtained by solving a standard traveling salesman problem (TSP). Dunn and Frahm (2009) deployed a hierarchical uncertainty-driven model designed to select viewpoints based on the model's covariance structure, appearance and camera characteristics. Viewpoints were selected which best reduce the existing model's 3D uncertainty, however, only viewpoints locations were estimated within the approach, rather than an entire collision-free trajectory between the viewpoints. Similarly, Schmid et al. (2012) developed a view planning heuristic that considers coverage, a maximum viewing angle, and an overlapping constraint imposed by MVS reconstruction techniques allowing for 3D reconstructions. However, the set of calculated viewpoints is highly redundant, leading to a multitude of images and, again, connecting the selected viewpoints by a feasible trajectory was not the scope of the research. Instead of using hand-crafted heuristics, several works developed machine learning methods to learn heuristics that allow predicting the confidence in the output of an MVS without executing it (Devrim Kaba et al., 2017; Hepp et al., 2018a; Mostegel et al., 2016). Although learning-based methods are capable of substituting complex hand-crafted modeling of multi-view geometries, they are less flexible in providing a generalized method applicable for various camera intrinsics and different reconstruction requirements, such as a desired GSD.

Recent works have claimed that the resulting 3D models obtained from a unique *exploitation* flight do not exhibit satisfying reconstruction quality and still could include uncertainties and gaps in hardly observable parts of the object (Huang et al., 2018; Peng and Isler, 2019). They proposed to iteratively refine the model with repeated *exploitation* flights while taking into account the remaining model uncertainty between each flight. With this procedure, additional views of previously unsatisfying object details can drastically improve the entire reconstruction quality.

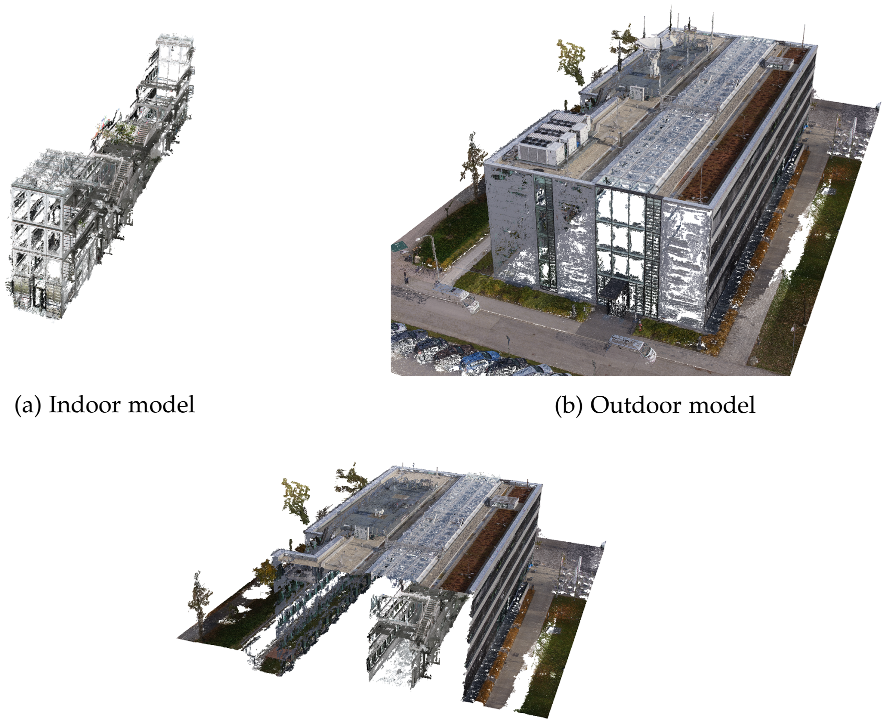
Lately, efficient methodologies formulated the view planning problem as a discrete optimization task and exploited submodularity in the optimization process, standing for fast and reliable convergence, even for a large number of viewpoint hypotheses (Hepp et al., 2018b; Roberts et al., 2017). The main advantage of this idea is to jointly assess additional information gain of individual viewpoints for arbitrary viewpoint constellations in a global manner. This allows formulating the path planning task as an orienteering problem, which can be solved with simple greedy algorithms by optimizing a path that collects as many information rewards as possible for a specific path length. The results presented in these works revealed notable trajectories for generating high-fidelity image-based 3D reconstructions.

However, setting a suitable path length in the optimization may require expert knowledge and highly affects the trajectory estimation, since, due to the purely additive nature of an orienteering problem, adding additional views will never decrease the objective function. This might lead to abundant redundant views for overestimated path lengths and incomplete reconstructions for underestimated path lengths. Although the presented heuristics follow best practices for MVS requirements, they do not respect user-specific demands on the resulting 3D model. Those include the number of views and observations angles for each part of the object surface, and a unified model resolution, expressed by a consistent GSD for the entire object surface. Additionally, prior work so far solely considered purely geometric cues for flight planning of both small-scale and large-scale areas. With the vast progress in semantic segmentation using deep learning-based approaches, the applicability of neural networks for semantic segmentation of aerial and UAV imagery was demonstrated in several works (Chen et al., 2018a; Kaiser et al., 2017; Marmanis et al., 2016). Including semantic cues in the flight planning would greatly impact flight safety since it allows to define inadmissible airspaces above prohibited or hazardous objects, such as other buildings, highly frequented streets, vehicles, railways, or water bodies.

Summarizing, UAV path planning for photogrammetric applications in practice still relies on 2D flight paths on the basis of available (but potentially outdated) topographic maps, while only a few include a rough DSM. Off-the-shelf mission planners compute simple geometric patterns to achieve a desired spatial resolution, which is only valid for the ground level. For the reconstruction of elevated areas or high-rising objects, the obtained GSD varies with the scene depth. Therefore, a consistent GSD for all reconstruction parts is not feasible with such standard planners. Automated methods are getting more advanced and can be used for an automated image acquisition even in full 3D space with respect to MVS image acquisition necessities. Among them, online methods might be used for an entirely automated exploration flight in case the environment is completely unknown. However, these approaches tend to perform rather slow exploration flights, potentially miss to cover all object parts, and do not aim for high-resolution reconstruction results. In case a coarse 3D proxy model of the environment is available, offline methods exploit the knowledge of the scene layout and the accessibility of free and occupied airspaces. These methods optimize a detailed flight concerning high-quality reconstruction results, which, in terms of photogrammetric applications, is of higher interest than complete autonomy but eventually incomplete reconstruction results. However, most of these methods do not consider real photogrammetric parameters for trajectory optimization, and moreover, none of them consider the semantics of the surrounding environment. As shown in Section 2.1, increasing stringent regulations for the operation of UAVs are emerging and undesired maneuvers of the vehicle, as well as, outages of the aircraft may still occur. Therefore, UAV flight plans have to consider the environment by avoiding hazardous flight spaces and adhering non-flight areas. Including these necessities in an automated flight planning pipeline for full 3D model reconstruction has not been addressed in current research.

3.3 ALIGNMENT OF INDOOR AND OUTDOOR BUILDING MODELS

As presented in Section 2.3.1, recent image-based 3D reconstruction pipelines, renowned for generating detailed 3D building models from UAV imagery and the modeling of building interior, have seen substantial improvements. Nevertheless, only limited research was done on investigating an automatic alignment of corresponding 3D indoor and outdoor models reconstructed by individual image sequences, such as those shown in Figure 3.3. However, the existing demand for LoD-4 building models, as pointed out in Section 2.3.2, requires automated and accurate registration methods. Furthermore, successful registration of two complete surface models of the exterior and interior could allow drawing conclusions about volumetric parameters of the building, such as wall thickness, which would constitute a further step towards building information models (BIMs). However, the challenges encountered for obtaining a joint indoor and outdoor model usually



(a) Indoor model

(b) Outdoor model

(c) Aligned models

Figure 3.3: Task of generating complete 3D building models by aligning individual image-based indoor and outdoor building reconstructions. Corresponding reconstruction parts lack in insufficient and noisy visual overlap

include weak visual overlap, changes in illumination in transition areas, temporal changes between the acquisition times, incomplete and drifting reconstructions, and ambiguous alignment solutions.

The tremendous complexity faced for the registration of individual reconstruction models was demonstrated in the *Chillon Project* (Strecha et al., 2014), which aimed to entirely reconstruct the interior and exterior of a large-scale castle in Switzerland. Due to different camera models and acquisition modes, encompassing both terrestrial and aerial footage, an automated joint reconstruction process was not feasible for all parts, particularly when linking indoor and outdoor models. Instead, multiple sub-models were generated and manually transformed into the same reference frame afterward with the help of GCPs and manually selected tie points in the images. Although the resulting joint model revealed an impressive reconstruction of a large-scale and complex architectural object, it also demonstrated the extensive manual interaction required to connect multiple sub-models, and thus obtain a joint indoor and outdoor model.

Partly, joint sub-models encompassing both interior and exterior can be achieved when great attention is paid to the image acquisition in transition areas. If the transition is sufficiently reachable from both sides for image acquisition, such as this is the case for doorways, a multitude of images with very high visual overlap could prevent the reconstruction from disconnecting into multiple-sub models. However, images in such transition areas often suffer from severe illumination changes, hampering a successful feature correspondence estimation. By including and relying on higher-level image information, such as the use of semantics, improvements in image matching and visual localization could be achieved. Schönberger et al. (2018) proposed a semantic visual localization method that is able to match features over extreme appearance changes across viewpoints and time. Similarly, a semantic consistency score, developed by Toft et al. (2018), rates the consistency of each 2D-3D match and uses this score to prioritize more consistent matching during random sampling consensus (RANSAC)-based pose estimation. However, obtaining reliable and sufficient 2D-2D correspondences in transition areas is not guaranteed and still requires carefully designed and abundant images for linking both models. Additionally, if the connection is based on a single transition, it is hard to prevent drift between the two models, which could cause unpleasing geometric inconsistencies for large-scale reconstructions.

Therefore, an alignment between both models is more likely to succeed if the registration is conducted in 3D space by obtaining 3D transformations between the corresponding reconstructions. Traditionally, variants of iterative closest point (ICP)-based registration are utilized for deriving a 3D transformation that registers individual 3D point clouds. Applying ICP, however, assumes high geometric accuracy in the point clouds, low clutter, and in particular, a sufficient degree of overlap between the individual point cloud sets (Pomerleau et al., 2013). These requirements are highly likely missed in the case of linking indoor and outdoor models. First, as already described in Section 2.3.2, indoor models prevalently suffer from a large amount of noise, caused by unfavorable acquisition geometries, specular surfaces, and ambiguous correspondences. Although modern filtering techniques, such as those presented in the comprehensive review of Han et al. (2017), allow for a reliable clutter removal in the point clouds, the requirement of sufficient overlap between indoor and outdoor models usually does not apply, since the attainable overlap is by far too little as only window frames and doorways share the same geometry between the models. Experiments on recorded data have shown that ICP-based registration techniques fail for such extremely little overlap. To overcome the problem of deficient overlap, higher-level features and scene knowledge were utilized to ease the registration process. Cohen et al. (2015) exploited symmetries and repetitive structures of building façades, as well as semantic reasoning to find reasonable connection points of adjacent models that used for linking the models. Their method focused on merging multiple SfM reconstruction models of a single outdoor building model which could not be connected due to occlusions or insufficient visual overlap. While the disconnected models still represented the same building façade in their method, the strong symmetric assumptions are not valid regarding indoor and outdoor models.

However, in a later work, Cohen et al. (2016) explicitly addressed the task of indoor-outdoor model alignment. Their work was done around the same time as the proposed methodology in Appendix C (Koch et al., 2016a) was published. The author's approach followed a similar strategy by aligning the 3D models on the basis of detected windows in the 3D models. They utilized semantics for the detection of windows in indoor and outdoor image sequences and computed 3D positions of these openings with the estimated camera poses and sparse point clouds derived from SfM. The subsequent registration is based on finding an adequate configuration of multiple matching hypotheses between window sets in both models while including an intersection quality metric, that rejects hypotheses that would lead to protrusions of the indoor model. Although their registration approach is robust to noisy and missing window detections and the intersection quality metric avoids geometrical incorrect registrations, the accuracy of the alignment strongly depends on accurate semantic segmentation and low clutter in the sparse point cloud derived from the SfM.

A robust and accurate registration based on sparse or dense point clouds might be unsuccessful due to excessive noise and data gaps since the number of distinctive interest points are limited in poorly textured indoor environments. Another abstraction of a 3D scene was proposed in the work of Hofer et al. (2015) by representing a 3D environment with a set of 3D line segments. As a substitution of MVS approaches, the proposed Line3D method uses straight 2D line segments as underlying features and includes geometric constraints to match the extracted 2D lines from overlapping images. The resulting 3D line segments are especially useful for built environments, mainly consisting of piece-wise linear and planar structures. Therefore, an investigation into the use of such higher-level geometric line features for the task of registering indoor and outdoor building models is sought.

3.4 single-image depth estimation for scene reconstruction

3D information retrieval from images is fundamental for a variety of applications, including 3D modeling, scene understanding, and autonomous navigation. Therefore, it is indispensable for both UAV-photogrammetry and indoor modeling. Besides using active sensors, such as laser scanners and RGB-D cameras, image-based 3D information usually relies on the triangulation of 2D-2D correspondences across stereo images with known interior and relative orientations. Advances in 3D computer vision have relaxed the requirement of rigid stereo camera setups, enabling 3D camera pose estimation and 3D reconstruction for a monocular camera by leveraging camera motion between the image acquisitions. These developed SfM and simultaneous localization and mapping (SLAM) methodologies still constitute the current state-of-the-art in image-based monocular 3D reconstruction. However, actual trends are moving towards depth prediction based on a single view. Although this task, commonly known as monocular depth estimation (MDE), has been approached with a broad range of techniques, current deep learning approaches strive to substitute explicit physical and optical models with implicit learning of scene priors from extensive RGB-D datasets, allowing to estimate dense depth maps from a single view image, which is referred to as single-image depth estimation (SIDE). The following survey on both established and novel algorithms with a focus on most recent deep learning-based methods is part of the published paper enclosed in Appendix D (Koch et al., 2019b). A concise overview of the methods presented hereafter is listed in Table 3.2.

Multi-view depth perception is derived by geometric constraints from multiple observations of a scene using stereo camera setups or leveraging camera motion and constitutes the most established image-based depth retrieval method. The former rely on a prior calibration of the stereo setup and dense point correspondences across the stereo images to estimate depth via geometric triangulation. The task of optimal pixel-wise disparity estimation is usually addressed by local, semi-global, or global optimization methods (Szeliski, 2010). While local methods (Yoon and Kweon, 2006) evaluate pixel correspondences in a point-wise approach, yielding fast,

Table 3.2: Overview of different image-based depth estimation methods (from Koch et al. (2019))

Group	Approach	Method	Literature
Multi-view	Calibrated stereo setup	Local, semi-global, global	Felzenszwalb and Huttenlocher (2006); Hirschmuller (2005); Kolmogorov and Zabih (2001); Yoon and Kweon (2006)
	Unordered image stacks	SfM + MVS	Hartley and Zisserman (2003); Seitz et al. (2006); Szeliski (2010)
	Light-field cameras		Doorn et al. (2011); Heber and Pock (2016)
Single-view	Active	Shape from focus/defocus	Favaro and Soatto (2005); Suwajanakorn et al. (2015)
		Lightning conditions	Ackermann and Goesele (2015)
		Polarization cues	Kadambi et al. (2015); Ngo et al. (2015)
	Passive	Shape from shading	Horn (1970); Zhang et al. (1999)
		Atmospheric optics	Nayar and Narasimhan (1999)
	Learning-based	Parametric	Baig and Torresani (2016); Furukawa et al. (2017); Hane et al. (2015); Hoiem et al. (2007); Ladicky et al. (2014); Li et al. (2014); Liu et al. (2010); Ranftl et al. (2016); Saxena et al. (2006); Saxena et al. (2008); Saxena et al. (2009); Shi et al. (2015); You et al. (2014)
		Non-parametric	Choi et al. (2015); Karsch et al. (2014); Kong and Black (2015); Konrad et al. (2012); Konrad et al. (2013); Liu et al. (2014)
	Deep learning-based	Supervised	Chakrabarti et al. (2016); Eigen and Fergus (2015); Eigen et al. (2014); Fu et al. (2018); Hao et al. (2018); Heo et al. (2018); Hu et al. (2019); Kim et al. (2016); Laina et al. (2016); Lee et al. (2018); Li et al. (2015); Li et al. (2017); Liu et al. (2018); Liu et al. (2015); Liu et al. (2016); Ramamonjisoa and Lepetit (2019); Roy and Todorovic (2016); Wang et al. (2015); Wang et al. (2016); Xu et al. (2018); Yang and Zhou (2018); Zhuo et al. (2015); Zoran et al. (2015)
		Unsupervised	Garg et al. (2016); Godard et al. (2017); Kuznietsov et al. (2017); Ummenhofer et al. (2017); Yin and Shi (2018); Zhan et al. (2018); Zhuo et al. (2015)

but often inaccurate correspondences due to their sensitivity towards appearance changes and occlusions, global (Felzenszwalb and Huttenlocher, 2006; Kolmogorov and Zabih, 2001) and semi-global methods (Hirschmuller, 2005), on the other hand, make explicit smoothness assumptions and solve for a global optimization problem formulated as energy minimization frameworks. Those methods result into accurate and less noisy depth maps but require significantly increased computation times. A prominent representative for semi-global methods constitutes the well-known semi-global matching (SGM) algorithm (Hirschmuller, 2005). Methods that leverage monocular camera motion are utilizing SfM or SLAM methods to transform multiple single-view images to a stereo problem, which can be subsequently addressed by MVS methods (Szeliski, 2010). Extensive studies in the field of two or more frame stereo correspondence algorithms can be found in reputable literatures such as Scharstein and Szeliski (2002), Hartley and Zisserman (2003) and Seitz et al. (2006). A further line of approaches was developed with the emergence of light field cameras using an array of micro-lenses placed in front of the image sensor (Doorn et al., 2011; Heber and Pock, 2016).

Single-view active methods endeavor to ease the multi-view requirement by addressing the task of depth estimation by a sequence of images from the same perspective. Depth information is obtained either by variations of the camera parameters as for instance shape from focus/defocus methods (Favaro and Soatto, 2005; Suwajanakorn et al., 2015), lighting conditions of the scene, like photometric stereo (Ackermann and Goesele, 2015), or by utilizing polarization cues (Kadambi et al., 2015; Ngo et al., 2015).

Single-view passive methods further relax the requirement of image sequences or camera displacements and address the problem of depth estimation from a single camera shot. Most prominently, shape from shading (SfS) methods (Horn, 1970) exploit intensity or color gradients of a single image under the assumption of homogeneous lighting and Lambertian surface properties. Although these methods work on single-shots, they only perform well for largely known environments or synthetic data but rather poor on real images in unconstrained environments (Zhang et al., 1999). Another early approach aimed at exploiting light sources and illumination conditions, such as haze and fog in an image to recover the relative scene depth by relying on atmospheric optical models (Nayar and Narasimhan, 1999).

Single-view learning-based methods utilize machine learning techniques in order to implicitly learn scene priors on the basis of a large amount of aligned RGB and depth map pair training samples. As one of the first learning-based approaches, Torralba and Oliva (2002) focused on absolute depth estimation for a query image by incorporating the size of known objects depicted in the image. Instead of decomposing the image into its constituent elements, the absolute scene depth of the image is derived from the global image structure represented as a set of features from Fourier and wavelet transformations. The features of the query image were finally compared towards a model trained with 4000 images and corresponding scene depths in a cluster-weighted modeling approach. With the release of first large-scale RGB-D datasets (Geiger et al., 2012; Saxena et al., 2009; Silberman et al., 2012), data-driven approaches became feasible and rapidly began to outperform established model-based methods. A pioneer work of a supervised learning-based approach was firstly proposed by Saxena et al., 2006 by training a discriminatively-trained Markov random field (MRF) incorporating multi-scale local and global image features to infer depth. An extension of this work to 3D scene reconstruction was proposed in a subsequent work (Saxena et al., 2009). Since then, a variety of approaches have been proposed to exploit the monocular cues using hand-crafted features together with graphical models (Baig and Torresani, 2016; Furukawa et al., 2017; Hane et al., 2015; Hoiem et al., 2007; Li et al., 2014; Ranftl et al., 2016; Saxena et al., 2008; Shi et al., 2015; You et al., 2014). Enhanced results have been achieved by incorporating semantic labels in the depth prediction scheme (Ladicky et al., 2014; Liu et al., 2010).

Single-view non-parametric learning-based methods assume similarities between RGB values and depth cues across a large set of images (Choi et al., 2015a; Karsch et al., 2014; Kong and Black, 2015; Konrad et al., 2012; Konrad et al., 2013; Liu et al., 2014). First, similar images of the input image are retrieved from a RGB-D database by feature-based matching with the query image. Depth complements of the nearest neighbors are subsequently combined and either cross-bilateral filtered for smoothing the final depth map (Konrad et al., 2013), warped towards the input image using SIFT flow (Karsch et al., 2014; Liu et al., 2011) or optimized via a conditional random field (CRF) (Liu et al., 2014).

Single-view deep learning-based methods emerged with the undeniable influence of deep learning within the field of computer vision and shifted this research field towards the use of convolutional neural networks (CNNs) for depth estimation. Since 2014, several works have significantly improved SIDE performance by using deep models, demonstrating the superiority of deep features over hand-crafted features (Chakrabarti et al., 2016; Eigen and Fergus, 2015; Eigen et al., 2014; Fu et al., 2018; Kim et al., 2016; Laina et al., 2016; Lee et al., 2018; Li et al., 2015; Li et al., 2017; Liu et al., 2018; Liu et al., 2015; 2016; Roy and Todorovic, 2016; Wang et al., 2015; 2016b; Xu et al., 2018a; Zhuo et al., 2015; Zoran et al., 2015). These methods pursue the problem of SIDE as a regression problem by building upon successful architectures and learning a deep CNN to estimate the continuous depth map. The first work using deep models was proposed in the work of Eigen et al. (2014) in a two-scaled architecture. A coarse global prediction is performed with one network in a first stage, while another network locally refines the prediction in a successive second stage. An extension to this approach uses deeper models and additionally predicts normals and semantic labels (Eigen and Fergus, 2015). Some works have harnessed the power of pre-trained CNNs in the form of fully convolutional networks (FCNs) (Chakrabarti et al., 2016; Eigen and Fergus, 2015; Laina et al., 2016; Li et al., 2017). The convolutional layers from networks such as AlexNet (Krizhevsky et al., 2012), VGG (Simonyan and Zisserman, 2014) and ResNet (He et al., 2016) were fine-tuned, while the fully connected layers were re-learned from scratch to encode a spatial feature mapping of the scene. One main limitation using CNNs for depth prediction is the decrease of the output map resolution due to repeated pooling operations in the deep feature extractors. In order to preserve the local structures of output depth maps, several authors have attempted to cope with this problem by up-sampling (Chakrabarti et al., 2016; Eigen and Fergus, 2015; Li et al., 2017), up-convolution blocks (Laina et al., 2016), skip connections between the up-sampling blocks (Li et al., 2017) and space-increasing discretization (Fu et al., 2018). Improving the quality of predicted depth maps was also addressed by combining CNNs and graphical models, such as CRFs (Kim et al., 2016; Li et al., 2015; Liu et al., 2015; 2016; Wang et al., 2015; Xu et al., 2017a; 2018a). A deep convolutional neural field combining CNNs and CRFs in a unified framework was proposed in the works of Liu et al. (2015) and Liu et al. (2016) by estimating depth for superpixels generated in a preceding segmentation step, while enforcing smoothness between adjacent segments via a CRF. Li et al. (2015) and Wang et al. (2015) use hierarchical CRFs to refine their patch-wise CNN predictions from superpixel down to pixel level. CRFs can be further exploited by fusing the multi-scale information derived from inner layers of a CNN (Xu et al., 2017a; 2018a). A combination of CNNs and regression forests with very shallow architectures at each tree node reduces the need for big data (Roy and Todorovic, 2016). A further attempt exploits the Fourier frequency domain in a deep learning algorithm (Lee et al., 2018).

With the successful use of deep architectures for SIDE, authors have begun to focus on major challenges of this task, such as distorted depth discontinuities (Hao et al., 2018; Hu et al., 2019; Ramamonjisoa and Lepetit, 2019) and defects in predicting planar regions (Heo et al., 2018; Liu et al., 2018; Wang et al., 2016b; Yang and Zhou, 2018). Although the recovery of the actual scene geometry is a key requirement for many application fields, meaningful evaluation capabilities are lacking to accurately assess the impact of such spatially local enhancements of depth maps. Besides the deficiency of a holistic evaluation protocol, current RGB-D datasets utilized for testing the performance of SIDE methods do not comply with the required accuracies and quality specifications. A comprehensive survey of current datasets is presented in the attached paper in Appendix D.

Unsupervised and semi-supervised deep learning-based were recently introduced in order to cope with the problem of the unavailability of a sufficient quantity of training samples (Garg et al., 2016; Godard et al., 2017; Kuznietsov et al., 2017; Ummenhofer et al., 2017; Yin and Shi, 2018; Zhan et al., 2018; Zhou et al., 2017). This is accomplished by an intermediate task of a view synthesis and allows for training by only using stereo pairs as input with known baselines. These methods design reconstruction losses to estimate the disparity map by recovering a right view with a left view.

Synthetic data is another approach to address the lack of available training data. With the emergence of synthetic datasets, first work was done to exhibit the possibility to render noise-free and dense depth maps in a very large scale. However, the large domain gap between synthetic data and real data is still a very challenging task. First works in this field are trying to handle this gap (Guo et al., 2018; Zheng et al., 2018).

Ordinal depth prediction simplifies the absolute depth estimation by predicting dense relative depths from pairwise relationships (closer-than and further-than relationships) estimates for rare points in the input image (Chen et al., 2016; Zoran et al., 2015). It has been shown that this simplification leads to better results at the cost of absolute depth estimation. However, several applications only require relative or ordinal depth relations, such as 2D-to-3D conversion (Karsch et al., 2014), image refocusing (Anwar et al., 2017) or foreground-background segmentation (Camplani and Salgado, 2014).

Single-image depth estimation of uav images has only been scarcely addressed. A brief investigation about the capability of current SIDE methods for depth estimation of oblique UAV was presented in the work of Julian et al. (2017). After re-training existing SIDE methods with synthetic UAV images, the presented results exhibited coarse and blurry depth predictions which might not be accurate and reliable enough for direct applications. This study suggested a more detailed investigation and adaption of sophisticated methods for the task of UAV-SIDE. A more advanced method was proposed by Marcu et al. (2018), presenting an embeddable and fast CNN for joint depth, obstacles and safe landing areas estimation. A pixel-wise segmentation of UAV images into safe horizontal areas and hazardous vertical and other areas defines potential landing areas for the vehicle, while depth maps are predicted to guide the landing procedure. The results have demonstrated the benefit of jointly estimating depth and semantic maps, while depth predictions revealed reasonable geometric representation of the environment allowing for the localization and guidance of landing maneuvers. To address the lack of available training data in the domain of oblique UAV imagery, the authors made use of extracted image and depth map pairs from Google Earth.

3.5 CONTRIBUTIONS OF THIS THESIS

Based on the objectives defined in Section 1.2 and the current state-of-the-art of associated UAV-photogrammetry tasks presented in Sections 3.1 to 3.4, the contributions of this thesis are summarized as follows.

3.5.1 *Automatic Indirect Geo-registration of UAV Imagery*

As pointed out in Section 3.1, many attempts have been already made to enable accurate geo-registration of UAV imagery, in particular by matching UAV images towards already georeferenced image data, which would prevent the elaborative deployment of GCPs. However, prior work exhibits large deficits either in the achieved registration accuracy or in the robustness of the proposed methodology for diverse environments. Although works that rely on feature-based image matching strategies show superior results than area-based methods, most approaches have not considered the domain-specific differences in matching multi-modal aerial image data, such as geometric, radiometric, and temporal differences between UAV and aerial or satellite images, as well as homogeneous and repetitive structures that occur in remote sensing data. Based on a comprehensive analysis of feature-based multi-modal image matching, the limitations and bottlenecks causing the failure of such methods in many scenarios were revealed and subsequently addressed to enhance the matching quality. A novel feature-based image matching pipeline was developed for the specific task of accurately registering multi-modal aerial images. The proposed methodology includes a dense feature extraction approach, a one-to-many matching scheme, and a geometric match verification strategy that substitutes the precarious ratio-test.

3.5.2 *Semantic-aware 3D UAV Path Planning for 3D Reconstruction*

With regard to high-quality 3D reconstructions of complex scenes from UAV imagery, an accurate and efficient automated flight planning strategy is required for the task of acquiring SfM and MVS-compliant images from different viewpoints. Since UAVs and modern 3D reconstruction methods are capable of quickly obtaining a coarse 3D model of the environment from few overview images, a model-based flight planning strategy, that relies on a coarse DSM or 3D model, is preferable for generating accurate and high-quality 3D models, as already described in Section 3.2. However, previous works have not considered photogrammetric heuristics in the flight planning scheme, ensuring a consistent spatial resolution for the entire object surface after the reconstruction process. Moreover, UAV-photogrammetry often witnesses complex environments that have to be addressed in path planning as well. Practical considerations do not always allow for entirely free accessible airspaces, particularly regarding the semantics of the environment, which could be exploited to define hazardous and prohibited airspaces. This part's contributions include the incorporation of photogrammetric properties and semantics in an automated 3D flight planning methodology, yielding practically feasible, efficient, and safe UAV paths, avoiding restricted airspaces but ensuring high-quality 3D reconstructions in desired spatial resolutions.

3.5.3 *Automatic Building Model Alignment*

In order to comply with the previous efforts in the creation of LoD-4 building models (*cf.* Sections 2.3.1 and 2.3.2), which cover both indoor and outdoor parts, methodologies for accurate alignment between individual models are demanded. As presented in Section 3.3, only a few works have dealt with this subject which is particularly difficult due to the lack of visual correspondences between the individual models. The contributions of the proposed work involve the exploitation of a 3D line representation of the scene for the identification of shared geometric structures in both models. Besides hypothesis generation for potentially matchable planar structures and robust hypothesis matching and verification in 2D, an accurate 3D line-based optimization is presented, enabling registration accuracies in centimeter-range between interior and exterior models with only a small amount of shared 3D structures.

3.5.4 *Evaluation of Single-image Depth Estimation Methods for Indoor Reconstruction*

Based on SIDE's tremendous progress stemming from the adoption of deep learning in this research area (*cf.* Section 3.4), further meaningful evaluation criteria are required to accurately assess the performance in terms of its applicability across different disciplines. Aside from global statistics, local geometric properties of the depth maps are crucial for evaluating the performance in recovering the actual scene geometry. Accordingly, new evaluation metrics have been introduced, comprising the assessment for different scene depths, at distinctive depth discontinuities, and for planar surfaces. Since these metrics require highly accurate ground truth data that cannot be provided by any other established dataset, a new high-quality RGB-D benchmark dataset has been generated intended for independent validation of

SIDE methods. A thorough examination of current state-of-the-art approaches has revealed new insights into these deep learning methods and has identified remaining challenges in this field.

4 SUMMARY OF THE CONTRIBUTIONS FOR UAV PHOTOGRAMMETRY

The objectives of this thesis are addressed in three peer-reviewed journal papers and one peer-reviewed conference paper. This chapter briefly introduces the articles with the following main contributions:

- Section 4.1 is devoted to an automatic and accurate image-based georeferencing methodology for unmanned aerial vehicle (UAV) images
- Section 4.2 presents an automatic and semantically-aware 3D UAV flight planning approach for photogrammetric image acquisition
- Section 4.3 addresses the geometric alignment of individual interior and exterior building models towards the generation of level of detail (LoD)-4 building models
- Section 4.4 demonstrates a holistic evaluation protocol for the currently very active field of single-image depth estimation (SIDE) and presents a comprehensive analysis of the performance, as well as the identification of remaining challenges for the applicability of SIDE in current or future applications

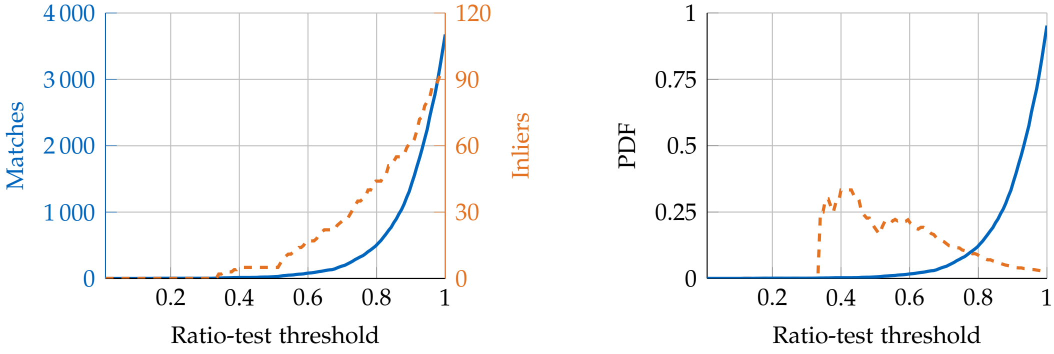
4.1 AUTOMATIC IMAGE-BASED UAV GEOREFERENCING

Appendix A demonstrates the failure of advanced handcrafted feature-based image matching methods for the task of UAV georeferencing with multi-modal image data, such as aerial or satellite imagery. Based on these findings, a methodology for a robust and accurate image matching is proposed.

4.1.1 *Analysis and Bottlenecks of Feature-based Methods*

As a result of comprehensive investigations with various image pair samples of UAV, aerial, and satellite images, the failure of SIFT-based image matching can be explained by a combination of different parts of the algorithm, namely detection and description of the feature points, as well as the feature matching scheme. The bottlenecks can be concluded as follows:

1. The scale-invariant property of SIFT features can not resolve the enormous scale differences between UAV and aerial images
2. The theoretical number of correct feature matches among detected keypoints is vastly reduced by the ratio-test, often resulting in too few remaining matches
3. Correct matches are often not found among nearest neighbors in feature space, but among the top *k*-nearest neighbors
4. The rotation invariance of SIFT is not as good as it has been considered



(a) Absolute numbers of matches (—) and correct inliers (- -)

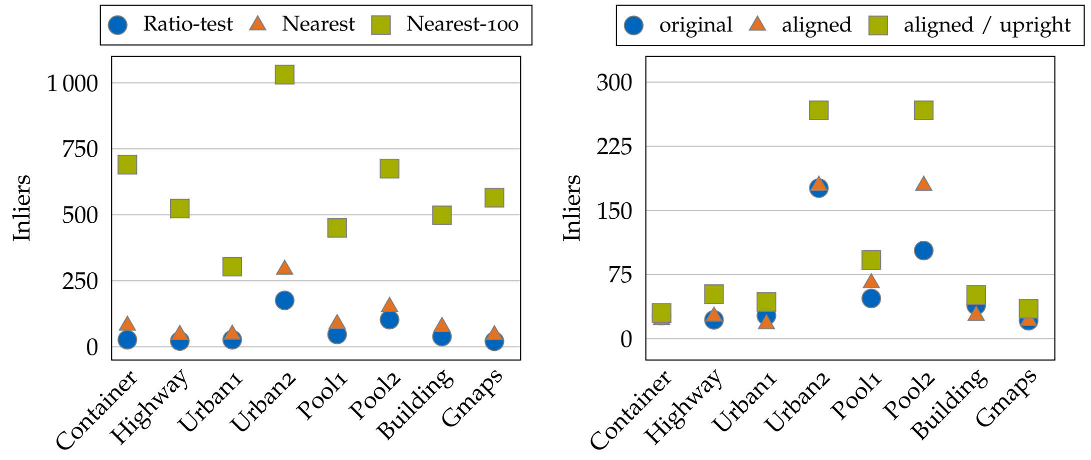
(b) PDFs for correct (- -) and incorrect (—) matches

Figure 4.1: Influence of different ratio-test thresholds for the task of multi-modal UAV and aerial image matching. Number of remaining matches after applying ratio-test and number of correct matches among them (a). Probability density functions (PDFs) of correct and incorrect matches (b)

5. The number of extracted feature points is too small for reliable image matching. Aerial images often depict weakly textured and homogeneous surfaces, which result in an unsatisfying amount and distribution of feature points

Despite the scale-invariant property of the SIFT detector, UAV and aerial images often differ in their spatial resolution up to a factor of ten, which generally can not be resolved by the octave representation in SIFT, resulting in too few keypoint detections in higher octaves. This finding was already pointed out by several researches (Lin et al., 2007; Onyango et al., 2017) and is usually tackled by eliminating the scale difference in the UAV images in advance, making use of either a global navigation satellite system (GNSS) or utilizing barometer altitude measurements.

The widely used ratio-test addresses the elimination of ambiguous feature matches with similar descriptor distances and has been proven to improve the image matching quality of close-range image pairs (Kaplan et al., 2016). However, UAV images often exhibit surfaces with homogeneous and repetitive textures resulting in similar descriptors of detected keypoints. Together with the overall small amount of detected feature points, the ratio-test often eliminates valuable feature matches among them. Figure 4.1 demonstrates the influence of different ratio-test thresholds in eliminating true matches. Based on image pair samples with ground-truth fundamental matrices (considering equally scaled image pairs capturing the same coverage by image resizing and cropping), raw matches after applying the ratio-test were filtered according to the epipolar constraint to identify true inliers and outliers. A maximum number of around 100 correct matches could be found when only the first nearest neighbor was considered (equivalent to a threshold of one) (*cf.* Figure 4.1a). Comparing the number of inliers to the total amount of around 4000 matches, as depicted in Figure 4.1b, a very low ratio of inliers is observable, which would highly likely lead to erroneous registration results. By increasing the impact of the ratio test (equivalent to lower values of the threshold), many correct matches were rejected due to a high similarity to other keypoint descriptors, while the ratio of outliers decreased at the same time. According to the obtained results, the best ratio of inliers is suggested for threshold values between 0.3 and 0.5. However, the absolute number of correct matches for



(a) Theoretical number of inliers after matching with ratio-test, without ratio-test and without ratio-test for 100 nearest neighbors

(b) Theoretical number of inliers after matching un-aligned, pre-aligned and pre-aligned image pairs with upright SIFT features

Figure 4.2: Analysis of SIFT matching performance after rejecting ratio-test and considering *k*-nearest neighbors (a), and fixing orientation cues in SIFT features after pre-alignment of the images (b). Matching was performed on equally scaled UAV images and cropped aerial images to the same image content as the UAV images

these values was found below ten and, therefore, not sufficient for a reliable matching result.

However, investigations on eight different image pairs from various acquisition campaigns have shown that the number of theoretical matches nearly doubles without applying the ratio-test, approving the assumption of rejecting a large number of correct matches by the ratio-test. Moreover, considering not only the nearest neighbors in the feature space but taking into account the one hundred nearest neighbors, the number of theoretical matches could be increased up to a factor of 15. Figure 4.2a compares the number of matches for the different matching schemes on the different image pairs. Considering the number of detected keypoints, the theoretical mean inlier ratio for the methods is 0.013, 0.023, and 0.14, respectively. This investigation demonstrates the bottleneck of the ratio-test, which eliminates a vast amount of correct matches. On the other hand, the SIFT descriptor often fails in identifying correct matches for feature points sharing similar structures. Remote sensing image data often exhibit repetitive and homogeneous structures, while temporal changes might change the descriptor response in a way that corresponding feature points are not nearest neighbors in feature space. However, by increasing the number of putative nearest neighbors, the correct match is often found among the top *k*-nearest neighbors, although the number of mismatches substantially increases as well by incorporating a one-to-many matching scheme. Figure 4.3 shows the increase of obtained true matches for various image samples when considering multiple nearest neighbors of feature correspondences.

Lastly, investigations on the orientation estimation of the SIFT keypoints have demonstrated that a pre-alignment of the images and the utilization of up-right orientations in the SIFT descriptor almost doubled the number of correct matches, as depicted in Figure 4.2b.

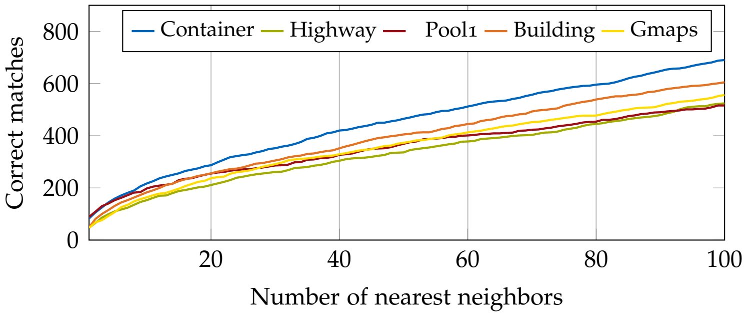


Figure 4.3: Cumulative number of theoretical correct matches considering multiple nearest neighbors in the feature matching for different multi-modal image pairs from the acquired dataset

4.1.2 *Proposed Image Matching Approach*

According to the findings of the image matching failure presented in Section 4.1.1, the proposed matching approach is designed to eliminate each of the exposed bottlenecks by a novel *feature extraction method*, a *one-to-many matching scheme*, and the substitution of the ratio-test with a *geometric match verification*.

Making use of the sensor information onboard of the UAV, first, the scale difference between both images can be estimated and approximately eliminated in advance, and second, a pre-alignment for the image rotation can be achieved with compass information.

In order to boost the amount of matches — a necessity for weakly textured surfaces and for aiding automated decision systems whether image pairs are matching properly or not — a novel *dense feature detection* scheme extracts a vast amount of feature points uniformly distributed in the images by making use of a superpixel segmentation. The boundaries of simple linear iterative clustering (SLIC) (Achanta et al., 2012) superpixels mostly define strong variations in the local neighborhood's intensities, such as edges and corners, which are suitable for representing hypothetical feature points to be found in the corresponding images. Due to the generation step of SLIC, the well-distributed generation of the superpixel ensures a uniform distribution of detected feature points, and all pixels on the boundaries of the segmented superpixels are adopted as feature points for which a SIFT descriptor is computed. Based on the findings in Section 4.1.1, scale-spaces and orientations are fixed in the descriptor. A *one-to-many matching* scheme keeps the k-nearest neighbors after exhaustively matching all keypoint descriptors using Euclidean distance calculation to ensure that correct matches can be even found for corresponding keypoints that do not show nearest descriptor distances. The *geometric verification* step addresses to find correct matches among the plethora of putative matches produced by the dense feature extraction and one-to-many matching scheme. Postulating that both UAV and reference images capture the same planar scene and differences in their scales and rotations have already been eliminated, the transformation between the two aligned images can simply be approximated as a 2D translation $\boldsymbol{t} = \left(t_\mathrm{x}, t_\mathrm{y}\right) \in \mathbb{R}^2$. For each UAV keypoint $\boldsymbol{p}^{\mathrm{uav}}_{i=1\ldots I} = (x^{\mathrm{uav}}_i, y^{\mathrm{uav}}_i) \in \mathbb{R}^2$ and every corresponding match hypothesis $j_{1\ldots k}$ of the k-nearest neighbors in the reference image $\boldsymbol{p}^{\mathrm{ref}}_{i,j=1\ldots k} = \left(x^{\mathrm{ref}}_{i,j}, y^{\mathrm{ref}}_{i,j}\right) \in \mathbb{R}^2$, pairwise coordinate differences $\Delta x_{i,j} = x^{\mathrm{uav}}_i - x^{\mathrm{ref}}_{i,j}$ and $\Delta y_{i,j} = y^{\mathrm{uav}}_i - y^{\mathrm{ref}}_{i,j}$ are computed. The estimation of the unknown translation

$$\boldsymbol{t} = \underset{t_{\mathrm{x}}, t_{\mathrm{y}}}{\arg\max} \sum_{i=1}^{I} \sum_{j=1}^{k} \omega_x \left(\hat{t}_{\mathrm{x}}, \Delta x_{i,j}, r\right) \cdot \omega_y \left(\hat{t}_{\mathrm{y}}, \Delta y_{i,j}, r\right) \tag{4.1}$$

is based on maximizing the amount of putative keypoint matches that satisfy

$$\omega_x \left(\hat{t}_{\mathrm{x}}, \Delta x_{i,j}, r\right) = \begin{cases} 1, & |\hat{t}_{\mathrm{x}} - \Delta x_{i,j}| \leq r \\ 0, & |\hat{t}_{\mathrm{x}} - \Delta x_{i,j}| > r \end{cases} \tag{4.2}$$

and

$$\omega_y \left(\hat{t}_{\mathrm{y}}, \Delta y_{i,j}, r\right) = \begin{cases} 1, & |\hat{t}_{\mathrm{y}} - \Delta y_{i,j}| \leq r \\ 0, & |\hat{t}_{\mathrm{y}} - \Delta y_{i,j}| > r, \end{cases} \tag{4.3}$$

where r defines a threshold related to the scene depth. A histogram voting scheme is utilized for recovering the unknown translation parameters.

In the case of imprecise or unknown image rotation, an extension of the proposed method enables the registration of even un-aligned image pairs. The extracted feature points in the UAV image are rotated around the image center for different discretized rotation values, and the same geometric match verification procedure, as described above, is performed for each specific rotation. The maximum number of raw matches for each rotation is kept and compared towards other rotations values. Correct image rotations are expected to result in noticeably pronounced peaks than incorrect rotations.

A subsequent refinement step solves for truly one-to-one matches and addresses for imprecise superpixel boundaries in both images. For all corresponding matching hypotheses, the corresponding patch is searched in the local neighborhood around the feature points using normalized cross-correlation (NCC). The refinement optimizes all matching hypotheses to the correct location and eliminates duplicate matches. These remaining raw matches can be subsequently used to estimate projective transformations, and thus to geo-register UAV images towards the reference images with the help of height maps or orthorectified mosaics with a high-resolution digital surface model (DSM).

4.1.3 *Performance Assessment with Real-world Data*

Various evaluations have demonstrated the robustness and accuracy of the proposed method, which will be briefly described in the following. The first part compares the raw 2D feature matching performance of the proposed method against the baseline of SIFT matching on a diverse multi-modal image dataset. In the second part, registration of individual orthomosaics generated from aerial and UAV images was conducted, and the registration accuracy was compared against the utilization of ground control points (GCPs) from terrestrial real-time kinematic (RTK)-GNSS measurements. Finally, an application scenario shows how to use the proposed method for enriching 3D building models generated from aerial images with UAV imagery.

Table 4.1 compares the matching performance of SIFT and the proposed method. Besides a multitude of matches that could be found with the proposed method, the registration accuracy was by far more robust and accurate than applying SIFT, which

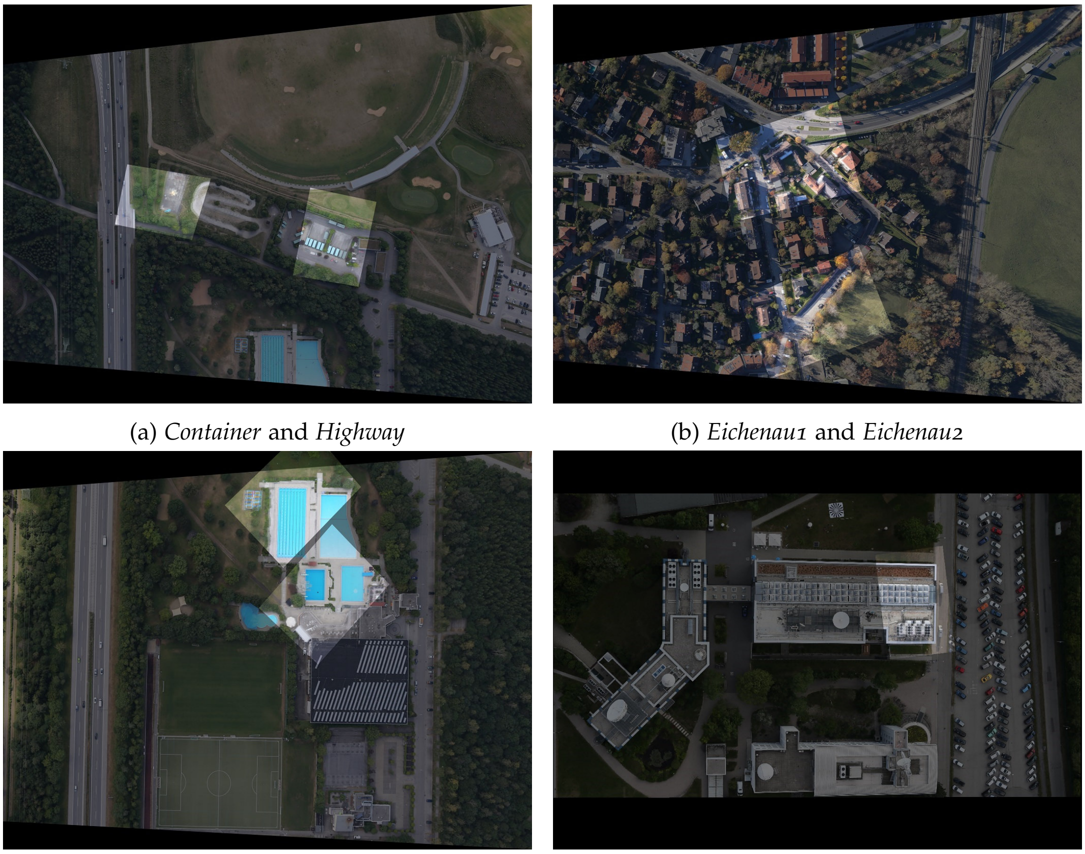
Table 4.1: Comparison of the matching results using standard SIFT and the proposed method. Number of raw matches after applying SIFT and the proposed method for all scenarios. Inliers after estimating fundamental matrix (F) and homography (H) and mean errors according to ground-truth fundamental matrix (F) and homography (H)

Scenario	**Raw Matches**		**Fundamental Matrix (F)**				**Homography (H)**			
			Inliers		Error (in px)		Inliers		Error (in px)	
	SIFT	prop.	SIFT	prop.	SIFT	prop.	SIFT	prop.	SIFT	prop.
Container	58	8264	14	4876	666.26	**2.59**	9	2835	1767.55	**7.01**
Highway	49	1979	15	1184	1996.30	**2.79**	9	1230	2210.20	**1.20**
Pool1	162	6593	52	3599	**0.83**	1.87	33	2188	**1.63**	1.87
Pool2	107	14 091	18	7555	618.54	**2.01**	10	4199	1308.02	**2.03**
Eichenau1	287	4018	45	1850	19.11	**4.35**	48	1165	3.63	**3.53**
Eichenau2	436	5846	140	3204	1.11	**1.09**	146	3077	**3.64**	4.65
EOC	446	6834	16	3949	959.87	**2.92**	6	2586	877.21	**3.18**
WV2	117	15 131	19	6290	175.73	**2.22**	19	6760	4.03	**3.57**
Building	553	9113	16	3526	595.06	**3.15**	11	1932	317.59	**2.36**
Gmaps	522	15 437	19	5120	195.34	**3.42**	8	3217	919.48	**2.82**

often failed in finding plausible matching correspondences. The estimated projective transformations were used to warp the UAV images to the reference images, as visualized in Figure 4.4.

Another series of experiments on two datasets have demonstrated the absolute registration accuracy by matching orthomosaics generated by individual image sequences from aerial and UAV images. After performing image matching with the proposed method, UAV camera poses could be estimated in the same reference frame as the georeferenced aerial orthomosaic exploiting the corresponding aerial DSM. The absolute geo-registration accuracy was assessed with the help of precisely measured GCPs in the test areas. Table 4.2 lists 3D coordinate differences for the GCPs as deviations of the registered UAV orthomosaic and reference orthomosaic ($\text{Error}_{\text{ref}}$). A global assessment of the registration accuracy is reported as deviations of terrestrial measured 3D GCP coordinates and the registered UAV orthomosaic ($\text{Error}_{\text{rtk}}$). The results have revealed geo-registration accuracies in the range of the ground sampling distance (GSD) of the reference images. The absolute georeferencing accuracy was slightly less accurate than terrestrial measurements, especially for the vertical component due to an erroneous DSM generated from the aerial images. A comparison of aerial and UAV-based DSMs after geo-registration is shown in Figures 4.5 and 4.6.

Lastly, an application has been conducted that utilizes UAV images for enriching building models generated from aerial views. Since aerial images often result in occlusions towards the building façades, oblique UAV images can complete the missing parts and increase the overall resolution. Making use of the multi-modal dataset presented in Koch et al. (2016), a sequence of nadir and oblique UAV images was used to generate a detailed 3D building model, while nadir views were automatically registered towards georeferenced aerial images with the proposed image matching method. After registration, an enhanced, complete and georeferenced 3D building model could be generated by merging the registered multi-modal images in a joint bundle adjustment and computing a dense 3D point cloud with an off-the-shelf multi-view stereo (MVS) pipeline. A superimposition of the registered 3D point clouds is shown in Figure 4.7.



(a) *Container* and *Highway*

(b) *Eichenau1* and *Eichenau2*

(c) *Pool1* and *Pool2*

(d) *Building*

Figure 4.4: Qualitative results of the proposed matching method showing superimpositions of UAV images registered towards aerial images

Table 4.2: Absolute geo-registration accuracy comparing the 3D coordinates of GCPs measured from geo-registered UAV images and georeferenced aerial images ($\text{Error}_{\text{ref}}$) towards RTK-GNSS measurements ($\text{Error}_{\text{rtk}}$)

GCP	***Eichenau***				***Germering***			
	$\text{Error}_{\text{ref}}$ (in m)		$\text{Error}_{\text{rtk}}$ (in m)		$\text{Error}_{\text{ref}}$ (in m)		$\text{Error}_{\text{rtk}}$ (in m)	
	Δxy	Δz	Δxy	Δz	Δxy	Δz	Δxy	Δz
1	0.51	−0.21	0.39	−1.74	0.15	−0.38	0.34	1.49
2	0.09	−0.15	0.41	−1.90	0.69	0.37	0.65	1.68
3	0.41	−0.36	0.83	−2.04	0.14	0.46	0.48	1.76
4	0.81	0.70	0.48	−1.91	0.77	0.26	0.80	1.71
5	0.49	−0.17	0.22	−1.81	0.21	0.50	0.50	0.75
6	0.32	−0.10	0.38	−1.63	0.19	0.18	0.40	1.30
7					0.42	−0.06	0.50	1.42
Abs. mean	0.44	0.28	0.45	1.83	0.36	0.32	0.52	1.44

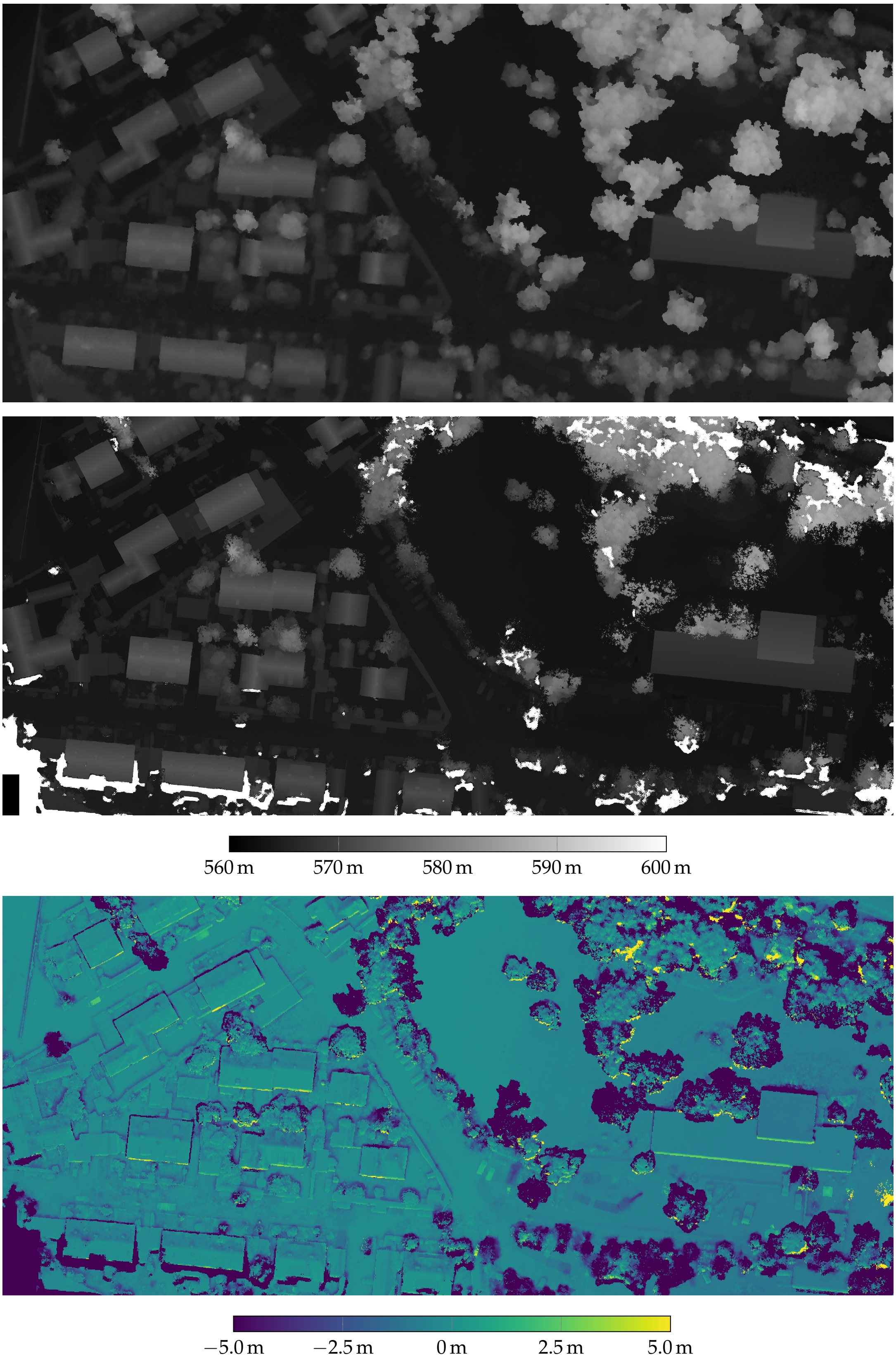


Figure 4.5: Comparison of a reference DSM of the *Eichenau* dataset obtained from georeferenced aerial images with 20 cm GSD (top) and a UAV-based DSM with 2 cm GSD (middle) obtained by registering UAV images towards the aerial images with the proposed matching method. Differences between both DSMs reveal the registration accuracy as color-coded height disparities (bottom)

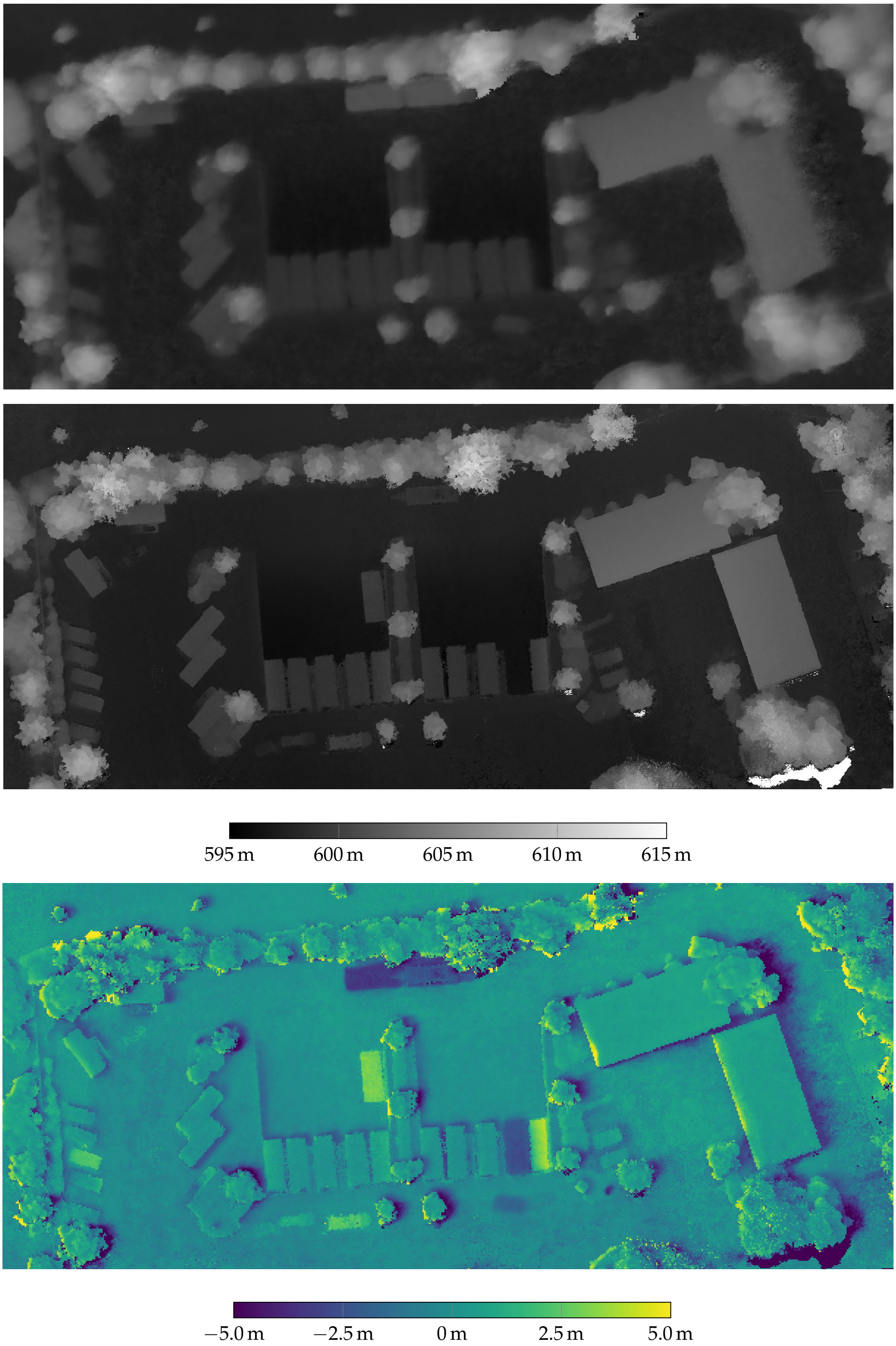
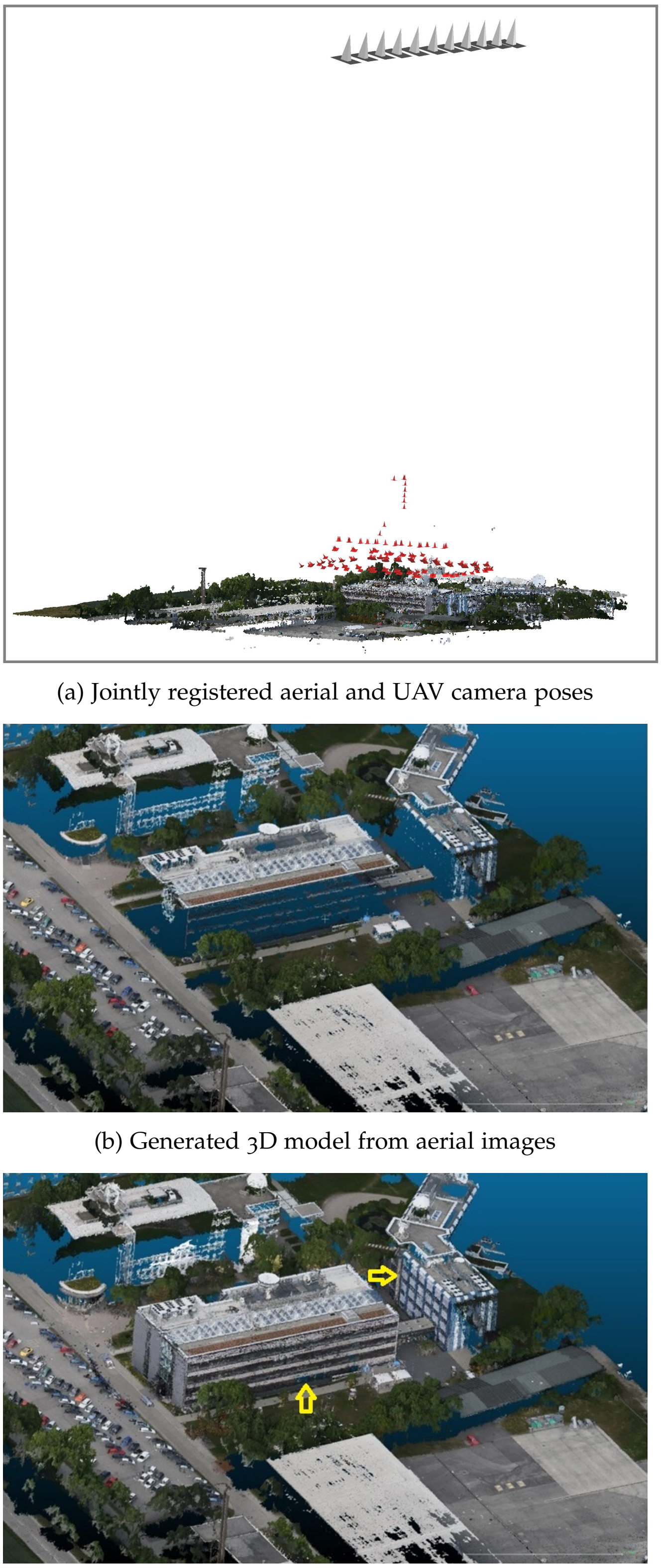


Figure 4.6: Comparison of a reference DSM of the *Germering* dataset obtained from georeferenced aerial images with 20 cm GSD (top) and a UAV-based DSM with 2 cm GSD (middle) obtained by registering UAV images towards the aerial images with the proposed matching method. Differences between both DSMs reveal the registration accuracy and temporal changes of the scene as color-coded height disparities (bottom)



(a) Jointly registered aerial and UAV camera poses

(b) Generated 3D model from aerial images

(c) Generated 3D model from aerial and UAV images

Figure 4.7: Enriching a 3D building model generated from aerial images with the help of UAV images. Camera poses of aerial and registered UAV images after applying the proposed image matching method (a). Dense point cloud derived from using aerial images solely (b) and merged point cloud derived by including UAV images (c)

4.2 Automatic and Semantically-Aware 3D UAV Path Planning for 3D Reconstruction

Appendix B proposes an automatic model-based 3D path planning pipeline designed for UAV image acquisition for small-scale photogrammetry applications, such as the generation of 3D building models. Besides incorporating photogrammetric parameters and minimization of the estimated path length, semantic cues of the surrounding are exploited for generating safe trajectories that avoid hazardous flight zones. The developed methodology consists of the following steps:

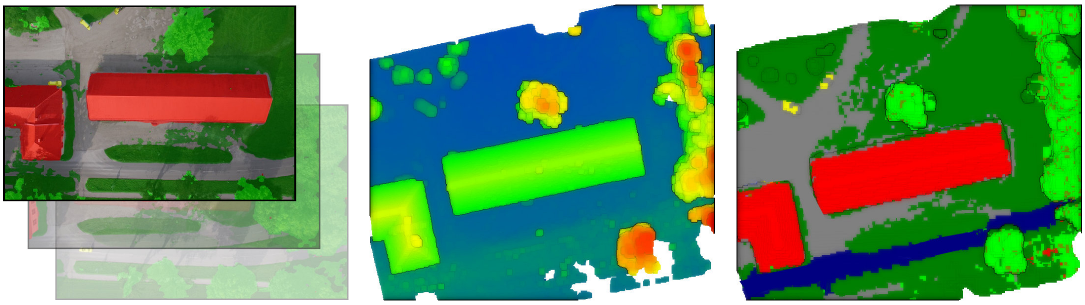
- Generation of a semantic 3D proxy model of the surrounding environment and extraction of the object of interest
- Generation of a multitude of discrete viewpoint hypotheses including camera locations and orientations
- Estimation of a UAV path formulated as a discrete submodular optimization problem by choosing a minimum number of connected viewpoint hypotheses for maximizing the reconstruction quality of the object while considering the semantic properties of the environment to avoid hazardous and restricted flight areas

4.2.1 *Semantically-enriched Prior Model Generation*

The methodology follows a model-based trajectory optimization approach for coverage maximization based on a proxy 3D model of the environment. The prior 3D model is generated from overlapping nadir images of a preceding safe overview flight covering the entire scenery around the target object to be reconstructed. Besides the computation of an approximate DSM from the overlapping images, the images are semantically segmented and fused into the 3D model for generating a semantic 3D proxy model, as exemplarily shown in Figure 4.8. The semantic cues represent relevant object classes that are later used for defining permissible and prohibited flight zones, such as areas above roads, buildings, cars, water basins, low vegetation, and trees. Optionally, the semantic proxy model can be augmented with additional geographic information, such as open street map (OSM) data in order to distinguish between different inter-object classes, such as street types. An exact georeferencing of the overview images, as possible with the proposed method in Section 4.1, can contribute to an exact alignment between the proxy model and extracted OSM objects. Based on the proxy model, a semantic-based 3D region growing approach facilitates the extraction of the target object based on a single manually selected seed point on the object's surface. A discretization of the 3D points of the object and the surrounding environment is conducted, and missing 3D points in occluded regions are completed under the assumption of vertical surfaces.

4.2.2 *Camera Viewpoint Hypotheses Generation*

A large amount of evenly distributed viewpoint candidates $\boldsymbol{c}_{i=1 \ldots I} \in \mathbb{R}^3$ is sampled in the free airspace inside a bounding box around the extracted object, excluding camera viewpoints that are closer to any surrounding obstacle than a predefined safety buffer. This safety buffer can be adapted according to the corresponding semantic



(a) Segmented nadir images

(b) DSM

(c) Semantic 3D proxy model

Figure 4.8: Generation of a semantically-enriched proxy 3D model used for the 3D path planning approach. Based on a sequence of semantically segmented overview UAV images (a), a geometric 3D model of the environment (b) is further enriched with the semantic maps of the input images (c)

labels of the environment to increase the distance towards unreliably reconstructed object classes, such as trees, which often lack completeness after photogrammetric reconstructions. For each viewpoint candidate, a vector containing the semantic labels of all proxy 3D points located below the camera viewpoints is stored. Furthermore, camera orientations $\boldsymbol{r}_{i=1\ldots I} \in \mathbb{R}^3$ are assigned for each viewpoint candidate pointing towards the target object's closest surface points $\boldsymbol{s}_{j=1\ldots J} \in \mathbb{R}^3$ while avoiding occlusions with obstacles by performing a visibility assessment of each viewpoint to identify 3D surface points that are visible from each specific viewpoint location considering the surrounding environment. The proposed viewpoint orientation strategy results in suitable viewpoint orientations applicable for different object outlines. These orientations avoid occluded views and produce almost fronto-parallel views with smooth transitions at object boundaries allowing for large image overlaps required for successful image registration. Finally, a visibility matrix is computed with respect to the camera intrinsics to assess the matchability between different camera viewpoints.

4.2.3 *Semantically-aware Model-based Path Planning Approach*

The path planning problem is expressed as finding a feasible UAV trajectory among the viewpoint hypotheses yielding a set of overlapping images for generating a complete and high-quality 3D reconstruction model of the targeted object. Besides reducing the estimated path to a reasonable total distance, the semantic cues of the 3D proxy model are exploited to either entirely prohibit certain flight areas or partially restrict flight maneuvers above specific objects. Due to discrete sampling of the camera poses, the viewpoints can be represented as an undirected weighted graph $G = (\mathcal{P}, \mathcal{E})$, composed of a set $\mathcal{P}$ of nodes as camera poses $\boldsymbol{p}_{i\ldots I} = (\boldsymbol{c}_i, \boldsymbol{r}_i) \in \mathcal{P}$ and edges $\mathcal{E} = \left\{ \boldsymbol{e}_k = \left(\boldsymbol{p}_i, \boldsymbol{p}_j \right) \right\}$ between adjacent viewpoints that satisfy a specific overlap constraint (*e.g.*, 75 %). Edges are associated with weights $\mathcal{W} = \left\{ \boldsymbol{w}_k = \left(w_k^{\text{eucl}}, w_k^{\text{sem}} \right) \right\}$ comprising the Euclidean distance $w_k^{\text{eucl}} \in \mathbb{R}$ between connected nodes and a semantic label cost $w_k^{\text{sem}} \in \mathbb{R}$. The latter is defined as the Euclidean distance of traversed ground surface points assigned with specific restricted object classes, which can be individually defined for the individual campaign. The objective of the path planning problem is defined as identifying an optimal trajectory

$$\begin{aligned} \mathcal{T}^* &= \arg\max_{\mathcal{T}} R(\mathcal{T}) \\ \text{subject to } \min &\sum_{e \in \mathcal{E}} w^{\text{eucl}}, \\ &\sum_{e \in \mathcal{E}} w^{\text{sem}} < L^{\text{sem}} \end{aligned} \tag{4.4}$$

among connected nodes in G, where $\mathcal{T} = \{\boldsymbol{p}_1, \boldsymbol{p}_2, ..., \boldsymbol{p}_n\} \subset \mathcal{P}$, and $R : \mathcal{P} \to \mathbb{R}$ is a function representing a reconstructability score of the entire target object from a given trajectory $\mathcal{T}$. The constraints in Equation 4.4 are used to minimize the obtained path length and restrict the path from exceeding an accumulated label cost limit L^{sem}. In order to evaluate an arbitrary trajectory $\mathcal{T}$ in terms of the expected reconstruction quality $R(\mathcal{T}) = \sum_{\mathcal{T}} I(\boldsymbol{p}(\mathcal{T}), \mathcal{S})$, a set of heuristics $I(\boldsymbol{p}_i, \boldsymbol{s}_j)$ is required that approximates the impact of an arbitrary camera pose $\boldsymbol{p}$ and stereo configurations for the reconstruction quality of an object surface point $\boldsymbol{s} \in \mathcal{S}$. The proposed heuristics follow best practices for MVS image acquisition, including a smooth distance-related function for maintaining an intended GSD, a smooth observation angle-based function for favoring fronto-parallel views towards the object surface, and a pairwise assessment of observation directions for all viewpoint hypotheses to ensure large parallax angles from the optimized views. Details on the developed heuristics can be found in Appendix B.

Inspired by the works of Smith et al. (2018) and Hepp et al. (2018), a suitable trajectory can be found by exploiting submodularity in the candidate view selection. Submodularity is a property of a set function $f : 2^{\mathcal{P}} \to \mathbb{R}$ that assigns each subset $\mathcal{T} \subseteq \mathcal{P}$ a value $f(\mathcal{T})$ (Krause and Golovin, 2014). $f(\cdot)$ is submodular if for every $\mathcal{T}_1 \subseteq \mathcal{T}_2 \subseteq \mathcal{P}$ and an element $\boldsymbol{p} \in \mathcal{P} \setminus \mathcal{T}_2$ it holds that $\Delta(\boldsymbol{p}|\mathcal{T}_1) \geq \Delta(\boldsymbol{p}|\mathcal{T}_2)$. Regarding the path planning problem, adding more viewpoint candidates to the trajectory, the marginal benefit of adding another viewpoint with a large overlap to the current set decreases. Adding the same viewpoint to a smaller set with limited coverage, on the other hand, leads to larger rewards. This requires $f(\cdot)$ being both monotone and non-decreasing stated as monotonicity, implying that further elements to the set cannot decrease its value. The marginal gain of a viewpoint candidate $\boldsymbol{p}$ towards a trajectory $\mathcal{T}$ is given by $\Delta(\boldsymbol{p}|\mathcal{T}) := f(\mathcal{T} \cup \boldsymbol{p}) - f(\mathcal{T})$. It has been shown that a simple greedy algorithm can be considered for providing a solution to the NP-hard maximization of submodular functions with a reasonable approximation guarantee (Krause and Golovin, 2014).

In order to limit the obtained reward for each surface point from numerous viewpoints to a maximum reconstructability score of 1, the submodular objective function f is constrained by

$$f\left(\boldsymbol{s}_j, \mathcal{T}\right) = \min\left(1, \sum_{\boldsymbol{p}_i \in \mathcal{T}} \frac{1}{v} I(\boldsymbol{p}_i, \boldsymbol{s}_j)\right), \tag{4.5}$$

where v reduces the obtained reward from a single view in order to enforce at least v different views capturing the same surface point $\boldsymbol{s}_j$. Since f is both monotone and non-decreasing, the individual rewards $I(\boldsymbol{p}_i, \mathcal{S}_{\boldsymbol{p}_i})$ for all viewpoint candidates can be transformed to tightly additive information rewards I_i^{add} by utilizing a simple greedy algorithm, that iteratively computes the marginal rewards of each viewpoint for the current reconstructability of each surface point and adds the viewpoint with the highest additive information reward I_i^{add} towards the output set. After executing the greedy method, each viewpoint candidate $\boldsymbol{p}_i$ is coupled with a marginal information reward I_i^{add} representing its value for the object's reconstructability. According to Roberts et al. (2017), a transformation of additive rewards into a standard additive orienteering problem allows to solve the objective as a mixed-integer programming (MIP) optimization problem. Adding path length and semantic restrictions to the objective function lead to

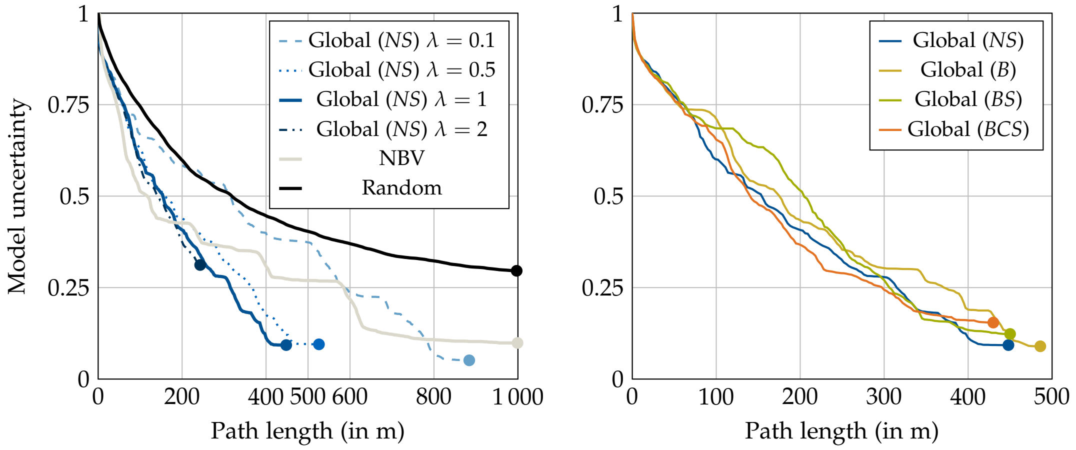
$$\begin{aligned} \mathcal{T}^* = \arg\max_{\mathcal{T}} & \sum_{\boldsymbol{p}_i \in \mathcal{T}} I_i^{\text{add}} - \lambda \sum_{\boldsymbol{e}_k \in \mathcal{E}} w_k^{\text{eucl}} \\ \text{subject to } & \sum_{\boldsymbol{e}_k \in \mathcal{E}} w_k^{\text{eucl}} < L^{\text{eucl}}, \\ & \sum_{\boldsymbol{e}_k \in \mathcal{E}} w_k^{\text{sem}} < L^{\text{sem}}, \end{aligned} \tag{4.6}$$

where I_i^{add} defines the additive rewards of the nodes along a path $\mathcal{T}$ with traversed Euclidean distances $\sum_{\boldsymbol{e}_k \in \mathcal{E}} w_k^{\text{eucl}}$ and traversed distances above semantical restricted airspaces $\sum_{\boldsymbol{e}_k \in \mathcal{E}} w_k^{\text{sem}}$. The regularization forces to reduce the maximum path length L^{eucl} for similar optimization results in shorter paths. The second constraint allows the optimization to select nodes in restricted but not prohibited airspaces, but encourages finding the most efficient and shortest path through these conditionally accessible airspaces by not exceeding a user-defined path length L^{sem} above restricted objects.

4.2.4 *Experimental Results*

The proposed path planning approach was evaluated in terms of feasibility of the estimated paths and reconstruction quality after utilizing the acquired images in an off-the-shelf 3D reconstruction pipeline. Due to the great effort of obtaining accurate ground truth 3D models of real objects and the lack of flexibility of generating different scenarios consisting of various environments, a synthetic scene was generated, composed of various objects arranged to a realistic and interchangeable scenery. The synthetic scene allows to render photo-realistic images to be processed with common 3D reconstruction pipelines and to compare the reconstruction results derived from different trajectories with exact ground truth.

First, a performance analysis of the proposed methodology, as well as the influence of the path length regularization term, were conducted with the synthetic scene. A series of estimated trajectories excluding semantic restrictions with different regularization parameters were compared towards both automated and manual baseline trajectories. Figure 4.9a depicts the expected model uncertainties concerning obtained path lengths for different values of λ. The results show that low regularization parameters yielded large expected model certainties, but also led to lengthy paths, while large values of λ result in shorter paths but reduced certainties of the reconstructability. A reasonable compromise of short path lengths and high model certainty could be realized for regularization parameters in the range of $\lambda = 1$, leading to a slight loss of 4.2% of the model certainty towards the optimized path with $\lambda = 0.1$, whereas



(a) Comparison of the reconstructability for different optimization approaches as a function of path length

(b) Comparison of semantically-aware optimization for $\lambda = 1$

Figure 4.9: Comparison of different optimization methods in terms of the expected model uncertainty for different path lengths assuming the same objective function. The effects of various regularization parameters are shown in blue and the performances of baseline approaches are depicted in gray and black (a). Note that $\lambda = 1$ leads to a balanced trade-off between short path lengths and a high model certainty. Comparison of the semantically-aware global optimization with $\lambda = 1$ for different restrictions on the airspace (b)

the path length has been reduced by half. A comparison of the optimized paths against a random and next-best-view (NBV) baseline revealed the superiority of the global approach, attributed to the exploitation of submodularity contributing to the selection of suitable viewpoints covering all parts of the object, and to the global optimization refining all viewpoints of the trajectory simultaneously, which led to less redundant acquisition views.

Subsequently, an investigation regarding the influence of different semantic constraints of the airspace on the path estimation and the reconstruction quality was carried out. Flight restrictions are two-fold: a hard restriction eliminates nodes and their corresponding edges above a certain semantic cue in the camera graph, while soft restrictions limit the path length to a maximum tolerable distance L^{sem} above specified semantic cues. Precisely, three semantically constrained trajectories were optimized with the following restrictions:

- *No semantics (NS)*: this path serves as a baseline and only considers geometric constraints
- *Building (B)*: hard restriction for airspaces above other buildings
- *Building & Street (BS)*: in addition to (*B*), airspaces above streets were partially restricted to maximum path length of $L = 12\,\text{m}$, approximately twice the width of a regular street
- *Building & Car & Street (BCS)*: In addition to (*BS*), hard restrictions above cars were imposed

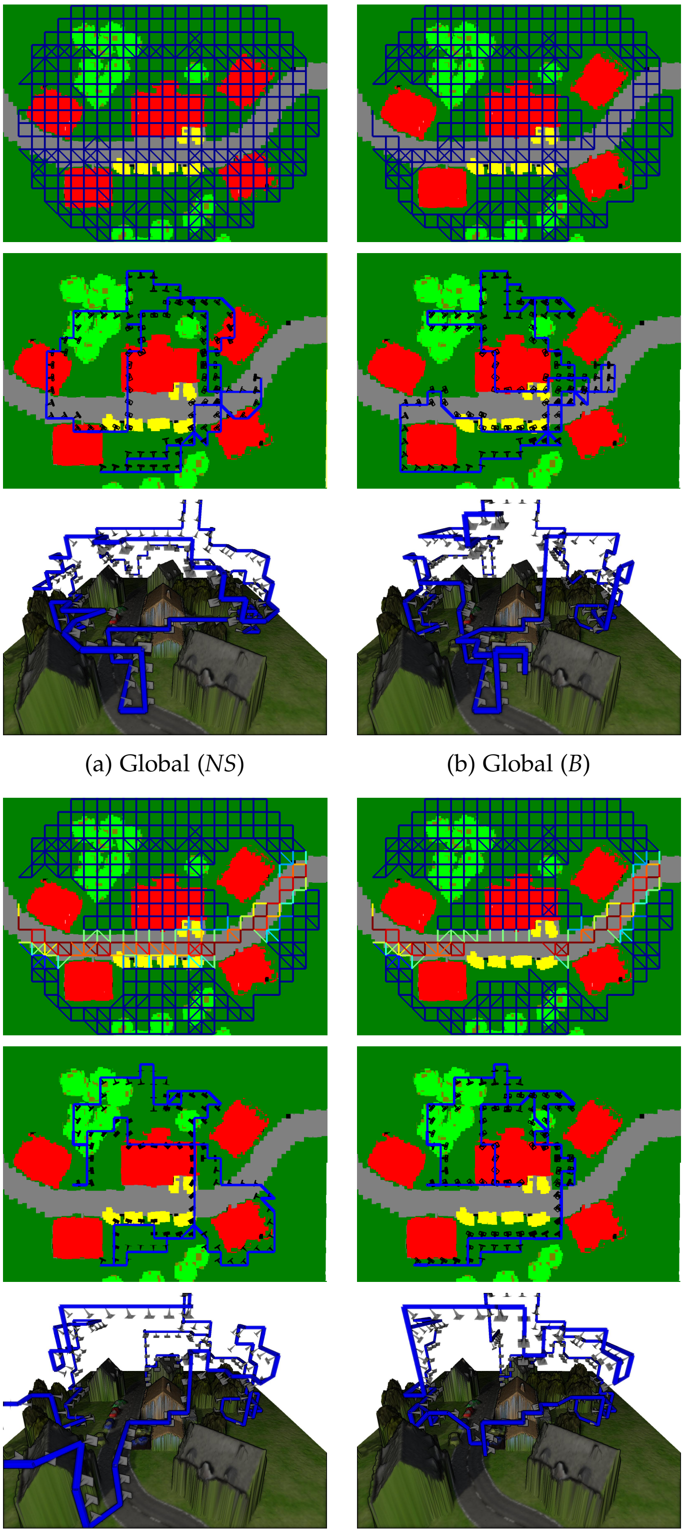
The semantic constraints affected the camera graph's generation, resulting in a limited number of accessible nodes imposed by hard restrictions and only conditionally accessible nodes imposed by soft restrictions. Figure 4.9b reveals that, despite

Table 4.3: Quantitative evaluation of the reconstruction results for the synthetic scene obtained from different path planning methods. Point density is reported as the percentage of reconstructed points that have a shorter distance towards their nearest neighbor than the demanded GSD = 2 cm, as well as one and a half times the distance ($1.5 \cdot \text{GSD} = 3\,\text{cm}$). The reconstruction errors are stated for $d_1 = 5\,\text{cm}$ and $d_2 = 10\,\text{cm}$. The proposed globally optimized paths are superior towards the baseline methods while featuring a shorter path. The severely limited free airspace due to different semantic restrictions only lead to a slight drop in the reconstruction quality

Method	**Images**	**Density (%) ↑**		**Precision (%) ↑**		**Completeness (%) ↑**		**F-Score (%) ↑**	
		GSD	1.5·GSD	d_1	d_2	d_1	d_2	d_1	d_2
Circle 30 m	100	46.9	73.3	88.8	96.3	79.2	91.8	83.7	94.0
Circle 20 m	100	29.7	60.3	89.7	95.8	84.0	93.9	86.7	94.8
Random	321	94.1	98.9	96.4	98.6	83.3	91.0	89.4	94.6
Greedy NBV	323	96.9	99.8	96.8	98.7	86.5	92.7	91.4	95.6
Global (*NS*)	148	97.6	99.9	96.7	98.9	91.1	95.7	93.8	97.2
Global (*B*)	162	97.3	99.8	96.2	98.7	88.3	95.5	92.1	97.1
Global (*BC*)	148	97.6	99.8	96.4	98.7	89.4	94.8	92.8	96.7
Global (*BCS*)	152	97.3	99.8	96.5	98.8	87.7	95.1	91.9	96.9

further limitations of the airspace, only slight losses in the model certainty had to be expected from the optimized paths, indicating decent reconstruction results from a comparable number of acquired images. A visualization of the obtained trajectories is shown in Figure 4.10. While each of the estimated trajectories exhibited both oblique views from the top of the building and horizontal views capturing the buildings façades from low altitudes, restricted trajectories have been furthermore successfully avoided banned airspaces and found suitable locations to efficiently cross the restricted road. In terms of flight safety, these trajectories are by far more desirable than the unconstrained path, since risky long-term periods above hazardousness roads were mostly avoided.

The use of the synthetic model allows for conducting a quantitative and qualitative evaluation of the reconstruction quality from arbitrary viewpoints by rendering the obtained viewpoints and subsequently processing the images with an established 3D reconstruction pipeline, such as Pix4D (Pix4Da). Besides trajectories derived from the proposed methodology, the evaluation included both random and NBV trajectories, as well as manual circular flights at two different altitudes and radii. The quality of the reconstructed point clouds was quantitatively assessed towards the ground truth model using the quantities of precision, completeness, and F-score. Furthermore, an assessment of the point density, which is required to be consistent along the entire object surface, was conducted by computing geometrical distances between neighboring reconstructed points. A quantitative evaluation regarding the reconstruction errors and point density error is listed in Table 4.3, and a visualization of the spatial occurrences of these errors is shown in Figure 4.11. While circular baseline paths revealed unsatisfying reconstruction results in terms of a low point density and gaps in the reconstruction due to occlusions from overhangs of the roof and balcony, the unconstrained global optimization (*NS*) yielded the best reconstruction quality for all investigated errors. The distance-based heuristics led to close-up views, resulting in a high global point density for more than 97% of all reconstructed points of the building. Comparing the completeness error, lower circular flights yielded less optical occlusions, which, however, were limited by the surrounding environment.



(a) Global (*NS*)

(b) Global (*B*)

(c) Global (*BS*)

(d) Global (*BCS*)

Figure 4.10: Visualization of the optimized paths for different semantic restrictions on the airspace for the synthetic scene. Nadir view of the entire camera graph as accessible and traversable UAV viewpoints (top). Color-coded edges represent associated semantical costs w_k^{sem} for the corresponding applied restrictions. Visualization of the optimized camera paths (middle) and 3D perspective with the RGB proxy model (bottom)

Paths considering the proxy generally performed better in terms of completeness, since low altitude viewpoints could be selected from the free airspace, however globally optimized paths revealed significantly better completeness, especially for occluded areas. It is worth noting that the globally optimized paths did not exceed a path length of 490 m, acquiring a maximum amount of 162 images for (*B*), while both random and NBV paths were limited to 1000 m resulting in 321 and 323 viewpoints, respectively. Comparing the results of different semantic restrictions on the airspace, only a minor decrease in terms of completeness was notable, matching the expected model uncertainty in Figure 4.9b. Regarding the precision of the reconstruction—a quality measure according to the noise of the reconstruction depending on the camera constellations—it can be noted that all paths considering viewpoints from the proposed camera graph achieved comparable good values, proving the suitability of the proposed viewpoint generation process.

An assessment of the proposed methodology for real-world applications was carried out at two sites, comprising differently shaped buildings in complex environments, consisting of other buildings, high vegetation, parked cars, and a crossing trunk road. Based on few nadir-directed overview images, semantic 3D proxy models served as a basis for the subsequent path estimation, which considered the semantic cues for prohibiting flyovers above other buildings and the trunk road. Figure 4.12 summarizes the conducted experiments, showing the initial 3D proxy models, the optimized path, samples of the automatically acquired images, and the final 3D reconstruction models. These experiments have shown that the proposed path planning methodology is capable of generating precise 3D flight plans to create reliable and accurate 3D models while considering the surrounding environment to plan safe trajectories. The methodology simplifies the realization of photogrammetric UAV campaigns since no prior scene knowledge is required, and the execution of the flight can be carried out in an automatic manner. Furthermore, the expected accuracy of the reconstruction from the optimized path can be provided before the actual flight, allowing to modify the path or to plan subsequent trajectories.

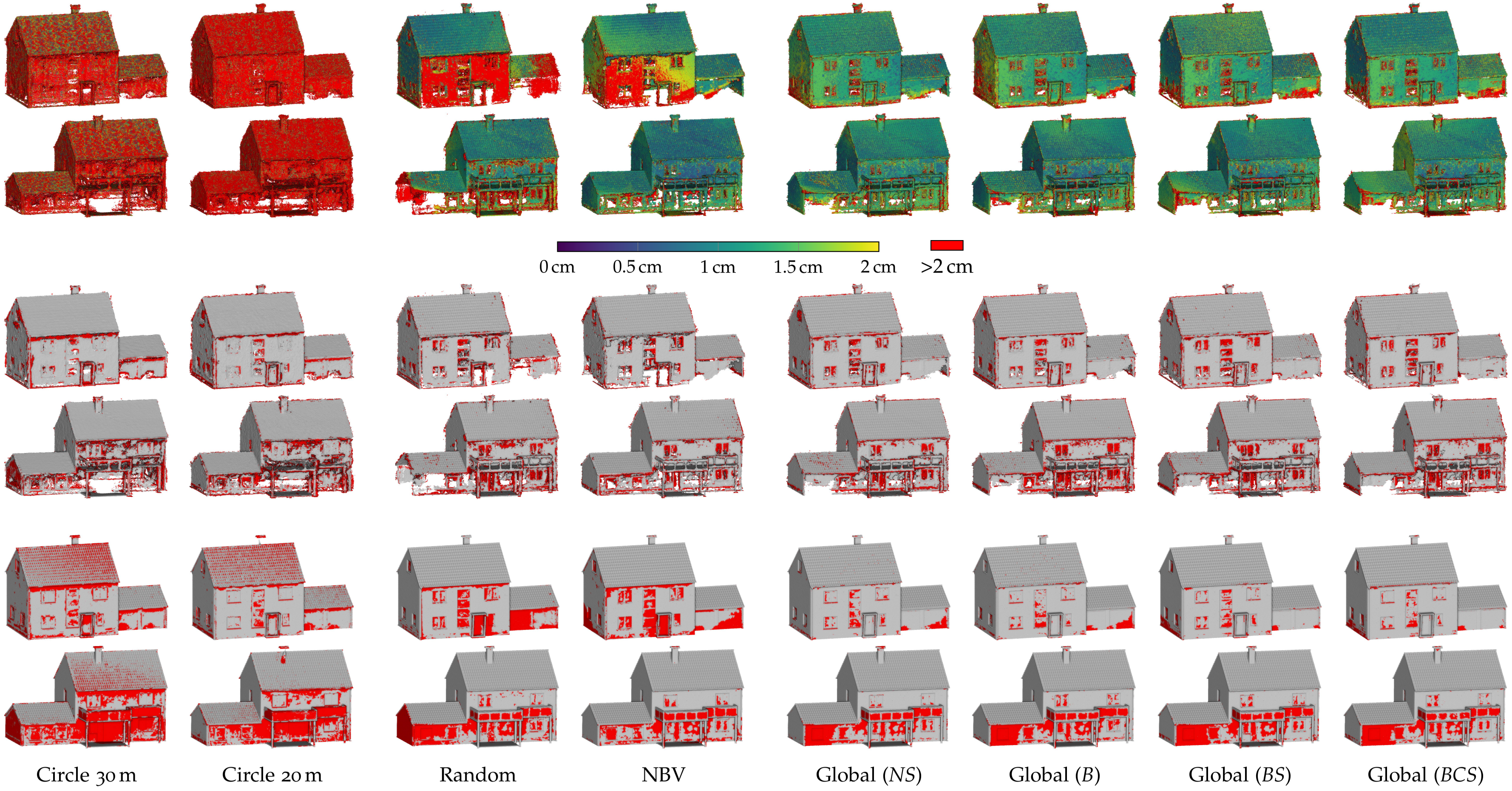
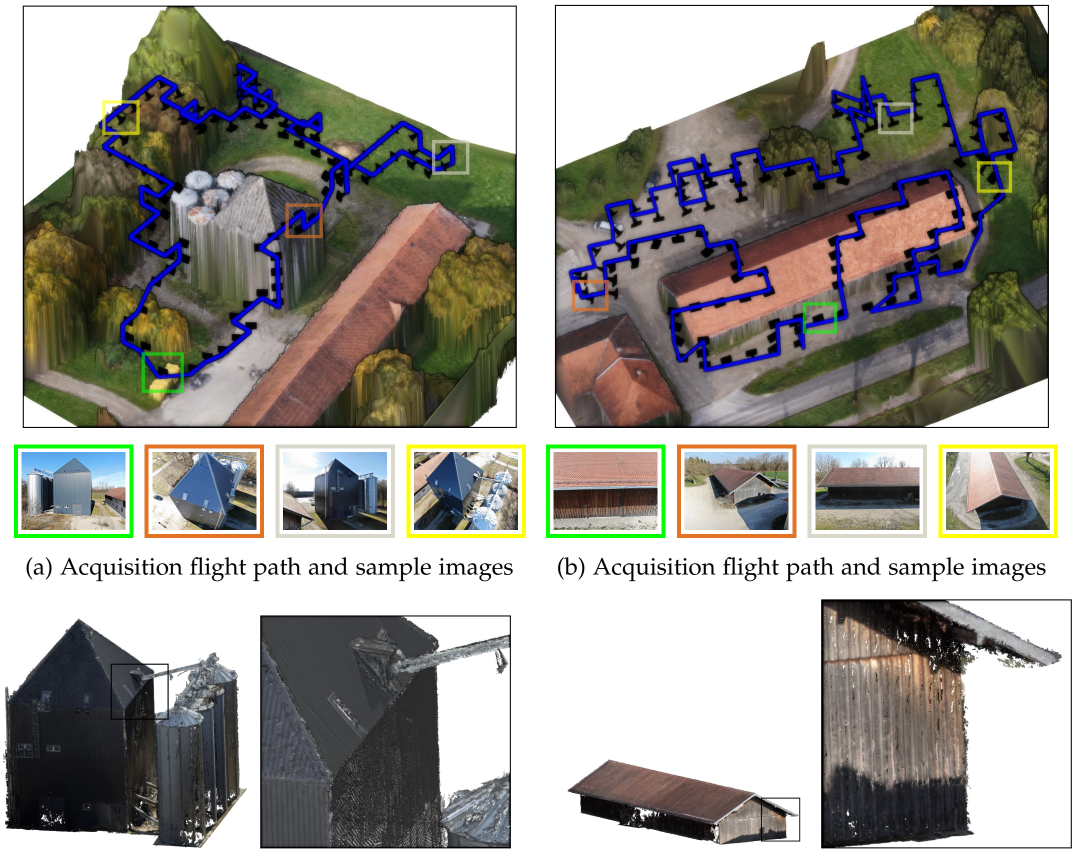


Figure 4.11: Qualitative comparison of the reconstruction results on a dense point cloud for the synthetic scene using different methodologies (columns). The first two rows show the point density as colored distances towards adjacent points on the front and rear side of the building, while red points indicate distances above the required GSD of 2 cm. The reconstruction errors of precision and completeness for $d = 5\,\text{cm}$ are visualized in rows three and four, and rows five and six, respectively, wherein red points indicate erroneous areas



(a) Acquisition flight path and sample images

(b) Acquisition flight path and sample images

(c) Reconstruction model

(d) Reconstruction model

Figure 4.12: Real world experiments for the *Silo* scene (left) and *Farm* scene (right). Optimized trajectories (blue lines) and discrete image acquisition viewpoints (black cameras) are visualized in (a) and (b), including sample images from the acquisition flight for the highlighted viewpoints. Restricted areas include adjacent buildings for the *Silo* scene and adjacent buildings, as well as trunk roads for the *Farm* scene. Visualizations of the final 3D reconstruction models derived from the acquired images are depicted in (c) and (d)

4.3 AUTOMATIC ALIGNMENT OF INDOOR AND OUTDOOR BUILDING MODELS

Appendix C addresses the alignment of individual image-based 3D reconstructions of a building's interior and exterior. This step is considered as a necessity for generating LoD-4 building models. The proposed methodology exploits geometric correspondences rather than appearance-based correspondences due to little or missing visual overlap between both models and eventual independent image acquisitions with large temporal differences, . Using a calibrated camera in the reconstruction process leads to Euclidean (metric) 3D reconstructions of the resulting 3D models, which can be registered with a 3D similarity transformation aligning the indoor building model towards the exterior model.

A summary of the methodology is as follows:

- A 3D line representation of the individual reconstructed models reduces the amount of scene features and reveals geometric matching structures
- Identification of topological planar structures that are shared in both models and coarse alignment derived from a robust matching strategy
- Refinement of the alignment by a 3D line-based optimization

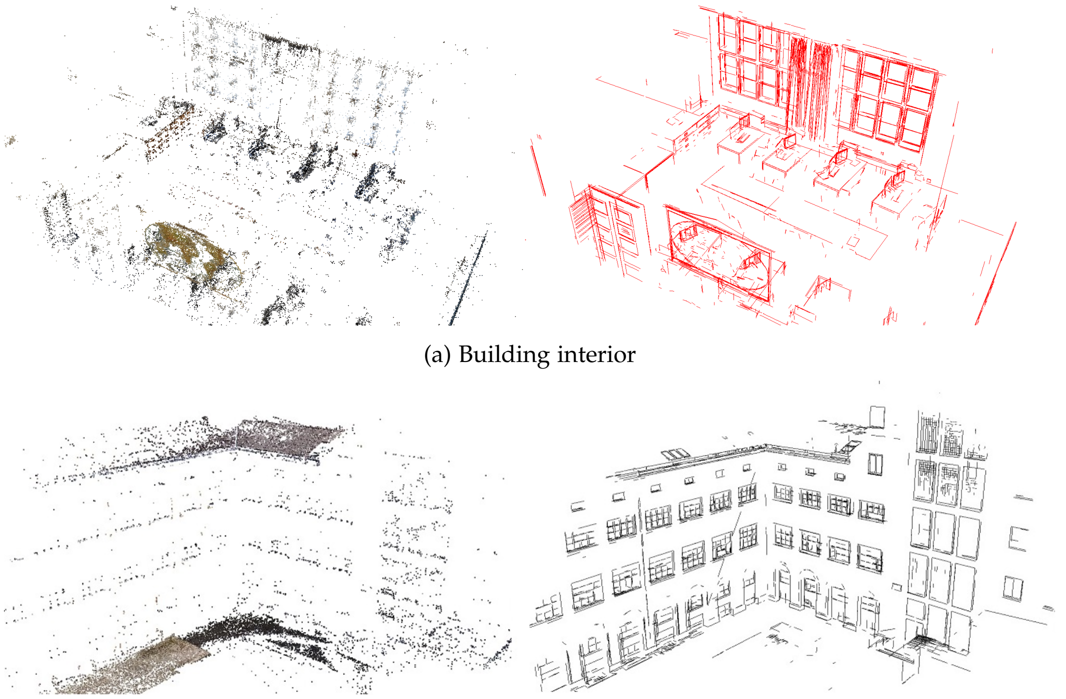
4.3.1 *Geometric Registration Approach*

Since the detection of shared geometric structures between the individual models based on 3D point clouds seems infeasible due to missing information in low textured areas, a 3D line-based scene representation (Hofer et al., 2015) can enrich the derived geometric information in terms of interpreting façades and windows with a reduced number of obtained 3D data (*cf.* Figure 4.13). Given two sets of 3D line segments $L_1 = \left\{l_1^1, ..., l_1^n\right\}$ and $L_2 = \left\{l_2^1, ..., l_2^m\right\}$, representing the building interior and exterior, the overall goal is to find a transformation $\mathbf{T} = (\mathbf{R}, \mathbf{t}, s)$ to align L_1 to L_2, where $\mathbf{t}$, $\mathbf{R}$, and s define the parameters of a 3D similarity transformation as a 3D translation vector, a 3×3 rotation matrix, and a scale. Each segment l is defined by its two endpoints. After identifying k corresponding line segments in L_1 and L_2, the parameters of $\mathbf{T}$ can be estimated by

$$\mathbf{T} = \arg\min_{T} \sum_{i=1}^{k} d\left(l_2^i, \pi\left(l_1^i, \hat{\mathbf{T}}\right)\right), \tag{4.7}$$

where $\pi\left(l, \hat{\mathbf{T}}\right)$ projects a line segment l with $\hat{\mathbf{T}}$, and $d\left(l_2, l_1\right)$ computes the length of the perpendicular of two 3D line segments extended to infinity.

As only a small subset out of several thousand pairs of 3D line segments in $L_1 \times L_2$ are expected to be correct 3D line matches, an exhaustive matching scheme is not applicable. Instead, the matching problem is reduced to 2D by first generating multiple 3D plane hypotheses for both models with an iterated 3D line-based random sampling consensus (RANSAC) scheme. Since window frames and doors are expected to be located on planar façades, the top k plane hypotheses will likely include the demanded geometric structures. After projecting corresponding 3D lines onto the extracted planes and a discretization step, a pairwise robust 2D binary matching is applied for each extracted 3D plane. The shape matching utilizes robust



(a) Building interior

(b) Building exterior

Figure 4.13: Scene representation as sparse 3D point clouds (left) and corresponding 3D line segments (right) for building interior (a) and exterior (b). The higher-level geometric features of 3D line segments expose potential registration elements such as window frames and doors

Chamfer distance and allows for registration of shared geometric structures in both models, such as windows and doors. The distance maps exhibit distinctive minimums, even in the presence of several non-shared structures in the edge maps. The local minimums in the resulting distance maps indicate potential matching locations of the indoor model, which allow for deriving a transformation matrix **T**. Since this coarse registration might be inaccurate due to the discretization and binary matching step, a refinement procedure identifies 3D line correspondences in 3D and optimizes **T** by minimizing Equation 4.7. An illustration of the proposed workflow is shown in Figure 4.14.

4.3.2 *Experimental Results*

The proposed alignment approach was evaluated on two acquired datasets consisting of UAV image sequences capturing the buildings' exterior and indoor images from a hand-held camera. The ability to obtain accurate registration results for unique building model configurations was demonstrated in the *EOC* dataset, shown in Figure 4.15. Through two shared openings on both sides of the building, corresponding building façades could be identified with the proposed matching approach, and joint optimization of both building sides has enhanced the registration accuracy (*cf.* Figure 4.15e) compared to the registration of solely one building façade (*cf.* Figure 4.15d). The obtained registration accuracy in terms of a mean perpendicular distance of matched 3D line pairs was 4.3 cm. However, the registration often might be ambiguous due to symmetric building shapes or missing reconstruction parts. The *TUM* dataset

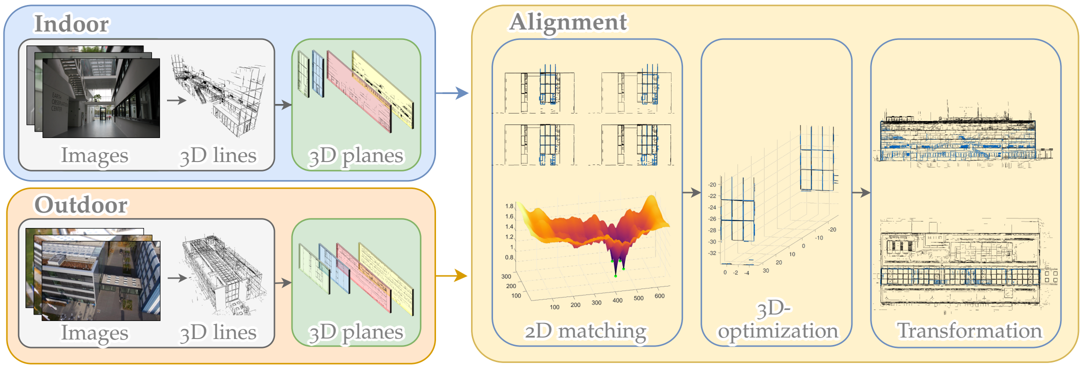
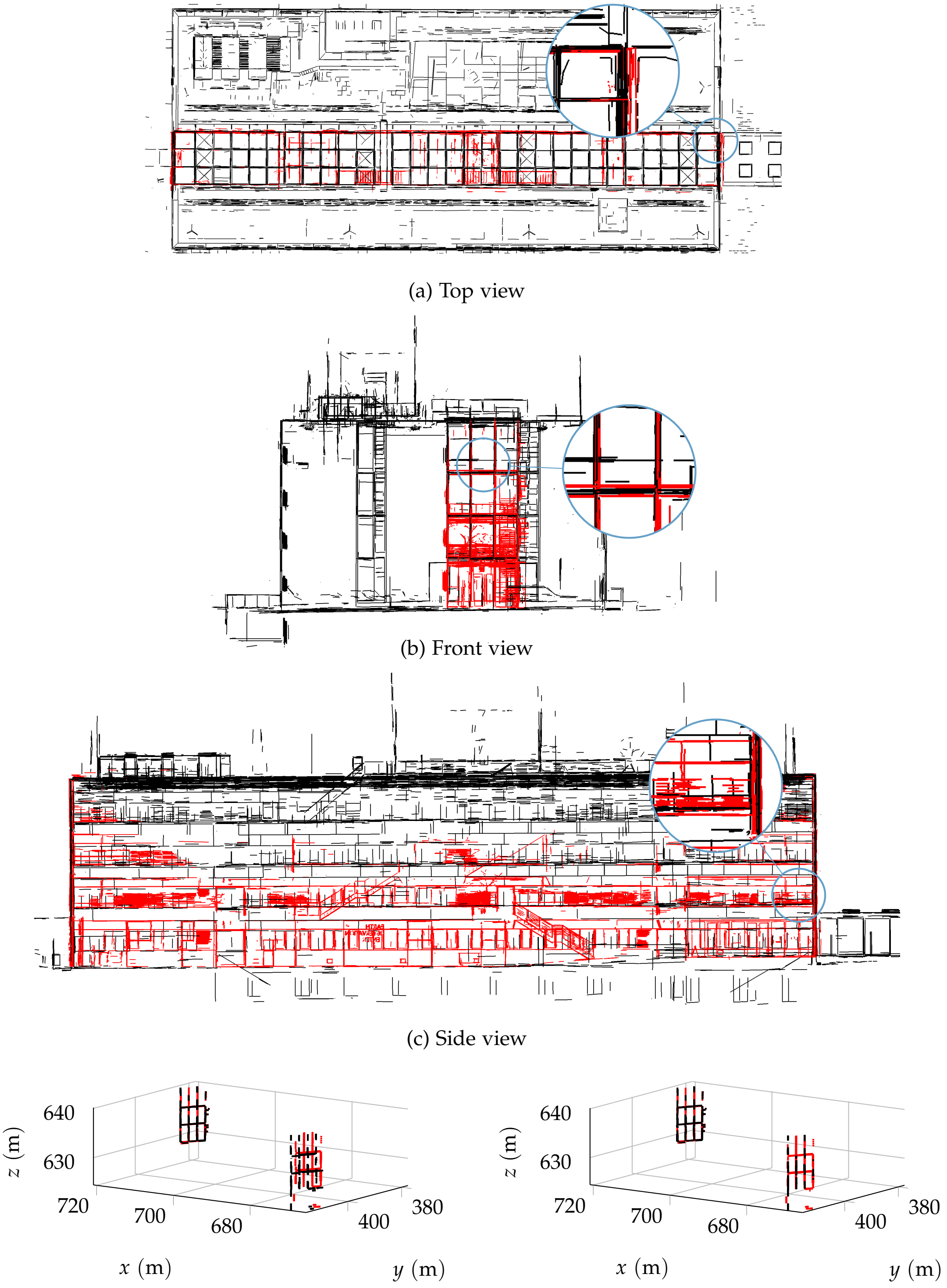


Figure 4.14: Workflow of the proposed method for aligning interior and exterior building models. Based on 3D line scene reconstructions for individual image sequences of the building interior and exterior, multiple planar structures are extracted. A coarse registration is obtained by pairwise binary matching of the extracted 3D planes, while a subsequent refinement step minimizes perpendicular distances between corresponding 3D line matches of both models

represents such a case where a single room has to be registered within a large façade with repetitive structure (*cf.* Figure 4.13). Although the identification of the true location of the room without human knowledge is impossible, the proposed methodology yielded multiple registration hypotheses from which a human operator could choose the correct hypothesis. An illustration of multiple matching hypotheses of the *TUM* dataset is depicted in Figure 4.16 and the obtained 3D alignment for the correct hypothesis (3) is shown in Figure 4.17.

Based on the conducted experiments, it can be concluded that the proposed method offers a useful tool towards the automated generation of LoD-4 building models obtained from individual indoor and outdoor image sequences. Since the methodology's premise is based on the registration of planar openings, it can not be performed on entirely curved or non-planar façades. However, this assumption applies to most buildings, since the method only requires a few planar structures. Even for a small number of corresponding 3D line correspondences, the 2D binary matching scheme enables a robust and accurate localization performance. The refinement achieved by the 3D line optimization significantly improves the coarse registration results—especially when multiple shared structures are utilized—and requires only a small amount of 3D line matches. The 3D line segments could be fragmented due to occlusions in the images, however, by minimizing perpendicular distances between 3D line matches instead of entire line segments, robust and accurate registration results can be realized as long as both horizontal and vertical 3D line matches are present in the set of line matches, which have been found in all of the conducted experiments. One critical assumption is that scale differences in both models have to be largely eliminated, otherwise the 2D binary matching has to be performed for different scale-spaces. Although the correct scale could be recovered for most experiments by including scale optimization in the binary matching, the computational complexity increases. By incorporating an approximate prior registration of both models, the matching approach can be reduced to a local search space and vastly decrease the computational effort.



(a) Top view

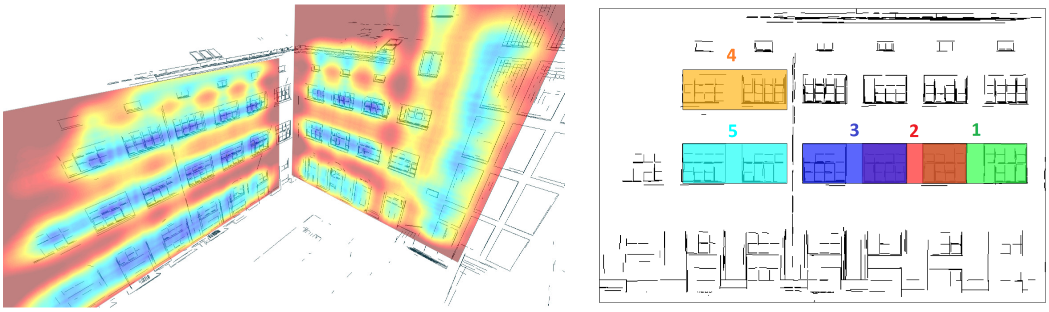
(b) Front view

(c) Side view

(d) Before joint optimization

(e) After joint optimization

Figure 4.15: Alignment result for the *EOC* dataset. 3D lines representing the transformed indoor model (—) and outdoor model (—) (a-c). 298 corresponding 3D line segments could be identified on both sides of the building out of 28*k* and 24*k* 3D lines of the indoor and outdoor model, respectively. A joint optimization of 3D line matches on both building sides substantially enhanced the global registration results towards a mean perpendicular distance between corresponding 3D lines of 4.3 cm (d-e)



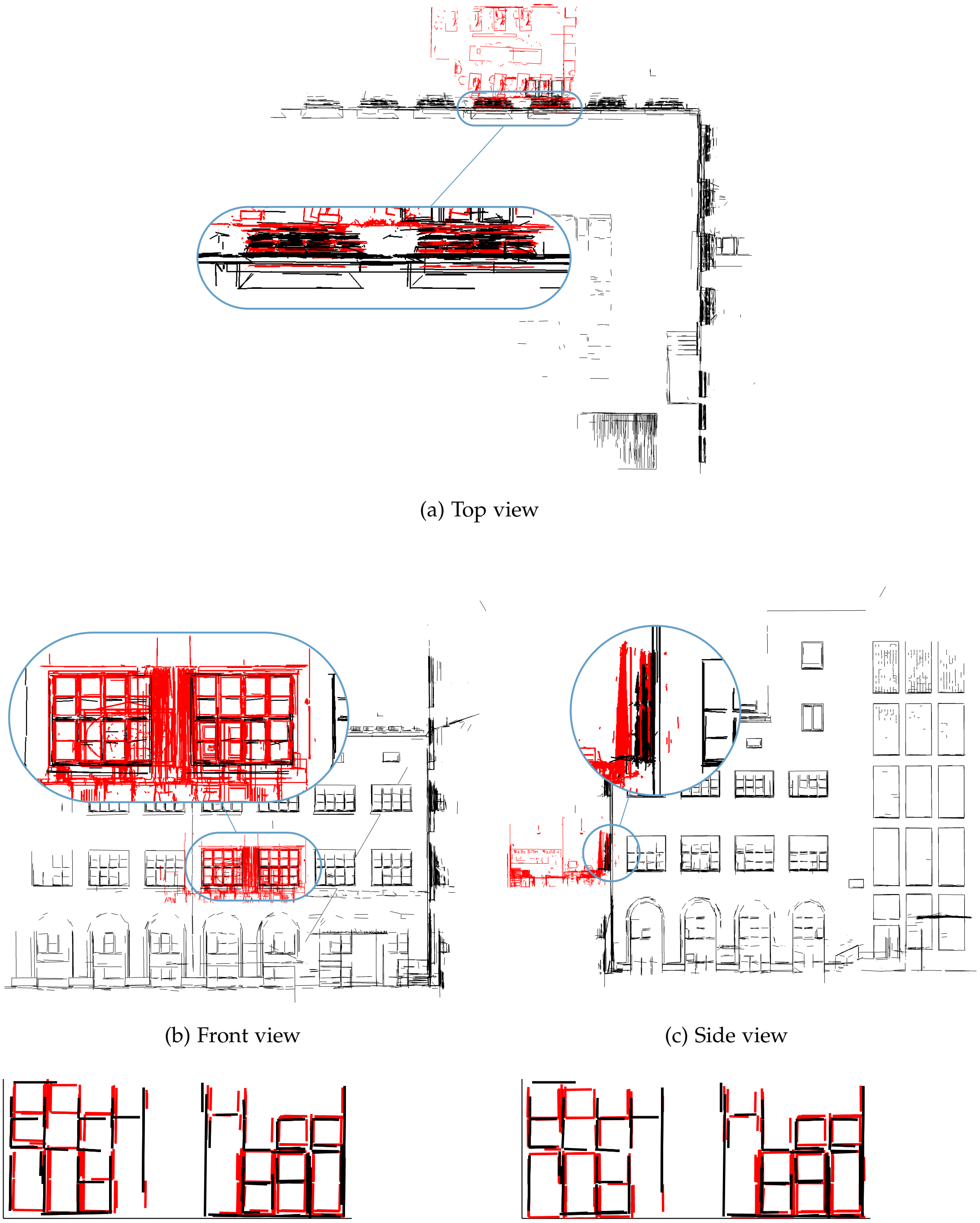
(a) 2D binary matching map with blue color indicating low distances

(b) Top-five registration hypotheses illustrated by colored and numbered boxes

Figure 4.16: Ambiguous registration of a single room to a repetitive building structure exemplified by the *TUM* dataset. Chamfer distance maps indicate possible locations of the indoor model (a). Top-five registration hypotheses (b) yield accurate localization results including the correct location (3) which can be manually selected by a human operator

4.3.3 *Comparison Towards Other Approach*

Cohen et al. (2016) proposed another methodology for the same task after release of the publication presented in this section. They leveraged semantic information for detecting windows in multiple views to obtain candidate matches for the alignment task. Unlike the proposed approach, their optimization accounts for scale differences in the models by comparing matching window outlines. They conducted experiments on various challenging datasets, and the alignment results have mostly shown plausible and accurate model registrations. However, the results strongly depended on the quality of the preceding semantic segmentation of the acquired images and the quality of the sparse reconstruction. Assessing the accuracy of the alignment on the basis of 3D points is hard to realize, however, utilizing a 3D line representation has revealed imprecise alignments for many scenes. The proposed methodology in this thesis can perfectly exploit these pre-alignments derived from the method of Cohen et al. (2016), and thus refine the registration results, leading to vast improvements in the resulting joint building models. Figures 4.18 to 4.20 illustrate the pre-alignment results derived from the method of Cohen et al. (2016) for additional datasets and the optimized alignment achieved from the proposed methodology. It is evident that the refined registration almost perfectly matches both models.



(a) Top view

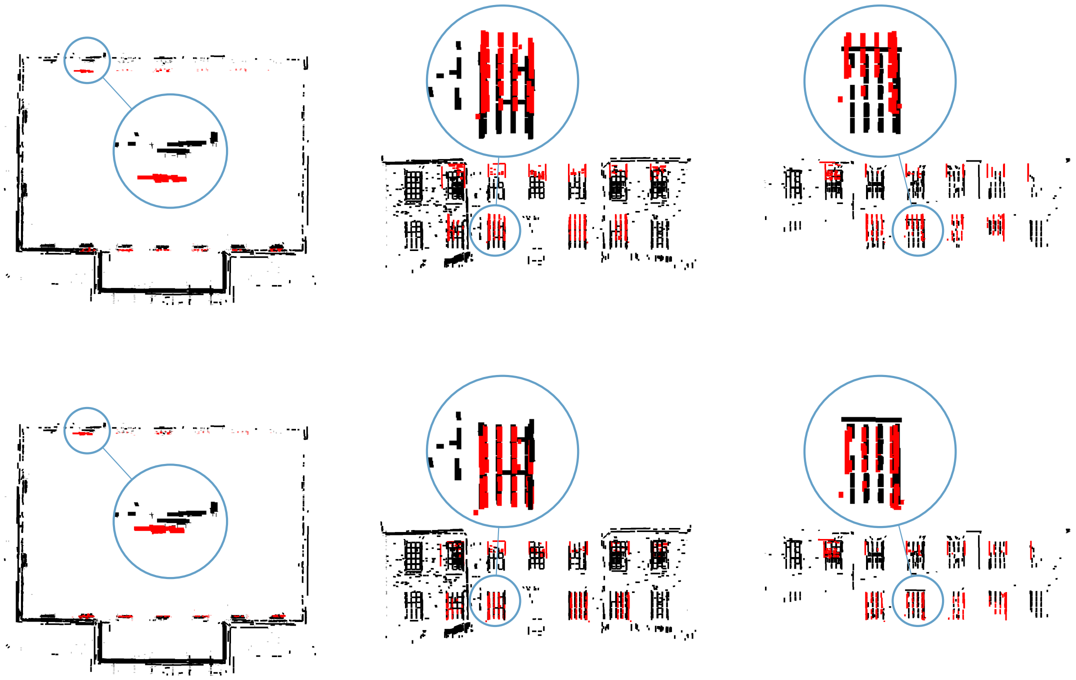
(b) Front view

(c) Side view

(d) Before 3D optimization

(e) After 3D optimization

Figure 4.17: Alignment result for the *TUM* dataset. 3D lines representing the transformed indoor model (—) and outdoor model (—) (a-c). Top-rank matching result includes 400 line matches out of 15*k* and 28*k* 3D lines of the indoor and outdoor model, respectively (d). Mean perpendicular distance between corresponding 3D lines after optimization is 4.5 cm (e)

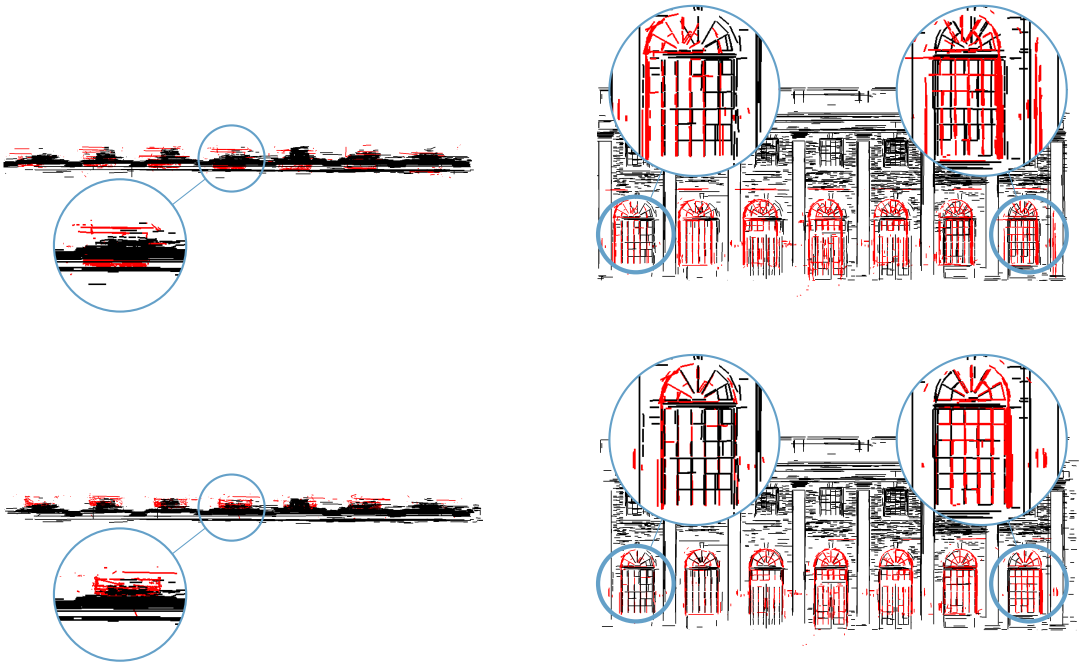


(a) Top view

(b) Front view

(c) Backside view

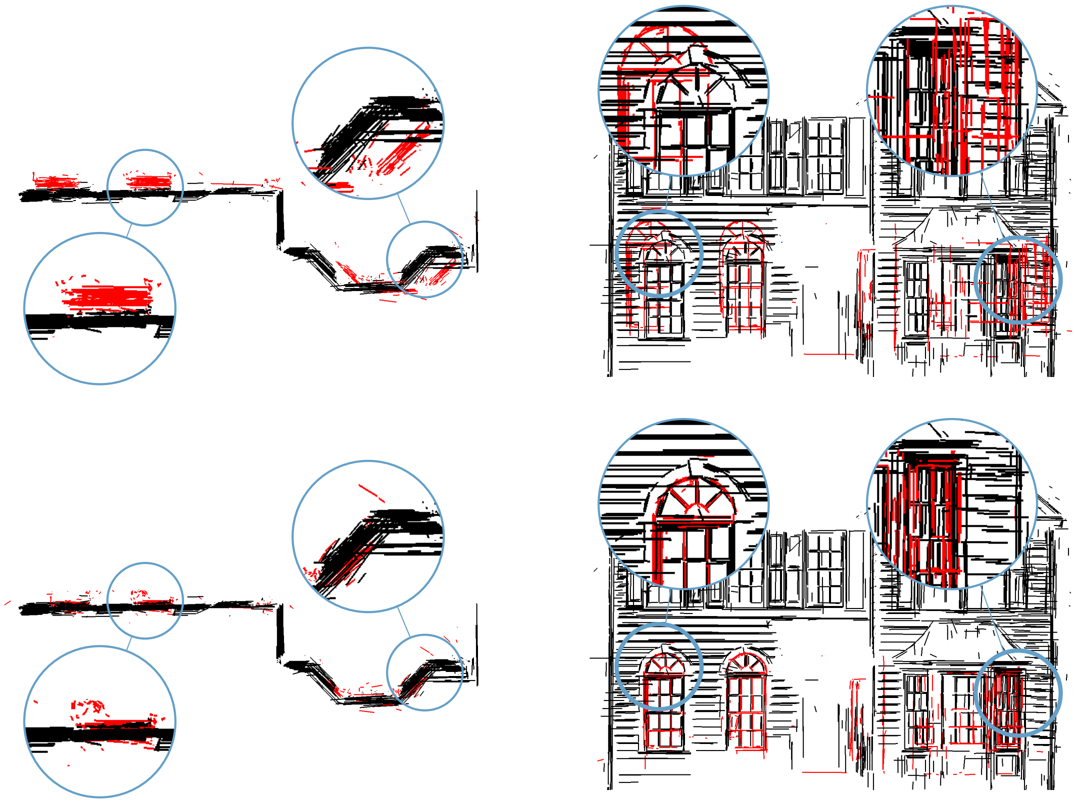
Figure 4.18: Alignment of indoor (🟥) and outdoor (⬛) models for the *Theatre* dataset (Cohen et al., 2016). Results derived from Cohen et al. (2016) (top) and the proposed method (bottom)



(a) Top view

(b) Front view

Figure 4.19: Alignment of indoor (🟥) and outdoor (⬛) models for the *Hall* dataset (Cohen et al., 2016). Results derived from Cohen et al. (2016) (top) and the proposed method (bottom)



(a) Top view

(b) Front view

Figure 4.20: Alignment of indoor (—) and outdoor (—) models for the *House-1* dataset (Cohen et al., 2016). Results derived from Cohen et al. (2016) (top) and the proposed method (bottom)

4.4 EVALUATION OF CNN-BASED SINGLE-IMAGE DEPTH ESTIMATION METHODS

Appendix D presents a new and holistic way of evaluating current advanced deep learning-based SIDE methods. Considering the developments presented in Section 3.4, the proposed evaluation scheme was designed to examine the potential of replacing complex and computationally expensive multi-view 3D vision methods with current single-image machine learning techniques. Related to UAV-based building reconstruction, such methods could facilitate the generation of a DSM from a single image without the requirement to acquire, register and, densely match multiple images. Such DSMs could be used as proxy models for subsequent trajectory planning. In addition, SIDE could help to generate 3D reconstructions of hardly accessible indoor environments and fill possible data gaps. In order to assess the applicability of such methods, meaningful evaluation is still lacking validity, comprising error-prone benchmark datasets and inadequate evaluation metrics. While recent approaches started to tackle individual characteristics of depth maps, such as accurate and sharp depth discontinuous and planar predictions of flat scene objects, established error metrics solely consider global statistics, which can hardly recover apparent distinct local differences in the depth map quality. As a result, several works faced problems to substantiate their improvements and had to rely on visual comparisons with other approaches (Hao et al., 2018; Hu et al., 2019; Liu et al., 2018). A more comprehensive and quantitative evaluation protocol is therefore needed for assisting further research in this field. At the same time, an investigation on the general performance of such methods could yield revealing insights about the applicability of SIDE methods and their potential to be used for UAV-based 3D building reconstruction tasks.

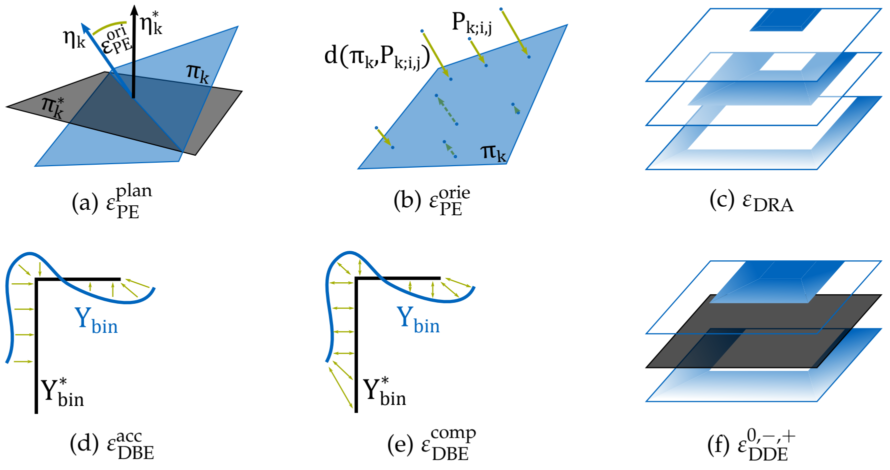
A summary of the contribution is as follows:

- Development of meaningful and geometrically interpretable evaluation metrics for SIDE methods
- Design and acquisition of a high-quality RGB-depth (RGB-D) benchmark dataset in accordance with the proposed error metrics
- Comprehensive analysis of the performance of current state-of-the-art SIDE methods

4.4.1 *Comparison of Established and Proposed Evaluation Metrics*

Evaluating the performance of SIDE methods requires accurate ground truth depth maps and error metrics that assess the discrepancy between the prediction and the reference depth map. Although depth maps reflect the complex 3D geometry of the captured scene, established error metrics merely allow for global assessments about the accuracy of predictions. Important geometric properties of depth maps, which are decisive for practical applications, are, however, largely overlooked. The following global statistics between a predicted depth map $\boldsymbol{Y}$ and its ground truth depth image $\boldsymbol{Y}^*$ with T depth pixels are considered as established metrics:

Absolute relative difference: $\operatorname{rel}(\boldsymbol{Y}, \boldsymbol{Y}^*) = \frac{1}{T}\sum_{i,j}\left|y_{i,j} - y^*_{i,j}\right| / y^*_{i,j}$



(a) ε_{PE}^{plan} (b) ε_{PE}^{orie} (c) ε_{DRA}

(d) ε_{DBE}^{acc} (e) ε_{DBE}^{comp} (f) $\varepsilon_{DDE}^{0,-,+}$

Figure 4.21: Visualizations of the proposed error metrics. The flatness and orientations of predicted planar regions can be evaluated with the *planarity errors* (a and b). The *distance-related assessment* (c) applies standard metrics for different depth range intervals. The location accuracy and completeness of depth discontinuities is rated by the *depth boundary errors* (d and e), while the consistency of depth predictions with respect to a virtual depth plane can be assessed with the *directed depth errors* (f)

Squared relative difference: $\text{srel}\left(\boldsymbol{Y}, \boldsymbol{Y}^*\right) = \frac{1}{T}\sum_{i,j}\left|y_{i,j} - y^*_{i,j}\right|^2 / y^*_{i,j}$

RMS (linear): $\text{RMS}\left(\boldsymbol{Y}, \boldsymbol{Y}^*\right) = \sqrt{\frac{1}{T}\sum_{i,j}\left|y_{i,j} - y^*_{i,j}\right|^2}$

RMS (log): $\log\left(\boldsymbol{Y}, \boldsymbol{Y}^*\right) = \sqrt{\frac{1}{T}\sum_{i,j}\left|\log y_{i,j} - \log y^*_{i,j}\right|^2}$

Threshold: percentage of $\boldsymbol{Y}$ such that $\max\left(\frac{y_i}{y_i^*}, \frac{y_i^*}{y_i}\right) = \sigma < thr$

However, the following important quality criteria should be addressed within a holistic evaluation:

- Flatness of planar surfaces and correctness of estimated plane orientations for planar image regions.
- Precision and completeness of reconstructed depth discontinuities
- Consistency of ordinal depth relations of spatially separated objects
- Consideration of the absolute scene depth for which a methodology provides trustworthy predictions

A series of error metrics, as illustrated in Figure 4.21, was specifically designed addressing these properties, comprised of *planarity errors (PE)*, *depth boundary errors (DBE)*, *directed depth errors (DDE)*, and a *distance-related assessment (DRA)*. Additional annotations are required for each benchmark RGB-D image in order to allow for detailed investigations with the proposed error metrics, such as masks of planar surfaces depicted in the images and accurate boundaries of depth discontinuities. The computation of PEs includes the generation of ground truth 3D planes $\pi_k^* = \left(\boldsymbol{\eta}_k^*, o_k\right)$ of a masked planar surface k, comprised of a normal vector $\boldsymbol{\eta}_k^*$ and an offset to the origin o. The projection of the masked predicted depth map into 3D points $\boldsymbol{P}_{k;i,j}$

enables a robust determination of a 3D plane $\boldsymbol{\pi}_k = (\boldsymbol{\eta}_k, o_k)$, allowing an assessment in terms of the planarity of the predicted 3D plane $\varepsilon_{\text{PE}}^{\text{plan}}$ and deviation of the estimated 3D plane orientation towards the reference plane $\varepsilon_{\text{PE}}^{\text{orie}}$. The calculation of DBEs exploits the occurrence of distinct depth discontinuities obtained by edge extraction in the depth maps. Ground-truth depth map edges $\boldsymbol{Y}_{\text{bin}}^{*}$ are compared to extracted edges in the predicted depth map $\boldsymbol{Y}_{\text{bin}}$ via truncated Chamfer distance, yielding a measure of the precision $\varepsilon_{\text{DBE}}^{\text{acc}}$ and the completeness $\varepsilon_{\text{DBE}}^{\text{comp}}$ of reconstructed depth contours. Specifically, a Euclidean distance transform is applied to the predicted and ground truth edge image $\boldsymbol{E} = DT\left(\boldsymbol{Y}_{\text{bin}}\right)$ and $\boldsymbol{E}^{*} = DT\left(\boldsymbol{Y}_{\text{bin}}^{*}\right)$, respectively, allowing for an efficient shape matching scheme, while distances exceeding a given threshold are truncated to a maximum distance. The determination of DDEs requires an orthogonal reference depth plane in the viewing direction of the camera $\boldsymbol{\pi}_d^{*}$ at a defined distance. The predicted depth values are assessed in terms of their ordinal relation to this reference plane, yielding underestimated, overestimated, and correctly estimated depths as $\varepsilon_{\text{DDE}}^{-}$, $\varepsilon_{\text{DDE}}^{+}$, and $\varepsilon_{\text{DDE}}^{0}$.

The proposed error metrics can be summarized as follows:

PE (flatness): $\varepsilon_{\text{PE}}^{\text{plan}}\left(\boldsymbol{Y}_k\right) = \mathbb{V}\left[\sum_{\boldsymbol{P}_{k;i,j} \in \mathcal{P}_k} d\left(\boldsymbol{\pi}_k, \boldsymbol{P}_{k;i,j}\right)\right]$

PE (orientation): $\varepsilon_{\text{PE}}^{\text{orie}}\left(\boldsymbol{Y}_k, \boldsymbol{\pi}_k^{*}\right) = \text{acos}\left(\boldsymbol{\eta}_k^{\top} \cdot \boldsymbol{\eta}_k^{*}\right)$

DBE (accuracy): $\varepsilon_{\text{DBE}}^{\text{acc}}(\boldsymbol{Y}_{\text{bin}}, \boldsymbol{Y}_{\text{bin}}^{*}) = \frac{1}{\sum_i \sum_j y_{\text{bin};i,j}} \sum_i \sum_j e_{i,j}^{*} \cdot y_{\text{bin};i,j}$

DBE (completeness): $\varepsilon_{\text{DBE}}^{\text{comp}}(\boldsymbol{Y}_{\text{bin}}, \boldsymbol{Y}_{\text{bin}}^{*}) = \frac{1}{\sum_i \sum_j y_{\text{bin};i,j}^{*} + y_{\text{bin};i,j}} \sum_i \sum_j e_{i,j}^{*} \cdot y_{\text{bin};i,j} + e_{i,j} \cdot y_{\text{bin};i,j}^{*}$

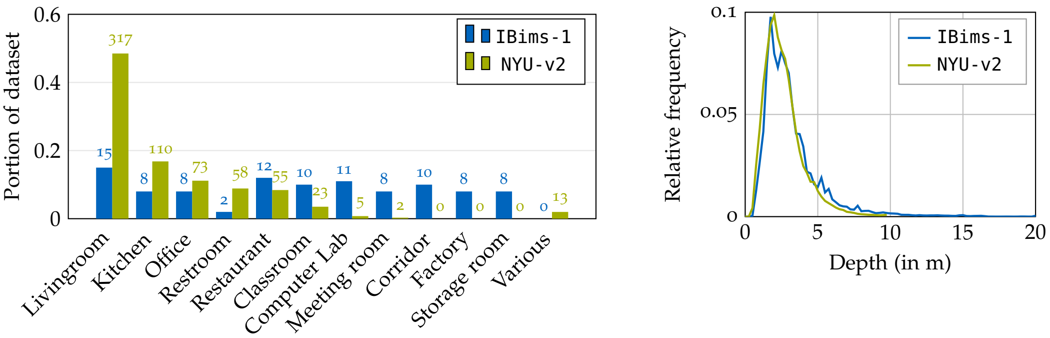
DDE (correct): $\varepsilon_{\text{DDE}}^{0}\left(\boldsymbol{Y}, \boldsymbol{Y}^{*}, \boldsymbol{\pi}_d^{*}\right) = \left|\left\{y_{i,j} | d_{\text{sgn}}(\boldsymbol{\pi}_d^{*}, \boldsymbol{P}_{i,j}) = 0 \land d_{\text{sgn}}(\boldsymbol{\pi}_d^{*}, \boldsymbol{P}_{i,j}^{*}) = 0\right\}\right| / T$

DDE (overestimated): $\varepsilon_{\text{DDE}}^{+}\left(\boldsymbol{Y}, \boldsymbol{Y}^{*}, \boldsymbol{\pi}_d^{*}\right) = \left|\left\{y_{i,j} | d_{\text{sgn}}(\boldsymbol{\pi}_d^{*}, \boldsymbol{P}_{i,j}) > 0 \land d_{\text{sgn}}(\boldsymbol{\pi}_d^{*}, \boldsymbol{P}_{i,j}^{*}) < 0\right\}\right| / T$

DDE (underestimated): $\varepsilon_{\text{DDE}}^{-}\left(\boldsymbol{Y}, \boldsymbol{Y}^{*}, \boldsymbol{\pi}_d^{*}\right) = \left|\left\{y_{i,j} | d_{\text{sgn}}(\boldsymbol{\pi}_d^{*}, \boldsymbol{P}_{i,j}) < 0 \land d_{\text{sgn}}(\boldsymbol{\pi}_d^{*}, \boldsymbol{P}_{i,j}^{*}) > 0\right\}\right| / T$

4.4.2 *Comparison of Established and Proposed RGB-D Datasets*

A customized acquisition setup was developed, comprised of a digital single-lens reflex (DSLR) camera and a high-quality laser scanner with an interchangeable custom panoramic tripod head ensuring coincidence of the optical center of the camera and the origin of the laser scanner coordinate system. The setup was used to generate a high-quality RGB-D benchmark dataset referred to as `IBims-1`, consisting of 100 aligned RGB image and depth map pairs featuring high resolution, low noise, and accurate depth discontinuous which are less affected of parallax-based occlusions compared to other datasets. The scenes were recorded in accordance with the `NYU-v2` (Silberman et al., 2012) dataset, which is still the most commonly used indoor RGB-D dataset. Both datasets exhibit a similar scene variety and depth distribution, however, `IBims-1` RGB-D image pairs provide an extended maximum scene depth and more challenging scenarios (*cf.* Figure 4.22). Further details about the acquisition process and components of the dataset, as well as a comparison towards other RGB-D datasets, are provided in the attached paper in Appendix D. An in-depth comparison of `NYU-v2` and `IBims-1` reveals the superiority of the proposed



(a) Distribution of samples for each scene type. Absolute numbers are given above

(b) Distribution of depth values

Figure 4.22: `IBims-1` dataset statistics compared to the `NYU-v2` (Silberman et al., 2012) dataset. Scene variety (a) and distribution of depth values (b)

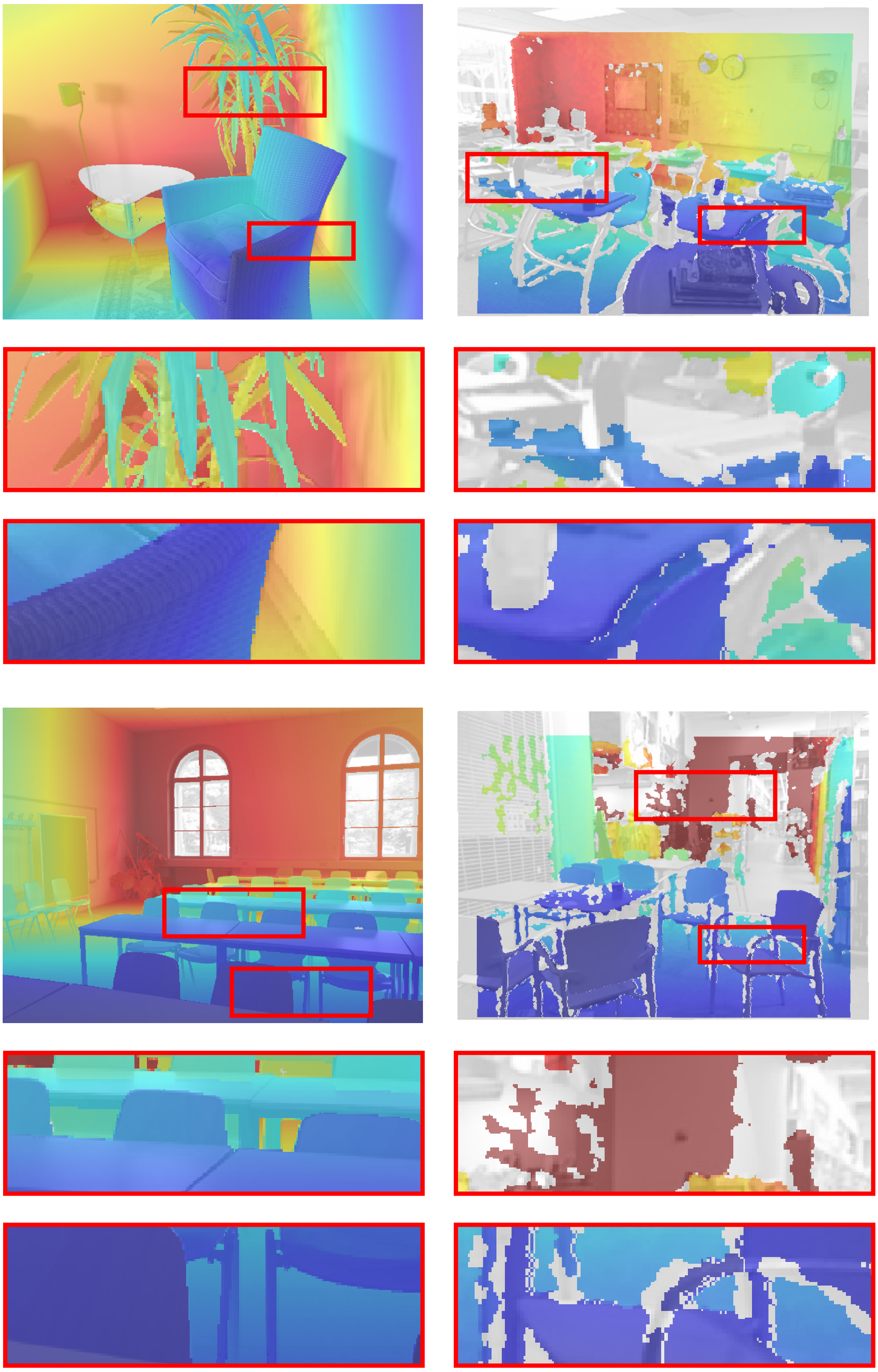
dataset, showing extremely accurate depth values and complete depth maps with less occlusion, allowing for comprehensive performance analysis of SIDE methods (*cf.* Figure 4.23).

4.4.3 *Evaluation and Analysis of Single-image Depth Estimation Methods*

A comprehensive analysis applying current state-of-the-art methodologies and proposed error metrics on the `IBims-1` dataset was conducted, revealing an in-depth performance comparison and insights into the behavior of convolutional neural network (CNN)-based SIDE methods. Quantitative results obtained from the investigated methods for `NYU-v2` and `IBims-1` are listed in Table 4.4. The following section summarizes the discovered findings and drawn conclusions with the help of selected samples. A more detailed description of the conducted experiments and comparison is provided in the attached paper in Appendix D.

Established Error Metrics indicate the comparability of the `IBims-1` dataset with the `NYU-v2` dataset, which was exclusively used for training the investigated methods. Although the acquisition of a novel dataset using alternative sensors naturally leads to a domain gap, the acquisition of `IBims-1` has focused on a comparability with `NYU-v2`, as reflected by similar camera parameters, scene depth, and scene variety. A comparison of global error statistics confirms the validity of the proposed dataset, as the results show slightly less accurate global errors with a comparable ranking of the methods. However, more challenging scenes with partially extended maximum scene depths lead to a slight drop of achieved accuracy, indicating modest overfitting effects on `NYU-v2` amongst others. Nevertheless, established global error metrics do not reveal certain properties of the predicted depth maps, and the discrimination of the methods is limited due to the proximity of the results.

Distance-related Assessment (DRA) allows a more in-depth analysis of the performance with regard to different scene depths. The results have shown a strong association between accuracy and the distribution of scene depths in the `NYU-v2` training dataset, as shown in Figures 4.22b and 4.24, and clearly demonstrate the effects of such imbalanced training datasets. Although differences in the investigated



(a) `IBims-1`

(b) `NYU-v2`

Figure 4.23: Visualization of registration accuracy and depth completeness of `IBims-1` (a) and `NYU-v2` (b). Overlay of grayscale RGB images and colored depth maps for various samples (invalid or missing depth values are depicted in grey). Top: full image. Middle and bottom row: detailed views

Table 4.4: Quantitative results for standard metrics on `NYU-v2` and standard metrics, proposed PE, DBE, and DDE metrics on `IBims-1` applying different SIDE methods (**best**, **second best**). Higher the better for ↑ and lower the better for ↓

Method	**Dataset**	**Standard Metrics ($\sigma_i = 1.25^i$)**						**PE (cm/°)**		**DBE (px)**		**DDE (%) for d =3 m**		
		rel ↓	$\log_{10}$ ↓	RMS ↓	σ_1 ↑	σ_2 ↑	σ_3 ↑	$\varepsilon_{\text{PE}}^{\text{plan}}$ ↓	$\varepsilon_{\text{PE}}^{\text{orie}}$ ↓	$\varepsilon_{\text{DBE}}^{\text{acc}}$ ↓	$\varepsilon_{\text{DBE}}^{\text{comp}}$ ↓	$\varepsilon_{\text{DDE}}^{0}$ ↑	$\varepsilon_{\text{DDE}}^{-}$ ↓	$\varepsilon_{\text{DDE}}^{+}$ ↓
Eigen et al. (2014)	NYU-v2	0.22	0.09	0.76	0.61	0.89	0.97	—	—	—	—	—	—	—
Eigen and Fergus (2015) (AlexNet)	NYU-v2	0.19	0.08	0.67	0.69	0.91	**0.98**	—	—	—	—	—	—	—
Eigen and Fergus (2015) (VGG)	NYU-v2	0.16	**0.07**	0.58	0.75	0.95	**0.99**	—	—	—	—	—	—	—
Laina et al. (2016)	NYU-v2	**0.14**	**0.06**	**0.51**	**0.82**	0.95	**0.99**	—	—	—	—	—	—	—
Liu et al. (2015)	NYU-v2	0.21	0.09	0.68	0.66	0.91	**0.98**	—	—	—	—	—	—	—
Li et al. (2017)	NYU-v2	**0.15**	**0.06**	0.53	0.79	**0.96**	**0.99**	—	—	—	—	—	—	—
Liu et al. (2018)	NYU-v2	**0.14**	**0.06**	**0.51**	0.81	**0.96**	**0.99**	—	—	—	—	—	—	—
Ramamonjisoa and Lepetit (2019)	NYU-v2	**0.14**	**0.06**	**0.46**	**0.84**	**0.97**	**0.99**	—	—	—	—	—	—	—
Eigen et al. (2014)	IBims-1	0.32	0.17	1.55	0.36	0.65	0.84	7.70	24.91	9.97	9.99	70.37	27.42	2.22
Eigen and Fergus (2015) (AlexNet)	IBims-1	0.30	0.15	1.38	0.40	0.73	0.88	7.52	21.50	4.66	8.68	77.48	18.93	3.59
Eigen and Fergus (2015) (VGG)	IBims-1	**0.25**	**0.13**	1.26	0.47	0.78	**0.93**	**5.97**	**17.65**	4.05	8.01	79.88	18.72	**1.41**
Laina et al. (2016)	IBims-1	0.26	**0.13**	1.20	0.50	0.78	0.91	**6.46**	19.13	6.19	9.17	81.02	17.01	1.97
Liu et al. (2015)	IBims-1	0.30	**0.13**	1.26	0.48	0.78	0.91	8.45	28.69	**2.42**	**7.11**	79.70	14.16	6.14
Li et al. (2017)	IBims-1	**0.22**	**0.11**	**1.09**	**0.58**	**0.85**	**0.94**	7.82	22.20	3.90	8.17	**83.71**	**13.20**	3.09
Liu et al. (2018)	IBims-1	0.29	0.17	1.45	0.41	0.70	0.86	7.26	**17.24**	4.84	8.86	71.24	28.36	**0.40**
Ramamonjisoa and Lepetit (2019)	IBims-1	0.26	**0.11**	**1.07**	**0.59**	**0.84**	**0.94**	9.95	25.67	**3.52**	**7.61**	**84.03**	**9.48**	6.49

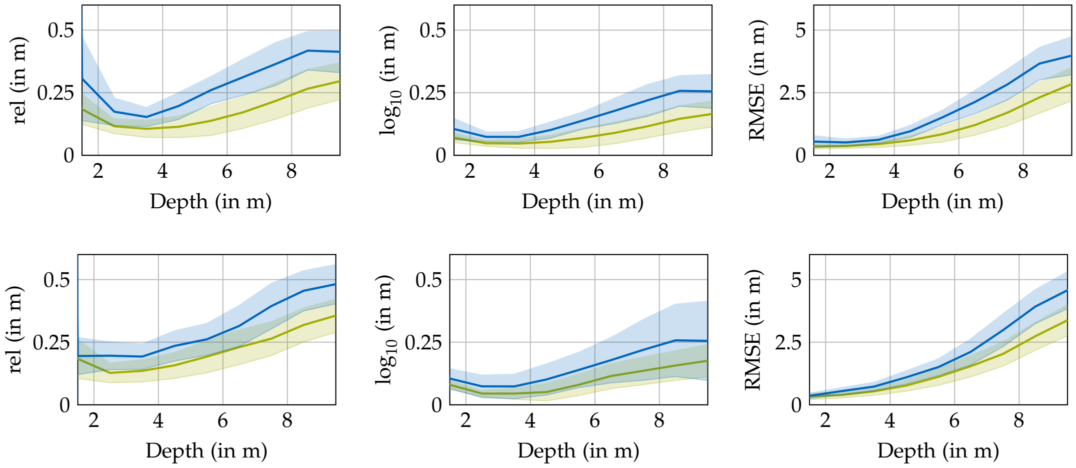
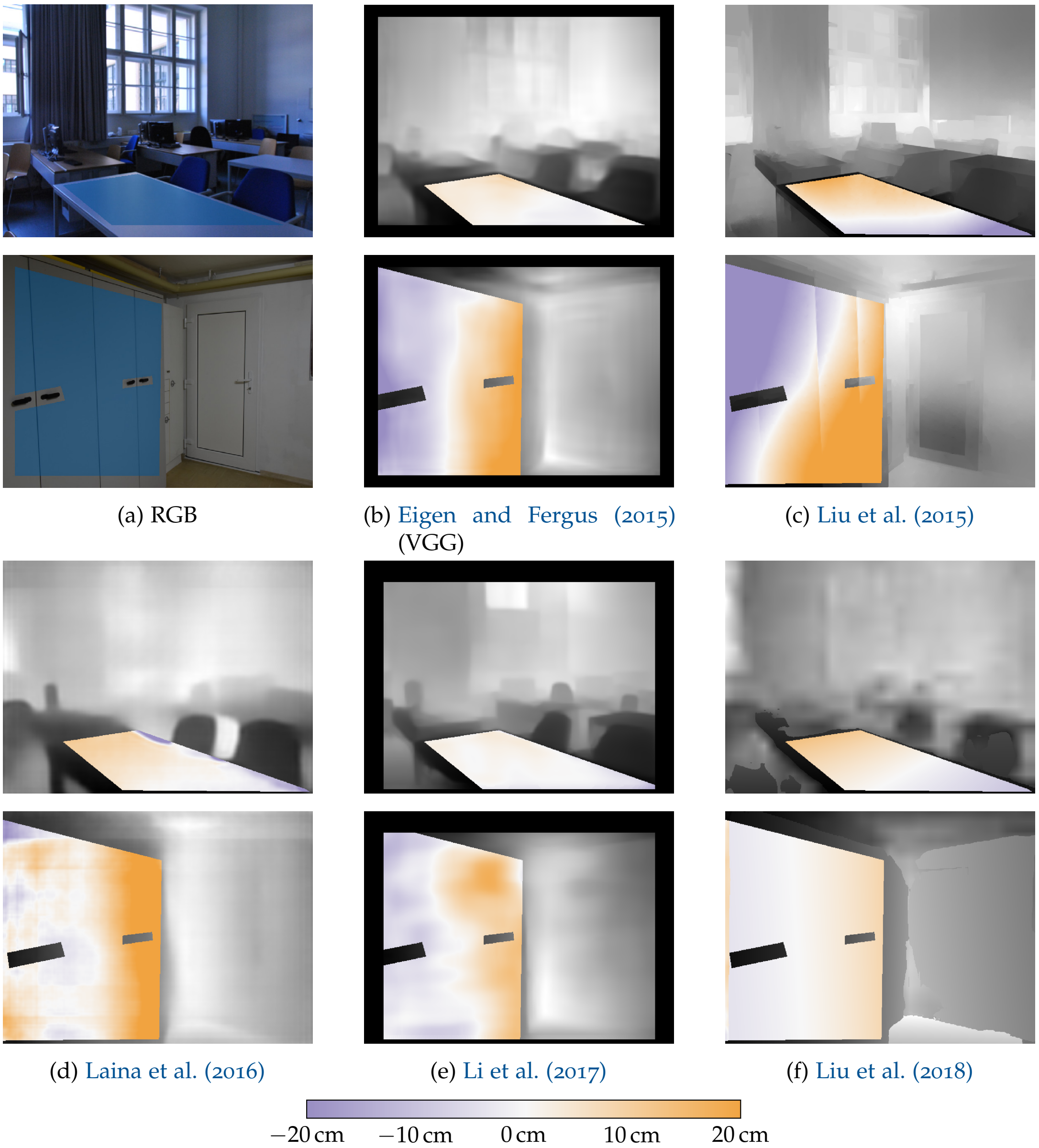


Figure 4.24: Distance-related global errors (left: relative error; mid: $\log_{10}$ error and right: root mean square error (RMSE)) for the shared depth range of NYU-v2 (mean: —,±0.5 std: 🟩) and IBims-1 (mean: —, ±0.5 std: 🟦) using the methods of Ramamonjisoa and Lepetit (2019) (top) and Li et al. (2017) (bottom)

methods could be revealed, the reliability of all methods substantially decreases with increased scene depths leading to large errors already for relatively short distances at 5 m. Further attempts for improving the performance of SIDE methods should address this challenge by introducing balanced training datasets or sophisticated ways to tackle this imbalance. A first attempt has been made by Jiao et al. (2018) in proposing an attention-driven loss for the network supervision that particularly improves depth predictions for distant regions.

Planarity Errors (PE) have exposed a surprisingly poor ability of the methods for predicting planar image regions. Although visual inspections of predicted depth maps suggest the planarity of planar objects such as walls or floors, quantitative results uncover large discrepancies, particularly in predicting the correct orientation of 3D planes disclosing deviations between 17° and 28°. This weakness is significantly less severe for horizontal planes such as floors and table tops, however, a reliable estimation of vertical planes in images could not be observed in the experiments, affecting the applicability of SIDE methods in practice. An illustration of the reconstruction of planar image regions with different methods is exemplarily shown in Figure 4.25. However, since some methods have focused on this specific task by usually performing a preceding segmentation of planar image regions and estimating 3D plane parameters for contiguous planar segments, such as this is the case for the method of Liu et al. (2018), the performance substantially increased, reflected by lower planarity errors. Nevertheless, segmentation errors have a direct impact on the quality of planar regions, often leading to severe errors when planar regions are missed or split into different segments.

Depth Boundary Errors (DBE) investigate the precision and completeness of reconstructed depth discontinuities. CNN-based methods often tend to produce smooth depth maps due to strided convolutions and spatial pooling operations in the network designs. This loss of local details results in a failure to reconstruct detailed depth boundaries, leading to spurious scene geometries. Figure 4.26 compares the



(a) RGB

(b) Eigen and Fergus (2015) (VGG)

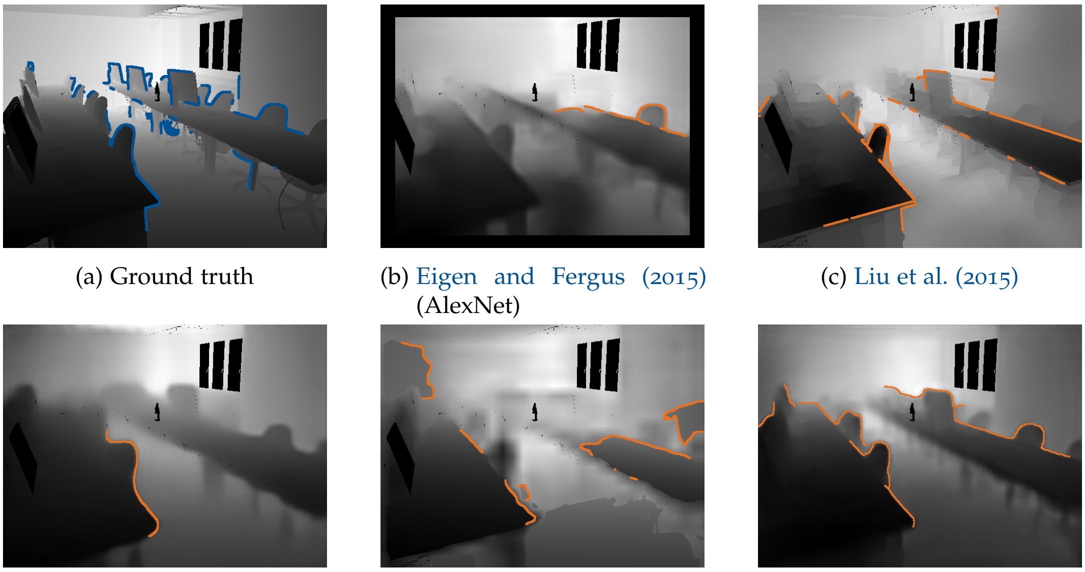
(c) Liu et al. (2015)

(d) Laina et al. (2016)

(e) Li et al. (2017)

(f) Liu et al. (2018)

Figure 4.25: Visual results after applying *planarity errors* (PEs) on different planar regions (top: table, bottom: wall). RGB with corresponding plane masks (🟦) (a). Predictions using different methodologies (b-f). Colors in the predictions correspond to orthogonal differences of projected depths towards the reference plane



(a) Ground truth

(b) Eigen and Fergus (2015) (AlexNet)

(c) Liu et al. (2015)

(d) Laina et al. (2016)

(e) Liu et al. (2018)

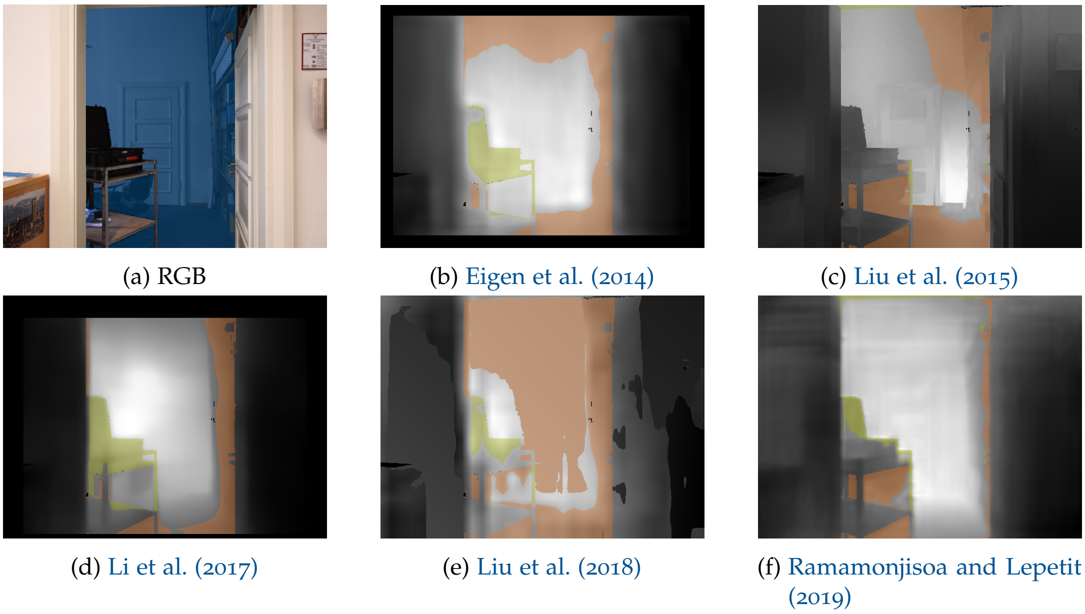
(f) Ramamonjisoa and Lepetit (2019)

Figure 4.26: Visual results after applying *depth boundary errors* (DBEs) on `IBims-1`. Overlay of ground truth depth map with ground truth edge (—) (a) and depth map predictions with extracted edges (colored) using different methods (b-f)

extracted depth discontinuities obtained from different methods. Approaches that particularly address this problem, such as the methods of Liu et al. (2015) and Ramamonjisoa and Lepetit (2019), evidently improve the precision and completeness of reconstructed depth discontinuities at the cost of falsely predicting depth at textured planar regions. This investigation suggests that gradients seem to serve as strong hints to the networks. Therefore, distinguishing between intensity changes due to real depth discontinuities and solely texture is still a major challenge in SIDE.

Directed Depth Errors (DDE) aim to identify predicted depth values that lie on the correct side of a predefined reference plane but also distinguishes between overestimated and underestimated predicted depths. The results have shown significant differences among the investigated methods, ranging from 70 % to 84 % correctly estimated depth values for a virtual reference plane at 3 m distance. Less accurate methods tended to underestimate depth rather than predicting depth overly far. While predictions at large object boundaries of the room layout, such as door frames, were predicted coherently, smaller objects, such as furniture, often encountered problems leading to a deteriorated scene geometry. Figure 4.27 illustrates the performance of selected methods on a sample image of `IBims-1`.

Geometric and Radiometric Augmentations of `IBims-1` were generated for investigating the robustness of the methods. While numerous radiometric augmentations, such as brightness, contrast, saturation, and color channel swaps, did not significantly influence the methods' performance, horizontal flips of the images led to a significant decrease in the accuracy. This finding suggests that the visibility of distinct scene structures such as floors or grounds on the bottom part of the image represents important scene priors for the networks. Complementing the conducted experiments



(a) RGB

(b) Eigen et al. (2014)

(c) Liu et al. (2015)

(d) Li et al. (2017)

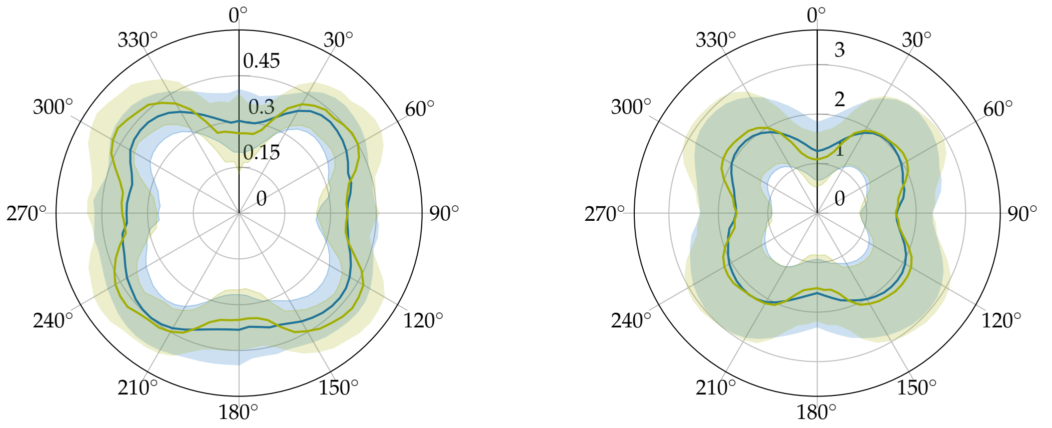
(e) Liu et al. (2018)

(f) Ramamonjisoa and Lepetit (2019)

Figure 4.27: Visual results after applying *directed depth errors* (DDEs) on `IBims-1`. Ground truth depth plane at $d = 3\,\text{m}$ separating foreground from background (🟦) (a). Differences between ground truth and predictions (b-f). Color coded are depth values that are either estimated too short (🟧) or too far (🟩)

in the paper, an additional experiment on various discrete image rotations has revealed the severity of such priors. Figure 4.28 demonstrates that already slight image rotations have led to substantial drops in the accuracy of the predictions. It is worth noting that the method of Li et al. (2015) did not use augmented rotated images for training their method, while minor augmentations of slight image rotations included in the training of the network proposed by Ramamonjisoa and Lepetit (2019) has not significantly improved the quality as well. While original upright images yielded best results, orthogonal rotations of 90°, 180°, and 270° have not worsened the results as severe as diagonal rotations, as best seen for the RMSE in Figure 4.28b. Since images of the `IBims-1` dataset do not exhibit rotations along the roll axis, horizontal and vertical surfaces, which dominate indoor scenes, remain image-axis aligned for orthogonal rotations. This confirms the assumption that CNN-based SIDE methods learn structural scene priors, such as horizontal ground planes and ceilings and vertical walls, windows, and doors, which guide the scene depth estimation. This examination provides insights into the learning process of CNN-based methods and highlights a remaining challenge in this field. Evaluating deteriorated images by adding different amounts of noise and blur to the input images has led to decreased qualities of the predictions for all methods. Specifically, the methods started to substantially react on these deteriorations for 10 % of Salt and Pepper noise, Gaussian noise with a variance of 0.01, and a Gaussian blur with $\sigma > 2$.

Textured Planar Surfaces and Variations in Scene Illumination were additionally investigated with an auxiliary part of the `IBims-1` dataset in order to gain deeper insights into the networks' behavior. By capturing close-up images of different patterns and pictures hang on a clearly visible wall, an in-depth analysis of the influence of gradients and illusory scene depth for SIDE methods could be

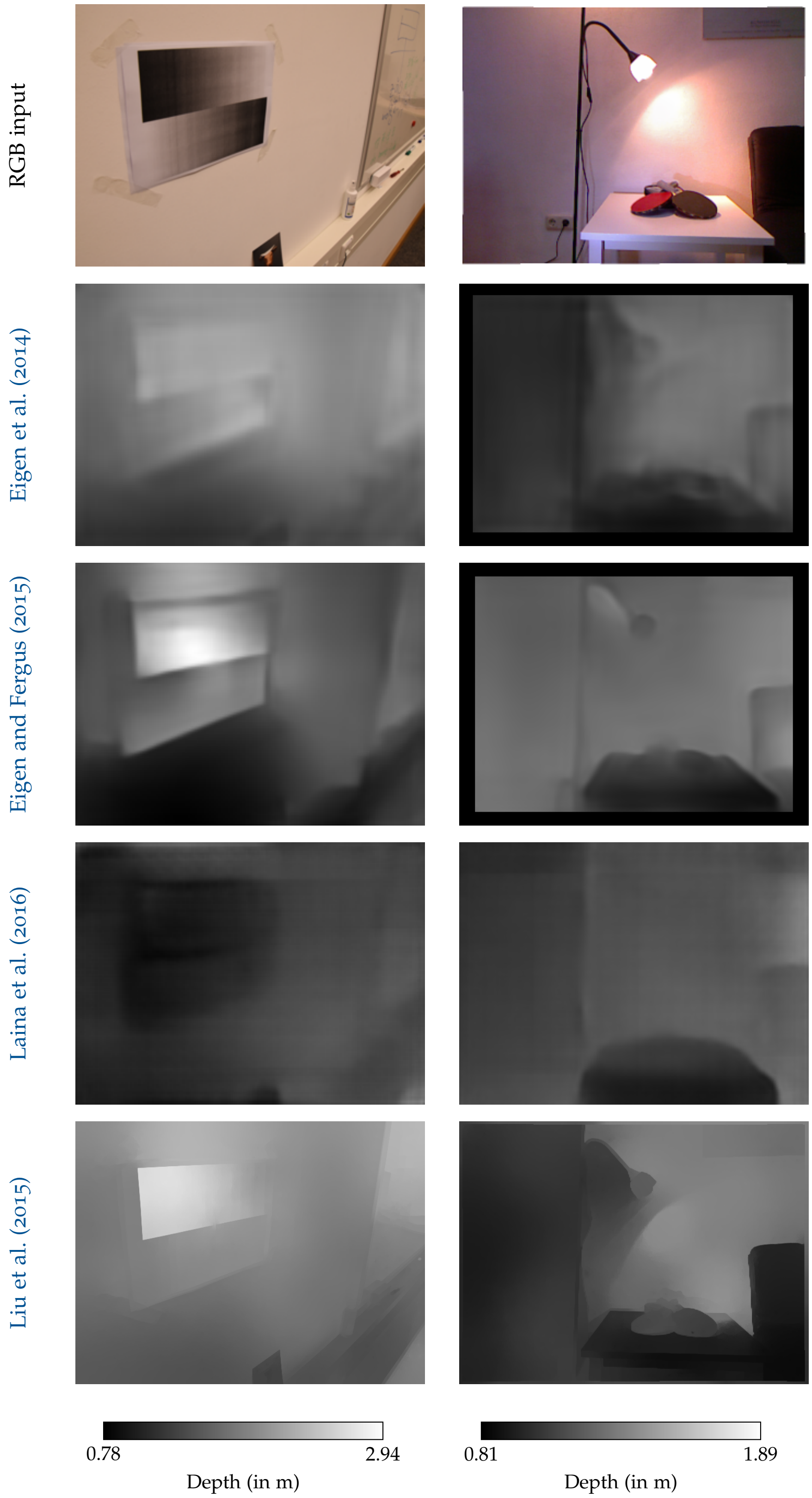


(a) Absolute relative error (in m)

(b) RMSE (in m)

Figure 4.28: Influence of image rotation on the performance of SIDE methods exemplary shown for the methods of Liu et al. (2015) (▬) and Ramamonjisoa and Lepetit (2019) (▬) using absolute relative error (a) and RMSE (b). Average errors (▬,▬) and ±0.5 standard deviation (■,■) on the `IBims-1` dataset

conducted. The results have shown that all investigated methods reacted to these planar textured patterns by predicting a non-existent depth in accordance with the gradients (*cf.* Figure 4.29a). This confirms the assumption that gradients and texture serve as strong hints for the network, however, the effect is unexpected severe and should be considered when applying such methods in practice, in particular for navigating autonomous robots. A similar finding has been revealed by comparing depth predictions of the same scene under various lighting conditions, ranging from natural to diffuse and directed artificial lighting. In particular, directional lighting caused severe changes in the predictions, as shown in the bottom row of Figure 4.29b, which again can be explained by gradients occurred from strong shadows on the wall. Although applications involving images that capture artificial patterns on a wall rarely seem to occur in practice, scene lighting is omnipresent and crucial for the quality of the predictions.



(a) Patterns

(b) Illumination

Figure 4.29: Predictions for a printed sample from the Pattern dataset (Asuni and Giachetti, 2014) on a planar surface (a) and for an artificially illuminated scene (b). Predictions using different methods (rows) of the input images (first row). Predicted depth maps are color-coded according to the colormaps shown in the last row

5 CONCLUSION

This thesis addressed different challenges in using unmanned aerial vehicle (UAV) imagery for photogrammetric applications and contributed towards the improvement, automation, and safety of photogrammetric survey campaigns, as well as for further processing of building models towards the generation of level of detail (LoD)-4 building models. The following chapter concludes with a summary of the main contributions in Section 5.1, an investigation of the applicability of the proposed methodologies in related UAV-based remote sensing fields (Section 5.2), and thoughts on potential future developments in Section 5.3.

5.1 SUMMARY

Although UAV-photogrammetry is already at an advanced stage of research and has proven its maturity for 3D modeling, witnessed by increasing popularity in scientific and industrial fields, the challenges addressed in this thesis have persisted.

The following contributions from this thesis can be drawn:

- An image-based georeferencing methodology for UAV images was proposed to accurately register UAV images towards aerial reference images. Based on an in-depth analysis of the matching failure using current feature-based methods for this task, the revealed bottlenecks include differences in image scale and rotation, the discriminative performance of feature descriptors, and the misuse of the ratio-test. A novel feature-based image matching strategy addresses the identified limitations by proposing a superpixel-based dense feature extraction strategy, a one-to-many feature matching scheme, and a global geometric verification strategy that allows to geo-register UAV imagery with pixel-level accuracy towards georeferenced image data, such as aerial images, orthophotos, or satellite images
- Integration of semantic cues into automatic 3D UAV path planning was presented, demonstrating the possibility of casting expert knowledge into algorithms for designing optimal trajectories for multi-view stereo (MVS) image acquisition, such as integrating photogrammetric properties and respecting flight safety by avoiding potential hazardous flight zones, such as roads, water basins, streets, or cars. The proposed model-based path planning strategy exploits a semantically-enriched 3D proxy model of the environment that defines accessible, partially accessible, and prohibited flight areas. The proposed methodology discretely optimizes for a short UAV image acquisition path allowing for detailed and complete 3D reconstruction models in an intended model resolution using standard 3D reconstruction pipelines on the acquired images
- An automatic methodology for aligning individual indoor and outdoor building models was proposed that focuses on accurate geometric registration of shared building parts, such as window frames and doors. A 3D-line based scene

representation allows to define multiple matching façade hypotheses between indoor and outdoor models, and a 2D binary matching scheme utilizing Chamfer distance proposes a set of transformation hypotheses. A refinement of the registration based on corresponding 3D line segments finally yields centimeter-level registration accuracies, leading to generate complete 3D building models in accordance with LoD-4 standards

- A novel evaluation protocol for assessing the performance of single-image depth estimation (SIDE) methods was proposed, comprising geometrically-relevant evaluation metrics and a high-quality indoor RGB-depth (RGB-D) dataset. The developed error metrics allow for obtaining reliable assessments about the suitability and the expected accuracy of such methods for certain applications. A comprehensive analysis of current SIDE methods has identified potential application areas and remaining challenges that should be addressed in further research in this field. The results have shown that the preservation of sharp depth discontinuities and planarity of actual planar image regions tend to be mutually contradictory due to the networks' sensitivity towards image gradients. The availability of merely imbalanced training datasets in terms of depth distribution is reflected in the weak performance in estimating distant regions

5.2 APPLICABILITY IN RELATED FIELDS

Although this thesis focused on the application of reconstructing as-built 3D building models, the proposed methods contribute to a much broader range of application fields. As already extensively outlined in Section 2.2, UAVs are currently becoming increasingly important in almost every field of remote sensing, including environmental monitoring, agriculture and forestry, cultural heritage preservation and beyond. Several of them could benefit from the proposed methodologies. A selection of various application areas is described hereafter:

- One task of *environmental monitoring* involves revealing temporal land cover changes. Depending on the temporal difference between acquisition campaigns, the mapped topology could have undergone dramatic changes in appearance. Accurate registration between the acquisition epochs is necessary for uncovering changes in the scene. Since the proposed image matching strategy has proven to be robust against temporal and radiometric changes and achieves registration accuracies at pixel-level even across different modalities, it can provide a valuable tool for this task
- A similar challenge arises when natural disasters destroy landscapes and human habitats, such as residences and infrastructure. In order to optimize *disaster management*, a quick and accurate assessment of the extend of destruction is required. Precise registration of a UAV-based orthomosaic of the affected environment with a reference map can help to rapidly localize destroyed objects and areas
- The increasing use of UAVs for assisting the excavation processes of *cultural heritages* requires high temporal acquisition frequencies for monitoring and planning purposes. Due to the enormous topological changes between the

short time periods and the spatial limitation of suitable locations for deploying ground control points (GCPs), the automated and robust image-based registration strategy could facilitate the excavation progress documentation without obstructing the excavation activities. With regard to the preservation of large-scale ancient sites, the proposed automated flight planning methodology could be applied to generate detailed 3D reconstructions of individual objects while avoiding hazardous flights above protected or currently excavated areas

- *Civil engineering* utilizes UAVs for monitoring and inspecting transportation systems, such as bridges. The proposed flight planning pipeline could be applied to generate a trajectory that allows detailed and close-up views for all parts of a bridge, while semantic restrictions can be applied to limit hazardous flight maneuvers above crowded roads

5.3 FUTURE WORK

Although the proposed methodologies have proven to enhance UAV-photogrammetry through a set of profound synthetic and real-world experiments, aspects of the approaches still can be refined while further challenges remain for subsequent research directions:

- One main limitation of the proposed image matching approach is its restriction to nadir images. A generalized approach for matching both nadir and tilted or oblique UAV images would involve the adaptation of the geometric match verification step to epipolar geometry

- Despite the relatively moderate effort required for generating a coarse proxy model of the environment for the path planning approach by acquiring a small amount of overview images, it would be desirable to reduce the initial acquisition to a single image. Despite the overwhelming progress in the field of SIDE, the applicability of these methods for UAV images is still at an early stage. In order to obtain accurate and reliable depth predictions from single images, domain adaptation as well as extreme scaling differences due to different flight altitudes pose further challenges in this field

- First efforts towards increasing the safety of UAV flights during an image acquisition campaign was initiated with the proposed semantically-aware path planning approach. Current legislation, however, often stipulate the permanent visibility of a UAV for the operating pilot. This condition should be integrated into an automated flight planning approach by *e.g.*, introducing additional visibility constraints or a joint optimization of both UAV and pilot paths

- Regarding the generation of satisfying photo-realistic 3D building models from UAV images, especially tightly built-up environments often lead to unavoidable gaps in the 3D model caused by occlusions from adjacent buildings or vegetation. A geometric completion of the building could be achieved by exploiting symmetrical features, while modern generative techniques, such as conditional generative adversarial networks (cGANs), can be used to enrich the occluded façades with realistic textures corresponding to those of the reconstructed building

BIBLIOGRAPHY

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., and Süsstrunk, S. (2012). "SLIC superpixels compared to state-of-the-art superpixel methods." IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 34(**11**), pp. 2274–2282.

Ackermann, J. and Goesele, M. (2015). "A survey of photometric stereo techniques." Foundations and Trends in Computer Graphics and Vision 9(**3-4**), pp. 149–254.

Adão, T., Hruška, J., Pádua, L., Bessa, J., Peres, E., Morais, R., and Sousa, J. (2017). "Hyperspectral imaging: a review on UAV-based sensors, data processing and applications for agriculture and forestry." Remote Sensing 9(**11**), p. 1110.

Agisoft. *Agisoft: Metashape*. `https://www.agisoft.com/`. Accessed: 2019-09-04.

Agüera-Vega, F., Carvajal-Ramírez, F., and Martínez-Carricondo, P. (2017). "Assessment of photogrammetric mapping accuracy based on variation ground control points number using unmanned aerial vehicle." Measurement 98, pp. 221–227.

Ahmed, O. S., Shemrock, A., Chabot, D., Dillon, C., Williams, G., Wasson, R., and Franklin, S. E. (2017). "Hierarchical land cover and vegetation classification using multispectral data acquired from an unmanned aerial vehicle." International Journal of Remote Sensing 38(**8-10**), pp. 2037–2052.

Ai, M., Hu, Q., Li, J., Wang, M., Yuan, H., and Wang, S. (2015). "A robust photogrammetric processing method of low-altitude UAV images." Remote Sensing 7(**3**), pp. 2302–2333.

Aicardi, I., Nex, F., Gerke, M., and Lingua, A. M. (2016a). "An image-based approach for the co-registration of multi-temporal UAV image datasets." Remote Sensing 8(**9**), p. 779.

Aicardi, I., Chiabrando, F., Grasso, N., Lingua, A. M., Noardo, F., and Spanò, A (2016b). "UAV photogrammetry with oblique images: first analysis on data acquisition and processing." International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS) XLI(**1**), pp. 835–842.

Albertz, J and Wiggenhagen, M (2009). *Guide for Photogrammetry and Remote Sensing*. Wichmann, Paderborn, Germany.

Alcantarilla, P. F., Bartoli, A., and Davison, A. J. (2012). "KAZE features." In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 214–227.

Alsadik, B., Gerke, M., and Vosselman, G. (2013). "Automated camera network design for 3D modeling of cultural heritage objects." Journal of Cultural Heritage 14(**6**), pp. 515–526.

Altwaijry, H., Veit, A., Belongie, S. J., and Tech, C. (2016). "Learning to detect and match keypoints with deep architectures." In: *Proceedings of the British Machine Vision Conference (BMVC)*.

Anwar, S., Hayder, Z., and Porikli, F. (2017). "Depth estimation and blur removal from a single out-of-focus image." In: *Proceedings of the British Machine Vision Conference (BMVC)*.

Apollonio, F., Ballabeni, A., Gaiani, M., and Remondino, F. (2014). "Evaluation of feature-based methods for automated network orientation." International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS) XL(**5**), pp. 47–54.

ArduPilot. *ArduPilot: Mission Planner*. `http://ardupilot.org/planner/`. Accessed: 2019-05-28.

Arefi, H., Engels, J., Hahn, M., and H., M. (2008). "Levels of detail in 3D building reconstruction from LiDAR data." International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS) XXXVII(**3**), pp. 485–490.

Asuni, N. and Giachetti, A. (2014). "Testimages: a large-scale archive for testing visual devices and basic image processing algorithms." In: *Smart Tools and Apps for Graphics - Eurographics Italian Chapter Conference*. The Eurographics Association, pp. 63–70.

Baig, M. H. and Torresani, L. (2016). "Coupled depth learning." In: *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–10.

Baiocchi, V., Dominici, D., and Mormile, M. (2013). "UAV application in post-seismic environment." International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS) XL(**1**), pp. 21–25.

Balsa-Barreiro, J. and Fritsch, D. (2018). "Generation of visually aesthetic and detailed 3D models of historical cities by using laser scanning and digital photogrammetry." Digital Applications in Archaeology and Cultural Heritage 8, pp. 57–64.

Bay, H., Ess, A., Tuytelaars, T., and Van Gool, L. (2008). "Speeded-up robust features (SURF)." Computer Vision and Image Understanding (CVIU) 110(**3**), pp. 346–359.

Becirevic, D., Klingbeil, L., Honecker, A., Schumann, H., Rascher, U., Léon, J., and Kuhlmann, H. (2019). "On the derivation of crop heights from multitemporal UAV based imagery." International Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS) IV(**2**), pp. 95–102.

Bejiga, M., Zeggada, A., Nouffidj, A., and Melgani, F. (2017). "A convolutional neural network approach for assisting avalanche search and rescue operations with UAV imagery." Remote Sensing 9(**2**), p. 100.

Bekele, D., Teutsch, M., and Schuchert, T. (2013). "Evaluation of binary keypoint descriptors." In: *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pp. 3652–3656.

Biljecki, F., Stoter, J., Ledoux, H., Zlatanova, S., and Çöltekin, A. (2015). "Applications of 3D city models: state of the art review." ISPRS International Journal of Geo-Information 4(**4**), pp. 2842–2889.

Bircher, A., Kamel, M., Alexis, K., Burri, M., Oettershagen, P., Omari, S., Mantel, T., and Siegwart, R. (2016). "Three-dimensional coverage path planning via viewpoint resampling and tour optimization for aerial robots." Autonomous Robots 40(**6**), pp. 1059–1078.

Bittner, K., Körner, M., Fraundorfer, F., and Reinartz, P. (2019). "Multi-task cGAN for simultaneous spaceborne DSM refinement and roof-type classification." Remote Sensing 11(**11**), p. 1262.

Blaha, M., Vogel, C., Richard, A., Wegner, J. D., Pock, T., and Schindler, K. (2016). "Large-scale semantic 3D reconstruction: an adaptive multi-resolution model for multi-class volumetric labeling." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3176–3184.

Border, R., Gammell, J. D., and Newman, P. (2018). "Surface Edge Explorer (SEE): planning next best views directly from 3D observations." In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1–8.

Brooks, C., Dobson, R. J., Banach, D. M., Dean, D., Oommen, T., Wolf, R. E., Havens, T. C., Ahlborn, T. M., Hart, B., et al. (2015). *Evaluating the use of unmanned aerial vehicles for transportation purposes*. Tech. rep. Michigan. Deptartment of Transportation. Office of Research and Best Practices.

Bundler. *Bundler: Structure from Motion (SfM) for Unordered Image Collections*. http://www.cs.cornell.edu/~snavely/bundler/. Accessed: 2019-04-20.

Buyukdemircioglu, M., Kocaman, S., and Isikdag, U. (2018). "Semi-automatic 3D city model generation from large-format aerial images." ISPRS International Journal of Geo-Information 7(**9**), p. 339.

Calonder, M., Lepetit, V., Ozuysal, M., Trzcinski, T., Strecha, C., and Fua, P. (2012). "BRIEF: computing a local binary descriptor very fast." IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 34(**7**), pp. 1281–1298.

Campana, S. (2017). "Drones in archaeology. State-of-the-art and future perspectives." Archaeological Prospection 24(**4**), pp. 275–296.

Camplani, M. and Salgado, L. (2014). "Background foreground segmentation with RGB-D Kinect data: an efficient combination of classifiers." Journal of Visual Communication and Image Representation 25(**1**), pp. 122–136.

Caroti, G., Martínez-Espejo Zaragoza, I., and Piemonte, A. (2015). "Accuracy assessment in structure from motion 3D reconstruction from UAV-born images: the influence of the data processing methods." International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS) XL(**1**), pp. 103–109.

Chakrabarti, A., Shao, J., and Shakhnarovich, G. (2016). "Depth from a single image by harmonizing overcomplete local network predictions." In: *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, pp. 2658–2666.

Chen, K., Fu, K., Yan, M., Gao, X., Sun, X., and Wei, X. (2018a). "Semantic segmentation of aerial images with shuffling convolutional neural networks." IEEE Geoscience and Remote Sensing Letters 15(**2**), pp. 173–177.

Chen, K., Lai, Y.-K., and Hu, S.-M. (2015). "3D indoor scene modeling from RGB-D data: a survey." Computational Visual Media 1(**4**), pp. 267–278.

Chen, W., Fu, Z., Yang, D., and Deng, J. (2016). "Single-image depth perception in the wild." In: *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, pp. 730–738.

Chen, Y., Wang, Y., Lu, P., Chen, Y., and Wang, G. (2018b). "Large-scale structure from motion with semantic constraints of aerial images." In: *Proceedings of the Chinese Conference on Pattern Recognition and Computer Vision (CCPR)*. Springer, pp. 347–359.

Chiabrando, F., Lingua, A., and Piras, M. (2013). "Direct photogrammetry using UAV: tests and first results." International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS) XL(**1**), pp. 81–86.

Chiang, K.-W., Tsai, M.-L., and Chu, C.-H. (2012). "The development of an UAV borne direct georeferenced photogrammetric platform for ground control point free applications." Sensors 12(**7**), pp. 9161–9180.

Chiang, K.-W., Tsai, G.-J., Li, Y.-H., and El-Sheimy, N. (2017). "Development of LiDAR-based UAV system for environment reconstruction." IEEE Geoscience and Remote Sensing Letters 14(**10**), pp. 1790–1794.

Choi, S., Min, D., Ham, B., Kim, Y., Oh, C., and Sohn, K. (2015a). "Depth analogy: data-driven approach for single image depth estimation using gradient samples." IEEE Transactions on Image Processing (TIP) 24(**12**), pp. 5953–5966.

Choi, S., Zhou, Q.-Y., and Koltun, V. (2015b). "Robust reconstruction of indoor scenes." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5556–5565.

Clothier, R. A. and Walker, R. A. (2015). "Safety risk management of unmanned aircraft systems." Handbook of unmanned aerial vehicles, pp. 2229–2275.

Cohen, A., Sattler, T., and Pollefeys, M. (2015). "Merging the unmatchable: stitching visually disconnected SfM models." In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2129–2137.

Cohen, A., Schönberger, J. L., Speciale, P., Sattler, T., Frahm, J.-M., and Pollefeys, M. (2016). "Indoor-outdoor 3d reconstruction alignment." In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 285–300.

Colomina, I. and Molina, P. (2014). "Unmanned aerial systems for photogrammetry and remote sensing: a review." ISPRS Journal of Photogrammetry and Remote Sensing 92, pp. 79–97.

Conte, G. and Doherty, P. (2009). "Vision-based unmanned aerial vehicle navigation using geo-referenced information." Journal on Advances in Signal Processing (EURASIP) 2009(**1**), 10:1–10:18.

Cramer, M. (2001). "On the use of direct georeferencing in airborne photogrammetry." In: *Proceedings of the International Symposium on Mobile Mapping Technology*. Citeseer.

Cummings, A. R., McKee, A., Kulkarni, K., and Markandey, N. (2017). "The rise of UAVs." Photogrammetric Engineering & Remote Sensing (PE&RS) 83(**4**), pp. 317–325.

DJI. *DJI: Flight Planner*. `https://www.djiflightplanner.com/`. Accessed: 2019-05-28.

Dahlke, D., Linkiewicz, M., and Meissner, H. (2015). "True 3D building reconstruction: façade, roof and overhang modelling from oblique and vertical aerial imagery." International Journal of Image and Data Fusion 6(**4**), pp. 314–329.

De Franchis, C., Meinhardt-Llopis, E., Michel, J., Morel, J.-M., and Facciolo, G. (2014). "An automatic and modular stereo pipeline for pushbroom images." International Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS) II(**3**), pp. 49–56.

Detert, M. and Weitbrecht, V. (2015). "A low-cost airborne velocimetry system: proof of concept." Journal of Hydraulic Research 53(**4**), pp. 532–539.

Devrim Kaba, M., Gokhan Uzunbas, M., and Nam Lim, S. (2017). "A reinforcement learning approach to the view planning problem." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6933–6941.

Dittmann, S, Thiessen, E, and Hartung, E (2017). "Applicability of different non-invasive methods for tree mass estimation: a review." Forest Ecology and Management 398, pp. 208–215.

Doorn, A. J. van, Koenderink, J. J., and Wagemans, J. (2011). "Light Fields and Shape from Shading." Journal of Vision 11(**3**), pp. 21.1–21.21.

Duan, L. and Lafarge, F. (2016). "Towards large-scale city reconstruction from satellites." In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 89–104.

Dunn, E. and Frahm, J.-M. (2009). "Next Best View Planning for Active Model Improvement." In: *Proceedings of the British Machine Vision Conference (BMVC)*, pp. 1–11.

Dwarakanath, D., Eichhorn, A., Halvorsen, P., and Griwodz, C. (2012). "Evaluating performance of feature extraction methods for practical 3D imaging systems." In: *Proceedings of the ACM Conference on Image and Vision Computing (CIVC)*, pp. 250–255.

Eigen, D. and Fergus, R. (2015). "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture." In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2650–2658.

Eigen, D., Puhrsch, C., and Fergus, R. (2014). "Depth map prediction from a single image using a multi-scale deep network." In: *Proceedings of Advances in Neural Information Processing Systems (NIPS)*. Vol. 2, pp. 2366–2374.

Eisenbeiss, H. and Zhang, L. (2006). "Comparison of DSMs generated from mini UAV imagery and terrestrial laser scanner in a cultural heritage application." International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS) V(**5**), pp. 90–96.

Elberink, S. O. and Vosselman, G. (2011). "Quality analysis on 3D building models reconstructed from airborne laser scanning data." ISPRS Journal of Photogrammetry and Remote Sensing 66(**2**), pp. 157–165.

Ellenberg, A, Branco, L, Krick, A, Bartoli, I, and Kontsos, A (2014). "Use of unmanned aerial vehicle for quantitative infrastructure evaluation." Journal of Infrastructure Systems 21(**3**), p. 04014054.

Engel, J., Schöps, T., and Cremers, D. (2014). "LSD-SLAM: Large-scale direct monocular SLAM." In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 834–849.

Fan, B., Du, Y., Zhu, L., and Tang, Y. (2010). "The registration of UAV down-looking aerial images to satellite images with image entropy and edges." In: *Proceedings of the International Conference on Intelligent Robotics and Applications (ICIRA)*. Springer, pp. 609–617.

Fan, X., Zhang, L., Brown, B., and Rusinkiewicz, S. (2016). "Automated view and path planning for scalable multi-object 3D scanning." ACM Transactions on Graphics (TOG) 35(**6**), p. 239.

Favaro, P. and Soatto, S. (2005). "A geometric approach to shape from defocus." IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 27(**3**), pp. 406–417.

Felzenszwalb, P. F. and Huttenlocher, D. P. (2006). "Efficient belief propagation for early vision." International Journal of Computer Vision (IJCV) 70(**1**), pp. 41–54.

Fernández-Hernandez, J., González-Aguilera, D., Rodríguez-Gonzálvez, P., and Mancera-Taboada, J. (2015). "Image-based modelling from unmanned aerial vehicle (UAV) photogrammetry: an effective, low-cost tool for archaeological applications." Archaeometry 57(**1**), pp. 128–145.

Fischer, A., Kolbe, T. H., and Lang, F. (1997). "Integration of 2D and 3D reasoning for building reconstruction using a generic hierarchical model." In: *Proceedings of the Workshop on Semantic Modeling for the Acquisition of Topographic Information from Images and Maps (SMATI)*, pp. 159–180.

Forlani, G., Dall'Asta, E., Diotri, F., Cella, U. M. d., Roncella, R., and Santise, M. (2018). "Quality assessment of DSMs produced from UAV flights georeferenced with on-board RTK positioning." Remote Sensing 10(**2**), p. 311.

Förstner, W. and Wrobel, B. P. (2016). *Photogrammetric computer vision*. Springer.

Fouhey, D. F., Gupta, A., and Hebert, M. (2013). "Data-driven 3D primitives for single image understanding." In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 3392–3399.

Fraser, C. S. (1984). "Network design considerations for non-topographic photogrammetry." Photogrammetric Engineering & Remote Sensing (PE&RS) 50(**8**), pp. 1115–1126.

Frías, E., Díaz-Vilariño, L., Balado, J., and Lorenzo, H. (2019). "From BIM to scan planning and optimization for construction control." Remote Sensing 11(**17**), p. 1963.

Frommholz, D., Linkiewicz, M., Meissner, H., Dahlke, D., and Poznanska, A. (2015). "Extracting semantically annotated 3D building models with textures from oblique aerial imagery." International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS) XL(**3**), pp. 53–58.

Fu, H., Gong, M., Wang, C., Batmanghelich, K., and Tao, D. (2018). "Deep ordinal regression network for monocular depth estimation." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2002–2011.

Furukawa, R., Sagawa, R., and Kawasaki, H. (2017). "Depth estimation using structured light flow–analysis of projected pattern flow on an object's surface." In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 4640–4648.

Furukawa, Y. and Hernández, C. (2015). "Multi-view stereo: a tutorial." Foundations and Trends in Computer Graphics and Vision 9(**1-2**), pp. 1–148.

Furukawa, Y., Curless, B., Seitz, S. M., and Szeliski, R. (2009). "Reconstructing building interiors from images." In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 80–87.

– (2010). "Towards internet-scale multi-view stereo." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1434–1441.

Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V., and Garcia-Rodriguez, J. (2017). "A review on deep learning techniques applied to semantic segmentation." arXiv preprint arXiv:1704.06857.

Garg, R., Carneiro, G., and Reid, I. (2016). "Unsupervised CNN for single view depth estimation: geometry to the rescue." In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 740–756.

Geiger, A., Lenz, P., and Urtasun, R. (2012). "Are we ready for autonomous driving? the kitti vision benchmark suite." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3354–3361.

Gerke, M. (2018). "Developments in UAV-Photogrammetry." Journal of Digital Landscape Architecture, pp. 262–272.

Gerke, M. and Przybilla, H.-J. (2016). "Accuracy analysis of photogrammetric UAV image blocks: influence of onboard RTK-GNSS and cross flight patterns." PFG: Journal of Photogrammetry, Remote Sensing and Geoinformation Science 2016(**1**), pp. 17–30.

Getzin, S., Wiegand, K., and Schöning, I. (2012). "Assessing biodiversity in forests using very high-resolution images and unmanned aerial vehicles." Methods in Ecology and Evolution 3(**2**), pp. 397–404.

Godard, C., Mac Aodha, O., and Brostow, G. J. (2017). "Unsupervised monocular depth estimation with left-right consistency." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6602–6611.

Goesele, M., Snavely, N., Curless, B., Hoppe, H., and Seitz, S. M. (2007). "Multi-view stereo for community photo collections." In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1–8.

Gómez-Candón, D., De Castro, A., and López-Granados, F. (2014). "Assessing the accuracy of mosaics from unmanned aerial vehicle (UAV) imagery for precision agriculture purposes in wheat." Precision Agriculture 15(**1**), pp. 44–56.

Gopalakrishnan, K., Gholami, H., Vidyadharan, A., Choudhary, A., and Agrawal, A. (2018). "Crack damage detection in unmanned aerial vehicle images of civil infrastructure using pre-trained deep learning model." International Journal of Traffic and Transportation Engineering 8, pp. 1–14.

Guo, X., Li, H., Yi, S., Ren, J., and Wang, X. (2018). "Learning monocular depth by distilling cross-domain stereo networks." In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 484–500.

Haala, N. and Rothermel, M. (2015). "Image-based 3D data capture in urban scenarios." Photogrammetric Week, pp. 119–130.

Haala, N. and Kada, M. (2010). "An update on automatic 3D building reconstruction." ISPRS Journal of Photogrammetry and Remote Sensing 65(**6**), pp. 570–580.

Haala, N., Rothermel, M., and Cavegn, S. (2015). "Extracting 3D urban models from oblique aerial images." In: *Proceedings of the IEEE Joint Urban Remote Sensing Event (JURSE)*, pp. 1–4.

Han, X.-F., Jin, J. S., Wang, M.-J., Jiang, W., Gao, L., and Xiao, L. (2017). "A review of algorithms for filtering the 3D point cloud." Signal Processing: Image Communication 57, pp. 103–112.

Hane, C., Ladicky, L., and Pollefeys, M. (2015). "Direction matters: Depth estimation with a surface normal classifier." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 381–389.

Hao, Z., Li, Y., You, S., and Lu, F. (2018). "Detail preserving depth estimation from a single image using attention guided networks." In: *Proceedings of the International Conference on 3D Vision (3DV)*. IEEE, pp. 304–313.

Hartley, R. and Zisserman, A. (2003). *Multiple view geometry in computer vision*. Cambridge University Press.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778.

Heber, S. and Pock, T. (2016). "Convolutional networks for shape from light field." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3746–3754.

Hedau, V., Hoiem, D., and Forsyth, D. (2009). "Recovering the spatial layout of cluttered rooms." In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1849–1856.

Heinly, J., Dunn, E., and Frahm, J.-M. (2012). "Comparative evaluation of binary features." In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 759–773.

Heng, L., Lee, G. H., Fraundorfer, F., and Pollefeys, M. (2011). "Real-time photo-realistic 3D mapping for micro aerial vehicles." In: *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4012–4019.

Heo, M., Lee, J., Kim, K.-R., Kim, H.-U., and Kim, C.-S. (2018). "Monocular depth estimation using whole strip masking and reliability-based refinement." In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 36–51.

Hepp, B., Dey, D., Sinha, S. N., Kapoor, A., Joshi, N., and Hilliges, O. (2018a). "Learn-to-Score: efficient 3D scene exploration by predicting view utility." In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 437–452.

Hepp, B., Nießner, M., and Hilliges, O. (2018b). "Plan3D: viewpoint and trajectory optimization for aerial multi-view stereo reconstruction." ACM Transactions on Graphics (TOG) 38(**1**), p. 4.

Hirschmuller, H. (2005). "Accurate and efficient stereo processing by semi-global matching and mutual information." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 2, pp. 807–814.

Hodgson, J. C., Baylis, S. M., Mott, R., Herrod, A., and Clarke, R. H. (2016). *Precision wildlife monitoring using unmanned aerial vehicles*. Tech. rep. Scientific Report, p. 22574.

Hofer, M., Maurer, M., and Bischof, H. (2015). "Line3D: efficient 3D scene abstraction for the built environment." In: *Proceedings of the German Conference on Pattern Recognition (GCPR)*. Springer, pp. 237–246.

Hoiem, D., Efros, A. A., and Hebert, M. (2007). "Recovering surface layout from an image." International Journal of Computer Vision (IJCV) 75(**1**), pp. 151–172.

Holzmann, T., Fraundorfer, F., and Bischof, H. (2016a). "Direct stereo visual odometry based on lines." In: *VISIGRAPP*, pp. 476–487.

– (2016b). "Regularized 3D modeling from noisy building reconstructions." In: *Proceedings of the International Conference on 3D Vision (3DV)*. IEEE, pp. 528–536.

Hoppe, C., Wendel, A., Zollmann, S., Pirker, K., Irschara, A., Bischof, H., and Kluckner, S. (2012). "Photogrammetric camera network design for micro aerial vehicles." In: *Proceedings of the Computer Vision Winter Workshop (CVWW)*. Vol. 8, pp. 1–3.

Horn, B. K. P. (1970). *Shape from shading: a method for obtaining the shape of a smooth opaque object from one view*. Tech. rep. Cambridge, MA, USA: MIT - AI.

Hu, J., Ozay, M., Zhang, Y., and Okatani, T. (2019). "Revisiting single image depth estimation: toward higher resolution maps with accurate object boundaries." In: *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1043–1051.

Huang, H., Brenner, C., and Sester, M. (2013). "A generative statistical approach to automatic 3D building roof reconstruction from laser scanning data." ISPRS Journal of Photogrammetry and Remote Sensing 79, pp. 29–43.

Huang, R., Zou, D., Vaughan, R., and Tan, P. (2018). "Active image-based modeling with a toy drone." In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1–8.

IV, G. P. J., Pearlstine, L. G., and Percival, H. F. (2006). "An assessment of small unmanned aerial vehicles for wildlife research." Wildlife Society Bulletin 34(**3**), pp. 750–758.

Ikehata, S., Yang, H., and Furukawa, Y. (2015). "Structured indoor modeling." In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1323–1331.

Jarzabek-Rychard, M. and Karpina, M. (2016). "Quality analysis on 3D building models reconstructed from UAV imagery." International Archives of the Photogrammetry, Remote Sensing ans Spatial Information Sciences (ISPRS) XLI(**1**), pp. 1121–1126.

Jiao, J., Cao, Y., Song, Y., and Lau, R. (2018). "Look deeper into depth: monocular depth estimation with semantic booster and attention-driven loss." In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 53–69.

Jing, W., Polden, J., Tao, P. Y., Lin, W., and Shimada, K. (2016). "View planning for 3D shape reconstruction of buildings with unmanned aerial vehicles." In: *Proceedings of the IEEE International Conference on Control, Automation, Robotics and Vision (ICARCV)*, pp. 1–6.

Juan, L. and Gwun, O. (2009). "A comparison of SIFT, PCA-SIFT and SURF." International Journal of Image Processing (IJIP) 3(**4**), pp. 143–152.

Julian, K., Mern, J., and Tompa, R. (2017). *UAV depth perception from visual images using a deep convolutional neural network*. Tech. rep. Stanford University.

Kada, M. and McKinley, L. (2009). "3D building reconstruction from LiDAR based on a cell decomposition approach." International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS) XXXVIII(**3**), pp. 47–52.

Kadambi, A., Taamazyan, V., Shi, B., and Raskar, R. (2015). "Polarized 3D: high-quality depth sensing with polarization cues." In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 3370–3378.

Kaiser, P., Wegner, J. D., Lucchi, A., Jaggi, M., Hofmann, T., and Schindler, K. (2017). "Learning aerial image segmentation from online maps." IEEE Transactions on Geoscience and Remote Sensing 55(**11**), pp. 6054–6068.

Kaplan, A., Avraham, T., and Lindenbaum, M. (2016). "Interpreting the ratio criterion for matching SIFT descriptors." In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 697–712.

Karel, W., Doneus, M., Briese, C., Verhoeven, G., and Pfeifer, N. (2014). "Investigation on the automatic geo-referencing of archaeological UAV photographs by correlation with pre-existing ortho-photos." International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS) XL(**5**), pp. 307–312.

Karsch, K., Liu, C., and Kang, S. B. (2014). "Depth transfer: Depth extraction from video using non-parametric sampling." IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 36(**11**), pp. 2144–2158.

Kerl, C., Sturm, J., and Cremers, D. (2013). "Robust odometry estimation for RGB-D cameras." In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3748–3754.

Khanal, S., Fulton, J., and Shearer, S. (2017). "An overview of current and potential applications of thermal remote sensing in precision agriculture." Computers and Electronics in Agriculture 139, pp. 22–32.

Kim, K. and Shan, J. (2011). "Building roof modeling from airborne laser scanning data based on level set approach." ISPRS Journal of Photogrammetry and Remote Sensing 66(**4**), pp. 484–497.

Kim, S., Park, K., Sohn, K., and Lin, S. (2016). "Unified depth prediction and intrinsic image decomposition from a single image via joint convolutional neural fields." In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 143–159.

Klosterman, S. and Richardson, A. (2017). "Observing spring and fall phenology in a deciduous forest with aerial drone imagery." Sensors 17(**12**), p. 2852.

Koch, T., Körner, M., and Fraundorfer, F. (2016a). "Automatic alignment of indoor and outdoor building models using 3D line segments." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR-WS)*, pp. 10–18.

Koch, T., d'Angelo, P., Kurz, F., Fraundorfer, F., Reinartz, P., and Körner, M. (2016b). "The TUM-DLR multimodal earth observation evaluation benchmark." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR-WS)*, pp. 698–705.

Koch, T., Körner, M., and Fraundorfer, F. (2019a). "Automatic and semantically-aware 3D UAV flight planning for image-based 3D reconstruction." Remote Sensing 11(**13**), p. 1550.

Koch, T., Liebel, L., Körner, M., and Fraundorfer, F. (2019b). "Comparison of monocular depth estimation methods using geometrically relevant metrics on the IBims-1 dataset." Computer Vision and Image Understanding (CVIU), under review.

Kolbe, T., Nagel, C., and Stadler, A. (2009). "CityGML–OGC standard for photogrammetry." Photogrammetric Week, pp. 265–277.

Kolmogorov, V. and Zabih, R. (2001). "Computing visual correspondence with occlusions using graph cuts." In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 508–515.

Kong, N. and Black, M. J. (2015). "Intrinsic depth: improving depth transfer with intrinsic images." In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 3514–3522.

Konrad, J., Brown, G., Wang, M., Ishwar, P., Wu, C., and Mukherjee, D. (2012). "Automatic 2D-to-3D image conversion using 3D examples from the internet." In: *Proceedings of the Stereoscopic Displays and Applications*. International Society for Optics and Photonics, 82880F.

Konrad, J., Wang, M., Ishwar, P., Wu, C., and Mukherjee, D. (2013). "Learning-based, automatic 2D-to-3D image and video conversion." IEEE Transactions on Image Processing (TIP) 22(**9**), pp. 3485–3496.

Krause, A. and Golovin, D. (2014). *Submodular Function Maximization.*

Kriegel, S., Rink, C., Bodenmüller, T., and Suppa, M. (2015). "Efficient next-best-scan planning for autonomous 3D surface reconstruction of unknown objects." Journal of Real-Time Image Processing 10(**4**), pp. 611–631.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "Imagenet classification with deep convolutional neural networks." In: *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, pp. 1097–1105.

Kumar Ramakrishnan, S. and Grauman, K. (2018). "Sidekick policy learning for active visual exploration." In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 413–430.

Kurz, F., Rosenbaum, D., Leitloff, J., Meynberg, O., and Reinartz, P. (2011). "Real time camera system for disaster and traffic monitoring." In: *Proceedings of the International Conference on Sensors and Models in Photogrammetry and Remote Sensing*.

Kurz, F., Rosenbaum, D., Meynberg, O., Mattyus, G., and Reinartz, P. (2014). "Performance of a real-time sensor and processing system on a helicopter." International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS) XL-1, pp. 189–193.

Kuznietsov, Y., Stückler, J., and Leibe, B. (2017). "Semi-supervised deep learning for monocular depth map prediction." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6647–6655.

Ladicky, L., Shi, J., and Pollefeys, M. (2014). "Pulling things out of perspective." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 89–96.

Lafarge, F. and Mallet, C. (2012). "Creating large-scale city models from 3D-point clouds: a robust approach with hybrid representation." International Journal of Computer Vision (IJCV) 99(**1**), pp. 69–85.

Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., and Navab, N. (2016). "Deeper depth prediction with fully convolutional residual networks." In: *Proceedings of the International Conference on 3D Vision (3DV)*. IEEE, pp. 239–248.

Lee, J.-H., Heo, M., Kim, K.-R., and Kim, C.-S. (2018). "Single-image depth estimation based on Fourier domain analysis." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 330–339.

Lee, S. C. and Nevatia, R. (2004). "Extraction and integration of window in a 3D building model from ground view images." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lehmann, J., Nieberding, F., Prinz, T., and Knoth, C. (2015). "Analysis of unmanned aerial system-based CIR images in forestry - a new perspective to monitor pest infestation levels." Forests 6(**3**), pp. 594–612.

Lehtola, V., Kaartinen, H., Nüchter, A., Kaijaluoto, R., Kukko, A., Litkey, P., Honkavaara, E., Rosnell, T., Vaaja, M., Virtanen, J.-P., et al. (2017). "Comparison of the selected state-of-the-art 3D indoor scanning and point cloud generation methods." Remote Sensing 9(**8**), p. 796.

Li, B., Shen, C., Dai, Y., Hengel, A. van den, and He, M. (2015). "Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1119–1127.

Li, J., Klein, R., and Yao, A. (2017). "A two-streamed network for estimating fine-scaled depth maps from single RGB images." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3372–3380.

Li, X., Qin, H., Wang, Y., Zhang, Y., and Dai, Q. (2014). "DEPT: depth estimation by parameter transfer for single still images." In: *Proceedings of the Asian Conference on Computer Vision (ACCV)*. Springer, pp. 45–58.

Lin, T.-Y., Cui, Y., Belongie, S., and Hays, J. (2015). "Learning deep representations for ground-to-aerial geolocalization." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5007–5015.

Lin, Y., Yu, Q., and Medioni, G. (2007). "Map-enhanced UAV image sequence registration." In: *Proceedings of the IEEE Workshop on Applications of Computer Vision (WACV)*, pp. 15–15.

Liu, B., Gould, S., and Koller, D. (2010). "Single image depth estimation from predicted semantic labels." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1253–1260.

Liu, C., Yuen, J., and Torralba, A. (2011). "Sift flow: dense correspondence across scenes and its applications." IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 33(**5**), pp. 978–994.

Liu, C., Yang, J., Ceylan, D., Yumer, E., and Furukawa, Y. (2018). "PlaneNet: piece-wise planar reconstruction from a single RGB image." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2579–2588.

Liu, F., Shen, C., and Lin, G. (2015). "Deep convolutional neural fields for depth estimation from a single image." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5162–5170.

Liu, F., Shen, C., Lin, G., and Reid, I. (2016). "Learning depth from single monocular images using deep convolutional neural fields." IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 38(**10**), pp. 2024–2039.

Liu, H., Li, C., Chen, G., Zhang, G., Kaess, M., and Bao, H. (2017). "Robust keyframe-based dense SLAM with an RGB-D camera." arXiv preprint arXiv:1711.05166.

Liu, M., Salzmann, M., and He, X. (2014). "Discrete-continuous depth estimation from a single image." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 716–723.

Lowe, D. G. (2004). "Distinctive image features from scale-invariant keypoints." International Journal of Computer Vision (IJCV) 60(**2**), pp. 91–110.

Luhmann, T., Robson, S., Kyle, S., and Boehm, J. (2013). *Close-range photogrammetry and 3D imaging*. Walter de Gruyter.

Macher, H., Landes, T., and Grussenmeyer, P. (2017). "From point clouds to building information models: 3D semi-automatic reconstruction of indoors of existing buildings." Applied Sciences 7(**10**), p. 1030.

Maes, W. H. and Steppe, K. (2018). "Perspectives for remote sensing with unmanned aerial vehicles in precision agriculture." Trends in Plant Science 24(**2**), pp. 152–164.

Majdik, A. L., Verda, D., Albers-Schoenberg, Y., and Scaramuzza, D. (2015). "Air-ground matching: appearance-based GPS-denied urban localization of micro aerial vehicles." Journal of Field Robotics 32(**7**), pp. 1015–1039.

Malihi, S., Valadan Zoej, M., and Hahn, M. (2018). "Large-scale accurate reconstruction of buildings employing point clouds generated from UAV imagery." Remote Sensing 10(**7**), p. 1148.

Manfreda, S., McCabe, M., Miller, P., Lucas, R., Pajuelo Madrigal, V., Mallinis, G., Ben Dor, E., Helman, D., Estes, L., Ciraolo, G., et al. (2018). "On the use of unmanned aerial systems for environmental monitoring." Remote Sensing 10(**4**), p. 641.

Marcu, A., Costea, D., Licaret, V., Pîrvu, M., Slusanschi, E., and Leordeanu, M. (2018). "SafeUAV: learning to estimate depth and safe landing areas for UAVs from synthetic data." In: *Proceedings of the European Conference on Computer Vision Workshops (ECCV-WS)*. Springer, pp. 43–58.

Marmanis, D., Wegner, J. D., Galliani, S., Schindler, K., Datcu, M., and Stilla, U. (2016). "Semantic segmentation of aerial images with an ensemble of CNNs." International Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS) III(**3**), pp. 473–480.

Matei, B. C., Sawhney, H. S., Samarasekera, S., Kim, J., and Kumar, R. (2008). "Building segmentation for densely built urban regions using aerial LiDAR data." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8.

Mendez, O., Hadfield, S., Pugeault, N., and Bowden, R. (2017). "Taking the scenic route to 3D: optimising reconstruction from moving cameras." In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Vol. 3, pp. 4687–4695.

Meng, Z., Qin, H., Chen, Z., Chen, X., Sun, H., Lin, F., and Ang Jr, M. H. (2017). "A two-stage optimized next-view planning framework for 3-D unknown environment exploration, and structural reconstruction." IEEE Robotics and Automation Letters 2(**3**), pp. 1680–1687.

Merino, L., Caballero, F., Martínez-De-Dios, J. R., Maza, I., and Ollero, A. (2012). "An unmanned aircraft system for automatic forest fire monitoring and measurement." Journal of Intelligent & Robotic Systems 65(**1-4**), pp. 533–548.

Michael, N., Shen, S., Mohta, K., Kumar, V., Nagatani, K., Okada, Y., Kiribayashi, S., Otake, K., Yoshida, K., Ohno, K., et al. (2012). "Collaborative mapping of an earthquake damaged building via ground and aerial robots." Journal of Field Robotics 29(**5**), pp. 832–841.

Micusik, B. and Kosecka, J. (2009). "Piecewise planar city 3D modeling from street view panoramic sequences." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2906–2912.

Moe, K., Toschi, I., Poli, D., Lago, F., Schreiner, C., Legat, K., and Remondino, F. (2016). "Changing the production pipeline - use of oblique aerial cameras for mapping purposes." International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS) XLI(**4**), pp. 631–637.

Mostegel, C., Rumpler, M., Fraundorfer, F., and Bischof, H. (2016). "UAV-based autonomous image acquisition with multi-view stereo quality assurance by confidence prediction." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR-WS)*, pp. 1–10.

Mostegel, C., Prettenthaler, R., Fraundorfer, F., and Bischof, H. (2017). "Scalable surface reconstruction from point clouds with extreme scale and density diversity." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 904–913.

Mueller, C. and Neumann, K. (2016). "Leica DMC III calibration and geometric sensor accuracy." International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS) XL(**3**), pp. 1–9.

Murtiyoso, A., Koehl, M., Grussenmeyer, P., and Freville, T. (2017). "Acquisition and processing protocols for UAV images: 3D modeling of historical buildings using photogrammetry." International Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS) IV(**2**), pp. 163–170.

Murtiyoso, A. and Grussenmeyer, P. (2017). "Documentation of heritage buildings using close-range UAV images: dense matching issues, comparison and case studies." The Photogrammetric Record 32(**159**), pp. 206–229.

Nagel, C., Becker, T., Kaden, R., Li, K.-J., Lee, J., and Kolbe, T. H. (2010). "Requirements and space-event modeling for indoor navigation - how to simultaneously address route planning, multiple localization methods, navigation contexts, and different locomotion types."

Nassar, A., Amer, K., ElHakim, R., and ElHelw, M. (2018). "A deep CNN-based framework for enhanced aerial imagery registration with applications to UAV geolocalization." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR-WS)*, pp. 1513–1523.

Natesan, S., Armenakis, C., and Vepakomma, U. (2019). "Resnet-based tree species classification using UAV images." International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS) XLII(**2**), pp. 475–481.

Nayar, S. K. and Narasimhan, S. G. (1999). "Vision in bad weather." In: *Proceedings of the IEEE International Conference on Computer Vision (CVPR)*. Vol. 2, pp. 820–827.

Neumann, K. J. (2011). "The Z/I DMC II–"Imaging Revolution"." Photogrammetric Week, pp. 97–101.

Nex, F, Remondino, F, Gerke, M, Przybilla, H.-J., Bäumker, M, and Zurhorst, A (2015). "ISPRS benchmark for multi-platform photogrammetry." International Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS) II(**3**), pp. 135–142.

Nex, F. and Remondino, F. (2014). "UAV for 3D mapping applications: a review." Applied Geomatics 6(**1**), pp. 1–15.

Ngo, T. T., Nagahara, H., and Taniguchi, R.-i. (2015). "Shape and light directions from shading and polarization." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2310–2318.

Nuske, S., Choudhury, S., Jain, S., Chambers, A., Yoder, L., Scherer, S., Chamberlain, L., Cover, H., and Singh, S. (2015). "Autonomous exploration and motion planning for an unmanned aerial vehicle navigating rivers." Journal of Field Robotics 32(**8**), pp. 1141–1162.

Ochmann, S., Vock, R., Wessel, R., and Klein, R. (2016). "Automatic reconstruction of parametric building models from indoor point clouds." Computers & Graphics 54, pp. 94–103.

Ono, Y., Trulls, E., Fua, P., and Yi, K. M. (2018). "LF-Net: learning local features from images." In: *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, pp. 6234–6244.

Onyango, F., Nex, F, Peter, M., and Jende, P (2017). "Accurate estimation of orientation parameters of UAV images through image registration with aerial oblique imagery."

International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS) XLII(**1**), pp. 599–605.

Ostrowski, W (2016). "Accuracy of measurements in oblique aerial images for urban environment." International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS) XLII(**2**), pp. 79–85.

Özdemir, E. and Remondino, F. (2019). "Classification of aerial point clouds with deep learning." International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS) XLII(**2**).

Padró, J.-C., Muñoz, F.-J., Planas, J., and Pons, X. (2019). "Comparison of four UAV georeferencing methods for environmental monitoring purposes focusing on the combined use with airborne and satellite remote sensing platforms." International Journal of Applied Earth Observation and Geoinformation 75, pp. 130–140.

Pajares, G. (2015). "Overview and current status of remote sensing applications based on unmanned aerial vehicles (UAVs)." Photogrammetric Engineering & Remote Sensing (PE&RS) 81(**4**), pp. 281–330.

Palazzolo, E. and Stachniss, C. (2018). "Effective exploration for MAVs based on the expected information gain." Drones 2(**1**), p. 9.

Park, S., Nolan, A., Ryu, D., Fuentes, S., Hernandez, E., Chung, H., and O'connell, M. (2015). "Estimation of crop water stress in a nectarine orchard using high-resolution imagery from unmanned aerial vehicle (UAV)." In: *Proceedings of the International Congress on Modelling and Simulation*, pp. 1413–1419.

Partovi, T., Huang, H., Krauß, T., Mayer, H., and Reinartz, P. (2015). "Statistical building roof reconstruction from Worldview-2 stereo imagery." International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS) XL(**3**), pp. 161–167.

Peng, C. and Isler, V. (2019). "Adaptive view planning for aerial 3D reconstruction." In: *Proceedings of the International Conference on Robotics and Automation (ICRA)*, pp. 2981–2987.

Pix4Da. *Pix4D: Professional Photogrammetry and Drone Mapping Software*. http://www.pix4d.com/. Accessed: 2019-04-22.

Pix4Db. *Pix4D: Capture*. https://pix4d.com/product/pix4dcapture/. Accessed: 2019-05-28.

Poli, D., Remondino, F., Angiuli, E., and Agugiaro, G. (2015). "Radiometric and geometric evaluation of GeoEye-1, WorldView-2 and Pléiades-1A stereo images for 3D information extraction." ISPRS Journal of Photogrammetry and Remote Sensing 100, pp. 35–47.

Pomerleau, F., Colas, F., Siegwart, R., and Magnenat, S. (2013). "Comparing ICP variants on real-world data sets." Autonomous Robots 34(**3**), pp. 133–148.

Precisionhawk. *Precisionhawk: Precision Flight*. https://www.precisionhawk.com/precisionflight/. Accessed: 2019-05-28.

Previtali, M., Scaioni, M., Barazzetti, L., and Brumana, R. (2014). "A flexible methodology for outdoor/indoor building reconstruction from occluded point clouds." International Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS) II(**3**), pp. 119–126.

Pu, S. and Vosselman, G. (2009). "Knowledge based reconstruction of building models from terrestrial laser scanning data." ISPRS Journal of Photogrammetry and Remote Sensing 64(**6**), pp. 575–584.

Puerta, J. and Fraundorfer, F. (2016). "Vision based safe navigation for search and rescue drones." In: *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1–8.

Qi, J., Song, D., Shang, H., Wang, N., Hua, C., Wu, C., Qi, X., and Han, J. (2016). "Search and rescue rotary-wing UAV and its application to the lushan ms 7.0 earthquake." Journal of Field Robotics 33(**3**), pp. 290–321.

Ramamonjisoa, M. and Lepetit, V. (2019). "SharpNet: fast and accurate recovery of occluding contours in monocular depth estimation." In: *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCV-WS)*, tbd.

Ranftl, R., Vineet, V., Chen, Q., and Koltun, V. (2016). "Dense monocular depth estimation in complex dynamic scenes." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4058–4066.

Rathinam, S., Kim, Z. W., and Sengupta, R. (2008). "Vision-based monitoring of locally linear structures using an unmanned aerial vehicle." Journal of Infrastructure Systems 14(**1**), pp. 52–63.

Remondino, F. and Gerke, M. (2015). "Oblique aerial imagery : a review." Photogrammetric Week, pp. 75–83.

Roberts, M., Dey, D., Truong, A., Sinha, S., Shah, S., Kapoor, A., Hanrahan, P., and Joshi, N. (2017). "Submodular trajectory optimization for aerial 3D scanning." In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 5324–5333.

Rosten, E. and Drummond, T. (2006). "Machine learning for high-speed corner detection." In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 430–443.

Roy, A. and Todorovic, S. (2016). "Monocular depth estimation using neural regression forest." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5506–5514.

Rublee, E., Rabaud, V., Konolige, K., and Bradski, G. (2011). "ORB: An efficient alternative to SIFT or SURF." In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2564–2571.

Rumpler, M., Irschara, A., and Bischof, H. (2011). "Multi-view stereo: redundancy benefits for 3D reconstruction." In: *Proceedings of the Workshop of the Austrian Association for Pattern Recognition (AAPR)*.

Rumpler, M., Daftry, S., Tscharf, A., Prettenthaler, R., Hoppe, C., Mayer, G., and Bischof, H. (2014). "Automated end-to-end workflow for precise and geo-accurate reconstructions using fiducial markers." International Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS) II(**3**), pp. 135–142.

Rupnik, E., Daakir, M., and Deseilligny, M. P. (2017). "MicMac–a free, open-source solution for photogrammetry." Open Geospatial Data, Software and Standards 2(**1**), p. 14.

Saari, H., Pellikka, I., Pesonen, L., Tuominen, S., Heikkilä, J., Holmlund, C., Mäkynen, J., Ojala, K., and Antila, T. (2011). "Unmanned aerial vehicle (UAV) operated spectral camera system for forest and agriculture applications." In: *Remote Sensing for Agriculture, Ecosystems, and Hydrology*. International Society for Optics and Photonics.

Saeed, A. S., Younes, A. B., Cai, C., and Cai, G. (2018). "A survey of hybrid unmanned aerial vehicles." Progress in Aerospace Sciences 98, pp. 91–105.

Sampath, A. and Shan, J. (2010). "Segmentation and reconstruction of polyhedral building roofs from aerial LiDAR point clouds." IEEE Transactions on Geoscience and Remote Sensing (TGRS) 48(**3**), pp. 1554–1567.

Sauerbier, M. and Eisenbeiss, H. (2010). "UAVs for the documentation of archaeological excavations." International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS) XXXVIII(**5**), pp. 526–531.

Saxena, A., Chung, S. H., and Ng, A. Y. (2006). "Learning depth from single monocular images." In: *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, pp. 1161–1168.

– (2008). "3-d depth reconstruction from a single still image." International Journal of Computer Vision (IJCV) 76(**1**), pp. 53–69.

Saxena, A., Sun, M., and Ng, A. Y. (2009). "Make3d: learning 3D scene structure from a single still image." IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 31(**5**), pp. 824–840.

Scharstein, D. and Szeliski, R. (2002). "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms." International Journal of Computer Vision (IJCV) 47(**1-3**), pp. 7–42.

Schindler, K. and Bauer, J. (2003). "A model-based method for building reconstruction." In: *Proceedings of the IEEE International Workshop on Higher-Level Knowledge in 3D Modeling and Motion Analysis*, pp. 74–82.

Schmid, K., Hirschmüller, H., Dömel, A., Grixa, I., Suppa, M., and Hirzinger, G. (2012). "View planning for multi-view stereo 3D reconstruction using an autonomous multicopter." Journal of Intelligent & Robotic Systems 65(**1-4**), pp. 309–323.

Schönberger, J. L., Hardmeier, H., Sattler, T., and Pollefeys, M. (2017). "Comparative evaluation of hand-crafted and learned local features." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1482–1491.

Schönberger, J. L., Pollefeys, M., Geiger, A., and Sattler, T. (2018). "Semantic visual localization." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6896–6906.

Schönberger, J. L. and Frahm, J.-M. (2016). "Structure-from-motion revisited." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4104–4113.

Seitz, S. M., Curless, B., Diebel, J., Scharstein, D., and Szeliski, R. (2006). "A comparison and evaluation of multi-view stereo reconstruction algorithms." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 519–528.

Shan, Q., Wu, C., Curless, B., Furukawa, Y., Hernandez, C., and Seitz, S. M. (2014). "Accurate geo-registration by ground-to-aerial image matching." In: *Proceedings of the International Conference on 3D Vision (3DV)*. IEEE, pp. 525–532.

Shen, S., Michael, N., and Kumar, V. (2012). "Autonomous indoor 3D exploration with a micro-aerial vehicle." In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9–15.

Shi, J., Tao, X., Xu, L., and Jia, J. (2015). "Break ames room illusion: depth from general single images." ACM Transactions on Graphics (TOG) 34(**6**), p. 225.

Silberman, N., Hoiem, D., Kohli, P., and Fergus, R. (2012). "Indoor segmentation and support inference from rgbd images." In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 746–760.

Simonyan, K. and Zisserman, A. (2014). "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556.

Smith, N., Moehrle, N., Goesele, M., and Heidrich, W. (2018). "Aerial path planning for urban scene reconstruction: a continuous optimization method and benchmark." In: *Proceedings of the ACM SIGGRAPH Conference and Exhibition on Computer Graphics and Interactive Techniques*, p. 183.

Snavely, N., Seitz, S. M., and Szeliski, R. (2006). "Photo Tourism: exploring photo collections in 3D." ACM Transactions on Graphics (TOG) 25(**3**), pp. 835–846.

– (2008). "Skeletal graphs for efficient structure from motion." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8.

Song, J., Wu, J., and Jiang, Y. (2015). "Extraction and reconstruction of curved surface buildings by contour clustering using airborne LiDAR data." Optik 126(**5**), pp. 513–521.

Stöcker, C., Bennett, R., Nex, F., Gerke, M., and Zevenbergen, J. (2017). "Review of the current state of UAV regulations." Remote Sensing 9(**5**), p. 459.

Strecha, C., Krull, M., and Betschart, S. (2014). *The Chillon Project: Aerial/ Terrestrial and Indoor Integration*. https://s3.amazonaws.com/mics.pix4d.com/KB/documents/Pix4D-White-Paper-Chillon-Project-.pdf. Accessed: 2019-04-22.

Stumberg, L. von, Usenko, V., Engel, J., Stückler, J., and Cremers, D. (2016). "Autonomous exploration with a low-cost quadrocopter using semi-dense monocular SLAM." arXiv preprint arXiv:1609.07835.

Sturm, J., Bylow, E., Kerl, C., Kahl, F., and Cremer, D. (2013). "Dense tracking and mapping with a quadrocopter." International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS) XL(**1**), pp. 395–400.

Sun, J., Li, B., Jiang, Y., and Wen, C.-y. (2016). "A camera-based target detection and positioning UAV system for search and rescue (SAR) purposes." Sensors 16(**11**), p. 1778.

Suveg, I. and Vosselman, G. (2000). "3D reconstruction of building models." International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS) IV(**2**), pp. 538–545.

– (2002). "Automatic 3D building reconstruction." In: *Three-Dimensional Image Capture and Applications V*. Vol. 4661. International Society for Optics and Photonics, pp. 59–70.

Suwajanakorn, S., Hernandez, C., and Seitz, S. M. (2015). "Depth from focus with your mobile phone." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3497–3506.

Szeliski, R. (2010). *Computer vision: algorithms and applications*. Springer Science & Business Media.

Tareen, S. A. K. and Saleem, Z. (2018). "A comparative analysis of SIFT, SURF, KARZE, AKAZE, ORB, and BRISK." In: *Proceedings of the IEEE International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, pp. 1–10.

Themistocleous, K., Agapiou, A., and Hadjimitsis, D. (2016). "3D documentation and BIM modeling of cultural heritage structures using UAVs: the case of the foinikaria church." International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS) XLII(**2**), pp. 45–49.

Themistocleous, K., Ioannides, M., Agapiou, A., and Hadjimitsis, D. G. (2015). "The methodology of documenting cultural heritage sites using photogrammetry, UAV, and 3D printing techniques: the case study of Asinou Church in Cyprus." In: *Proceedings of the International Conference on Remote Sensing and Geoinformation of the Environment (RSCy)*. Vol. 9535, p. 953510.

Tian, Y., Gerke, M., Vosselman, G., and Zhu, Q. (2010). "Knowledge-based building reconstruction from terrestrial video sequences." ISPRS Journal of Photogrammetry and Remote Sensing 65(**4**), pp. 395–408.

Toft, C., Stenborg, E., Hammarstrand, L., Brynte, L., Pollefeys, M., Sattler, T., and Kahl, F. (2018). "Semantic match consistency for long-term visual localization." In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 383–399.

Torralba, A. and Oliva, A. (2002). "Depth estimation from image structure." IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 24(**9**), pp. 1226–1238.

Toschi, I., Ramos, M., Nocerino, E., Menna, F., Remondino, F., Moe, K., Poli, D., Legat, K., and Fassi, F. (2017). "Oblique photogrammetry supporting 3D urban reconstruction of complex scenarios." International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS) XLII(**1**), pp. 519–526.

Tsai, C.-H. and Lin, Y.-C. (2017). "An accelerated image matching technique for UAV orthoimage registration." ISPRS Journal of Photogrammetry and Remote Sensing 128, pp. 130–145.

Turner, D., Lucieer, A., and Watson, C. (2012). "An automated technique for generating georectified mosaics from ultra-high resolution unmanned aerial vehicle (UAV) imagery, based on structure from motion (SfM) point clouds." Remote Sensing 4(**5**), pp. 1392–1410.

Turner, D., Lucieer, A., and Wallace, L. (2013). "Direct georeferencing of ultrahigh-resolution UAV imagery." IEEE Transactions on Geoscience and Remote Sensing (TGRS) 52(**5**), pp. 2738–2745.

Turner, E. and Zakhor, A. (2015). "Automatic indoor 3D surface reconstruction with segmented building and object elements." In: *Proceedings of the International Conference on 3D Vision (3DV)*. IEEE, pp. 362–370.

Tuttas, S., Braun, A., Borrmann, A., and Stilla, U. (2017). "Acquisition and consecutive registration of photogrammetric point clouds for construction progress monitoring using a 4D BIM." PFG: Journal of Photogrammetry, Remote Sensing and Geoinformation Science 85(**1**), pp. 3–15.

Ummenhofer, B., Zhou, H., Uhrig, J., Mayer, N., Ilg, E., Dosovitskiy, A., and Brox, T. (2017). "DeMoN: depth and motion network for learning monocular stereo." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 5, pp. 5038–5047.

Vacca, G., Dessì, A., and Sacco, A. (2017). "The use of nadir and oblique UAV images for building knowledge." ISPRS International Journal of Geo-Information 6(**12**), p. 393.

Van der Wal, T., Abma, B., Viguria, A., Previnaire, E., Zarco-Tejada, P., Serruys, P., van Valkengoed, E., and van der Voet, P. (2013). "Fieldcopter: unmanned aerial systems for crop monitoring services." In: *Proceedings of Precision Agriculture*, pp. 169–175.

Verhoeven, G., Wieser, M., Briese, C., and Doneus, M. (2013). "Positioning in time and space: cost-effective exterior orientation for airborne archaeological photographs." International

Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS) II(**5**), pp. 313–318.

Verma, V., Kumar, R., and Hsu, S. (2006). "3D building detection and modeling from aerial LiDAR data." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2213–2220.

Vosselman, G. (1999). "Building reconstruction using planar faces in very high density height data." International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS) II(**3**), pp. 87–94.

Vosselman, G., Dijkman, S., et al. (2001). "3D building model reconstruction from point clouds and ground plans." International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS) XXXIV(**3**), pp. 37–44.

Wang, K. and Frahm, J.-M. (2017a). "Fast and accurate satellite multi-view stereo using edge-aware interpolation." In: *Proceedings of the International Conference on 3D Vision (3DV)*. IEEE, pp. 365–373.

– (2017b). "Single view parametric building reconstruction from satellite imagery." In: *Proceedings of the International Conference on 3D Vision (3DV)*. IEEE, pp. 603–611.

Wang, K., Stutts, C., Dunn, E., and Frahm, J.-M. (2016a). "Efficient joint stereo estimation and land usage classification for multiview satellite data." In: *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WCACV)*, pp. 1–9.

Wang, P., Shen, X., Lin, Z., Cohen, S., Price, B., and Yuille, A. L. (2015). "Towards unified depth and semantic prediction from a single image." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2800–2809.

Wang, P., Shen, X., Russell, B., Cohen, S., Price, B., and Yuille, A. L. (2016b). "Surge: surface regularized geometry estimation from a single image." In: *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, pp. 172–180.

Wefelscheid, C., Hänsch, R., and Hellwich, O. (2011). "Three-dimensional building reconstruction using images obtained by unmanned aerial vehicles." International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS) XXXVIII(**1**), pp. 183–188.

Wen, X., Xie, H., Liu, H., and Yan, L. (2019). "Accurate reconstruction of the LoD3 building model by integrating multi-source point clouds and oblique remote sensing imagery." ISPRS International Journal of Geo-Information 8(**3**), p. 135.

Weng, Q., Quattrochi, D., and Gamba, P. E. (2018). *Urban remote sensing*. CRC Press. ISBN: ISBN 9781138054608.

Whelan, T., Kaess, M., Johannsson, H., Fallon, M., Leonard, J. J., and McDonald, J. (2015). "Real-time large-scale dense RGB-D SLAM with volumetric fusion." The International Journal of Robotics Research 34(**4-5**), pp. 598–626.

Wiechert, A., Gruber, M., and Ponticelli, M. (2011). "UltraCam: the new super-large format digital aerial camera." In: *Proceedings of the Annual Conference of the American Society for Photogrammetry and Remote Sensing (ASPRS)*, pp. 1–7.

Wu, B., Yu, B., Wu, Q., Yao, S., Zhao, F., Mao, W., and Wu, J. (2017). "A graph-based approach for 3D building model reconstruction from airborne LiDAR point clouds." Remote Sensing 9(**1**), p. 92.

Wu, B., Xie, L., Hu, H., Zhu, Q., and Yau, E. (2018). "Integration of aerial oblique imagery and terrestrial imagery for optimized 3D modeling in urban areas." ISPRS Journal of Photogrammetry and Remote Sensing 139, pp. 119–132.

Wu, C. *Visualsfm: a Visual Structure from Motion System*. `http://ccwu.me/vsfm/`. Accessed: 2019-04-20.

Xiang, T.-Z., Xia, G.-S., and Zhang, L. (2018). "Mini-UAV-based remote sensing: techniques, applications and prospectives." arXiv preprint arXiv:1812.07770.

Xiao, J., Fang, T., Zhao, P., Lhuillier, M., and Quan, L. (2009). "Image-based street-side city modeling." ACM Transactions on Graphics (TOG) 28(**5**), p. 114.

Xu, D., Ricci, E., Ouyang, W., Wang, X., and Sebe, N. (2017a). "Multi-scale continuous CRFs as sequential deep networks for monocular depth estimation." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 161–169.

Xu, D., Wang, W., Tang, H., Liu, H., Sebe, N., and Ricci, E. (2018a). "Structured attention guided convolutional neural fields for monocular depth estimation." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3917–3925.

Xu, Y., Pan, L., Du, C., Li, J., Jing, N., and Wu, J. (2018b). "Vision-based UAVs aerial image localization: a survey." In: *Proceedings of the ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*. ACM, pp. 9–18.

Xu, Y., Yao, W., Hoegner, L., and Stilla, U. (2017b). "Segmentation of building roofs from airborne LiDAR point clouds using robust voxel-based region growing." Remote Sensing Letters 8(**11**), pp. 1062–1071.

Yamazaki, F., Matsuda, T., Denda, S., and Liu, W. (2015). "Construction of 3D models of buildings damaged by earthquakes using UAV aerial images." In: *Proceedings of the Pacific Conference on Earthquake Engineering (PCEE)*, pp. 6–8.

Yan, S., Wu, C., Wang, L., Xu, F., An, L., Guo, K., and Liu, Y. (2018). "Ddrnet: depth map denoising and refinement for consumer depth cameras using cascaded CNNs." In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 151–167.

Yang, B., Xu, W., and Dong, Z. (2013). "Automated extraction of building outlines from airborne laser scanning point clouds." IEEE Geoscience and Remote Sensing Letters 10(**6**), pp. 1399–1403.

Yang, F. and Zhou, Z. (2018). "Recovering 3D planes from a single image via convolutional neural networks." In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 85–100.

Yang, Y., Lin, Z., and Liu, F. (2016). "Stable imaging and accuracy issues of low-altitude unmanned aerial vehicle photogrammetry systems." Remote Sensing 8(**4**), p. 316.

Yao, H., Qin, R., and Chen, X. (2019). "Unmanned aerial vehicle for remote sensing applications—a review." Remote Sensing 11(**12**), p. 1443.

Yi, K. M., Trulls, E., Lepetit, V., and Fua, P. (2016). "LIFT: learned invariant feature transform." In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 467–483.

Yin, Z. and Shi, J. (2018). "GeoNet: unsupervised learning of dense depth, optical flow and camera pose." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1983–1992.

Yoder, L. and Scherer, S. (2016). "Autonomous exploration for infrastructure modeling with a micro aerial vehicle." In: *Proceedings of the International Conference on Field and Service Robotics*. Springer, pp. 427–440.

Yoon, K.-J. and Kweon, I. S. (2006). "Adaptive support-weight approach for correspondence search." IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 28(**4**), pp. 650–656.

You, X., Li, Q., Tao, D., Ou, W., and Gong, M. (2014). "Local metric learning for exemplar-based object detection." IEEE Transactions on Circuits and Systems for Video Technology 24(**8**), pp. 1265–1276.

Yu, G. and Morel, J.-M. (2011). "ASIFT: an algorithm for fully affine invariant comparison." Image Processing On Line 1, pp. 11–38.

Yuan, Y., Huang, W., Wang, X., Xu, H., Zuo, H., and Su, R. (2019). "Automated accurate registration method between UAV image and Google satellite map." Multimedia Tools and Applications, pp. 1–19.

Zamir, A. R. and Shah, M. (2014). "Image geo-localization based on multiple nearest neighbor feature matching using generalized graphs." IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 36(**8**), pp. 1546–1558.

Zarco-Tejada, P. J., Guillén-Climent, M. L., Hernández-Clemente, R., Catalina, A, González, M., and Martín, P. (2013a). "Estimating leaf carotenoid content in vineyards using high resolution hyperspectral imagery acquired from an unmanned aerial vehicle (UAV)." Agricultural and Forest Meteorology 171, pp. 281–294.

Zarco-Tejada, P. J., Catalina, A, González, M., and Martín, P. (2013b). "Relationships between net photosynthesis and steady-state chlorophyll fluorescence retrieved from airborne hyperspectral imagery." Remote Sensing of Environment 136, pp. 247–258.

Zekkos, D., Greenwood, W., Lynch, J., Manousakis, J., Athanasopoulos-Zekkos, A., Clark, M., Cook, K. L., and Saroglou, C. (2018). "Lessons learned from the application of UAV-enabled structure-from-motion photogrammetry in geotechnical engineering." International Journal of Geoengineering Case Histories (ISSMGE) 4(**4**), pp. 254–274.

Zhan, H., Garg, R., Weerasekera, C. S., Li, K., Agarwal, H., and Reid, I. (2018). "Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 340–349.

Zhang, R., Tsai, P.-S., Cryer, J. E., and Shah, M. (1999). "Shape from shading: a survey." IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 21(**8**), pp. 690–706.

Zheng, C., Cham, T.-J., and Cai, J. (2018). "T2Net: synthetic-to-realistic translation for solving single-image depth estimation tasks." In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 798–814.

Zhou, T., Brown, M., Snavely, N., and Lowe, D. G. (2017). "Unsupervised learning of depth and ego-motion from video." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6612–6619.

Zhu, L., Shen, S., Gao, X., and Hu, Z. (2018a). "Large scale urban scene modeling from MVS meshes." In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 614–629.

Zhu, L., Suomalainen, J., Liu, J., Hyyppä, J., Kaartinen, H., Haggren, H., et al. (2018b). "A review: remote sensing sensors." Multi-purposeful Application of Geospatial Data, pp. 19–42.

Zhuo, W., Salzmann, M., He, X., and Liu, M. (2015). "Indoor scene structure analysis for single image depth estimation." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 614–622.

Zhuo, X., Koch, T., Kurz, F., Fraundorfer, F., and Reinartz, P. (2017). "Automatic UAV image geo-registration by matching UAV images to georeferenced image data." Remote Sensing 9(**4**), p. 376.

Zhuo, X., Fraundorfer, F., Kurz, F., and Reinartz, P. (2018). "Optimization of OpenStreetMap building footprints based on semantic information of oblique UAV images." Remote Sensing 10(**4**), p. 624.

Zoran, D., Isola, P., Krishnan, D., and Freeman, W. T. (2015). "Learning ordinal relationships for mid-level vision." In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 388–396.

ACKNOWLEDGMENTS

First of all I would like to thank my supervisor Prof. Richard Bamler for the opportunity to pursue my PhD at the Chair of Remote Sensing Technology at the Technical University of Munich (TUM) in close cooperation with the Remote Sensing Institute of the German Aerospace Center (DLR). Your supervision and the possibility of a cooperation with DLR was the ideal basis for this doctoral thesis.

My second sincere gratitude is owed to my topical supervisor Prof. Friedrich Fraundorfer. Your continuous support of my doctoral work and your dedicated and persistent guidance during all the ups and downs of my PhD was more than essential for the outcome of this work. Thank you for our fruitful academic discussions, private conversations and coffee breaks, whether virtual via Skype or real at DLR. Friedrich, you have been an excellent supervisor and I wish you all the best for the future.

Besides my advisors, many thanks go to the members of my committee: Prof. Xiaoxiang Zhu and Prof. Markus Gerke. I am deeply honored for having had such a great committee and for your valuable feedback on my thesis.

Further I would like to thank Prof. Dr. Humberto Sossa for hosting me at the Centro de Investigación en Computación at the Instituto Polytecnical Nacional in Mexico City. This research visit was a great experience for me in terms of sharing academic knowledge and gaining new research inputs. Thank you for this interesting and valuable stay and for introducing me into your great and vivid culture.

This dissertation would not have been possible without my colleagues from TUM, DLR, and IPN, who contributed to this work in various ways. I consider myself very lucky for having had such a diverse group of wonderful colleagues around me. Marco, I very much appreciate that you have joined our team. Your knowledge in machine learning and computer vision is an excellent complement in the field of remote sensing from which everyone of us has benefited. Thank you for your support, thoughtful comments and recommendations. Xiangyu and Lukas, also many thanks to you for your great work in our successful collaborative papers. Uwe, Xiaoxiang, Richard, Ludwig, Michael and Marco, thank you for your continuous commitment and dedication to unite our three chairs and thus create a unique working atmosphere that has not only led to invaluable and inspiring discussions between different research fields, but also to a strong social team spirit. With this I would like to thank Sandra, Lukas, Marc, Rong, Alex, Yusheng, Christian, Sebastian, Richard, Chunni, Gerald, Lloyd, Karl, Achy, Yilei, Hossein, Adrian, Ksenia, and Majid for all of our conversations, coffee breaks, lunches and refreshing cold beers in the evening. Christine and Irene, thank you for your immense support in all administrative affairs.

Last but not least, I could not have completed this dissertation without the support of my family and friends. Thank you for your wise counsels and endless support.

A

ZHUO, X., KOCH, T., KURZ F., FRAUNDORFER F., REINARTZ P. (2017) AUTOMATIC UAV IMAGE GEO-REGISTRATION BY MATCHING UAV IMAGES TO GEOREFERENCED IMAGE DATA. REMOTE SENSING 9 (4), 376.

This chapter represents a pre-print version of the published article with identical content. The original article appeared under doi:10.3390/rs9040376.

A.1 INTRODUCTION

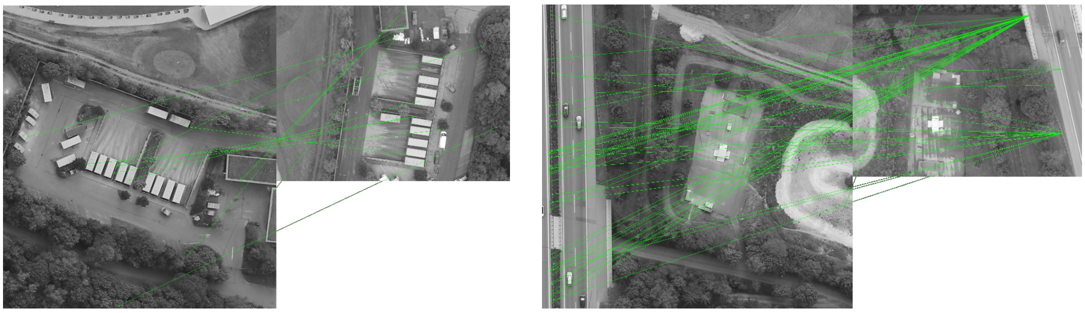
Emerging as novel image acquisition platforms, Unmanned Aerial Vehicles (UAVs) bridge the gap between aerial and terrestrial photogrammetry and offer an alternative to conventional airborne image acquisition systems. In comparison to airborne or satellite remote sensing, UAVs stand out for low cost, the utility to be used in hazardous or inaccessible areas and the ability to achieve high spatial and temporal resolutions. Table A.1 compares the main features of UAVs and manned aircrafts based on the surveys of (Eisenbeiß, 2009) and (Nex and Remondino, 2014). In contrast with manned aircrafts, UAVs have smaller coverage due to lower flight altitude, but they are able to achieve high ground sampling distance (GSD) with lower cost and better flexibility. While manned aircrafts require big landing fields and pilots, UAVs only need small landing sites and can be remotely controlled, therefore they can work even in hazardous areas and severe weather conditions. Hence, UAVs have been widely involved in remote sensing applications, such as disaster management, urban development, documentation of cultural heritage or agriculture management (Colomina and Molina, 2014).

Accurate geo-registration of UAV imagery is a prerequisite for UAV geolocalization and many photogrammetric applications, such as generating georeferenced orthophotos, 3D point clouds or DSMs. However, accurate geo-registration of UAV imagery is still an open problem. Limited by on-board payload restrictions, UAVs are equipped with lightweight GNSS/IMU systems, whose georeferencing accuracies are in the range of meters (Chiabrando et al., 2013) and far from the centimeter-level accuracy of airborne photogrammetry (Jacobsen et al., 2010; Zhao et al., 2014). In order to achieve higher geo-registration accuracy beyond hardware limits, we use a pre-georeferenced aerial or satellite image as a reference, and register the UAV image to the reference image with a novel feature-based image matching method.

In the field of image matching, numerous algorithms for different matching scenarios have been proposed in the last few decades. The biggest challenge for UAV and aerial image matching lies in the substantial differences in their scales, viewing directions and temporal changes. For instance, the flight altitude of UAV platforms is about 50 m − 120 m above the earth whereas aerial images are usually captured at 800 m − 1500 m from different viewing directions. Although state-of-the-art feature-based image matching methods are generally working fine for many different image pairs and are said to be invariant to changes in viewpoints, wider baselines and local changes of the scene, they surprisingly failed in many of our test cases. Figure A.1

Table A.1: Comparison between UAV and manned aircraft photogrammetry

	UAV Photogrammetry	**Manned Aircraft Photogrammetry**
Coverage	m^2 - km^2	km^2
Image resolution/GSD	$mm - cm$	$cm - dm$
Geo-registration possibility	low quality GNSS/IMU meter-level accuracy	high quality GNSS/IMU centimeter-level accuracy
Price and operating cost	low - moderate	high
Flexibility	applicable in hazardous areas works in cloudy/drizzly weather remotely controlled	less mobile weather-dependent pilot needed



(a) `Container`

(b) `Highway`

Figure A.1: Typical cases from the datasets (a) `Container` and (b) `Highway` showing the results of matching UAV and aerial images using SIFT, where, left of the subfigure is a downsampled UAV image and right is a cropped aerial image. Green lines indicate the matches detected by SIFT, almost all of them are wrong

illustrates two typical cases of UAV and aerial image matching using SIFT (Lowe, 2004).

Even though the scale difference has been eliminated by down sampling the UAV image towards the aerial image and the aerial image has been cropped to the same region as the UAV image, no reliable set of correct matches could be found in the similar looking image pairs. This finding motivated us to analyze the reasons for the failure and to develop a new image matching strategy facilitating a successful and robust matching of imagery with wide baselines and substantial geometrical and temporal changes. The obtained 2D matches are used for geo-registration of the UAV image with reference to the aerial image. The results demonstrate that our approach achieves decimeter-level co-registration accuracy and comparable absolute geo-registration accuracy as the reference image.

In summary, the main innovations of this paper cover following aspects:

- An exhaustive analysis of limiting cases of SIFT-based image matching for UAV and aerial image pairs. The reasons for the matching failure are identified by investigating the influence of different SIFT and ASIFT parameters, image rotations and the ratio-test.

- A novel feature-matching pipeline constituted of a dense feature detection scheme, a one-to-many matching strategy and a global geometric verification scheme.
- A comprehensive analysis of the matching quality with ground-truth correspondences and a demonstration of various experiments for evaluating absolute and relative accuracies of generated photogrammetric 3D products.

The paper is organized as follows: Section A.2 gives a review of related works; Section A.3 introduces limiting cases for SIFT matching and outlines the key factors accounting for the failure of the matching. Section A.4 proposes the novel feature matching method for a robust and reliable matching result for wide-baseline image pairs. In Section D.5, various experiments are carried out to validate the accuracy of the proposed matching method. Beside a qualitative and quantitative analysis of the obtained matches of UAV and aerial images, 3D errors of triangulated matches from geo-registered UAV images are compared towards 3D points from aerial imagery and towards terrestrial measured ground control points (GCPs). Additionally, DSMs generated from geo-registered UAV images and from aerial images are compared and a joint 3D point cloud is presented. Finally, Section A.6 discusses the applicability and limitations of the proposed method and Section A.7 concludes the paper and describes further applications.

A.2 RELATED WORK

The availability of georeferenced imagery is a prerequisite for many photogrammetric tasks, such as the generation of registered 3D point clouds, DSMs, orthorectification, mosaicking or 3D reconstructions of buildings. The key for precise georeferencing of the mentioned products lies in an accurate geo-registration of the captured images, which can be tackled in different ways. In the field of aerial photogrammetry high-end GNSS/IMU localization sensors are used which allows direct georeferencing of the images without the need of external GCPs or photogrammetric adjustments in a post-processing step. Many established systems in aerial photogrammetry have access to such accurate sensors and achieve centimeter-level registration accuracy. The relatively low-cost `DLR 3K` sensor system (Kurz et al., 2012) presents a camera frame carried by either a airplane or helicopter and consists of three Canon EOS 1Ds Mark II cameras looking in nadir, forward, and backward direction developed for real time disaster monitoring. The synchronized image acquisition and localization information provided by the expensive and heavy GNSS/IMU system (4 kg in total) allows for direct georeferencing accuracies of 10 cm (Kurz et al., 2014). The `Vexcel UltraCam` (*Vexcel UltraCam*) offers a high level optical sensor for high resolution aerial photogrammetry with more than 100 megapixel. Combined with the high-end `UltraNav`-GNSS/IMU system (*Vexcel UltraNav*), 5 cm accuracy for direct georeferencing can be achieved. Due to payload limitations, many commercial UAVs are usually equipped with lightweight sensors providing localization accuracies in the range of meters (Verhoeven et al., 2013), which is not sufficient enough for photogrammetric applications using direct georeferencing. An investigation regarding the ability of direct georeferencing with UAV systems shows that the geolocalization accuracy of current UAV systems is still too low to perform direct applications of photogrammetry at very large scale (Chiabrando et al., 2013).

For this reason, image-based methods are usually utilized to facilitate geo-registration of UAV imagery in centimeter-level accuracy. One way to augment geo-registration results is to deploy GCPs, which is even recommended for high-end devices due to the existence of systematic errors (Gerke and Przybilla, 2016). Nevertheless, the deployment of GCPs is often expensive, requires fieldwork operations and is unpractical or even impossible for hazardous or inaccessible regions. Due to the growing accessibility of high resolution aerial and satellite imagery, image matching approaches present a promising alternative for geo-registration. Here, geo-registration of UAV imagery is done by matching UAV images with georeferenced databases, such as 3D models, aerial images, orthophotos or satellite images. An accurate geo-registration of UAV images depends on the accuracy and reliability of the image matching result. Although image matching is a long-standing problem and lots of research has been performed in this area, still many cases exist where established methods fail or perform poorly. The task of matching UAV and aerial images can be characterized by wide baselines, large differences in viewpoints, and geometrical as well as temporal changes. Among intensity-based and frequency-based matching methods, local feature-based matching methods perform best with regard to these matching conditions (Zitová and Flusser, 2003). Among various feature-based matching algorithms, SIFT (Lowe, 2004) stands out for its robust scale and rotation invariant property. Although many variants and alternatives have been developed, such as its approximation SURF (Bay et al., 2008) and the binary descriptor BRIEF (Calonder et al., 2010), investigations demonstrate that SIFT is still more robust to viewpoint changes and common image disturbances than both BRIEF and SURF (Calonder et al., 2012). ORB (Rublee et al., 2011), which is a combination of the FAST detector (Rosten et al., 2010) and the BRIEF binary descriptor is a good choice for real-time applications but several evaluations state that it can not reach the repeatability and discriminative properties of SIFT (Bekele et al., 2013; Dwarakanath et al., 2012; Heinly et al., 2012; Juan and Gwun, 2009). KAZE (Alcantarilla et al., 2012) is a new development and succeeds especially in presence of deformable objects. As a variant of SIFT, a full affine invariant matching framework ASIFT (Yu and Morel, 2011) was proposed to handle big differences in viewpoints by simulating a series of transformed images to cover the whole affine space. In the case of matching images with large differences in viewpoints, ASIFT has more robust performance than SIFT, which was also confirmed in the evaluation presented in Apollonio et al. (2014).

Apart from feature-based wide baseline matching, other concepts also investigate different methods for geo-registration of UAV imagery. Intensity-based methods, like an on-board correlation-based method to register UAV images towards aerial images in case of GNSS outages (Conte and Doherty, 2009) or deformable template matching with image edges and entropy as feature representation (Fan et al., 2010) do usually not perform well in case of temporal and geometrical changes. More recent work also focus on matching terrestrial and aerial images showing extremely large viewpoints changes. A new feature representation using a Convolutional Neural Network (CNN) is learned for geolocalizing ground-level images with an aerial reference database (Lin et al., 2015). However, manual interventions are needed to estimate the scale for ground-level queries, and the absolute orientation of the query image can hardly be estimated. Shan et al. (2014) synthesizes aerial views from pre-aligned Google Steet View images using depth maps and corresponding camera poses, which are then matched with aerial images using SIFT. A similar approach is presented by Majdik et al. (2015), where UAV images are matched with geo-tagged street view images using

ASIFT achieving meter-level global accuracy. However, only low altitudes and oblique images facing building façades are considered for the geo-registration. Aicardi et al. (2016) adopts an image-based approach for co-registering multi-temporal UAV image datasets, however, it only estimates the relative transformation between the epochs, while the absolute transformation of the epoch is not solved. Finally, Xu et al. (2016) presents an fast and efficient way for UAV image mosaicking without the explicit computation of camera poses, however the image mosaics are not geo-registered.

Although considerable attempts and progress have been made regarding this topic, many of them rely on intensity-based matching methods, which are proven to be unstable in case of geometric or temporal changes. In this sense, robust image matching against large scale and viewpoint differences is the key to solve the problem for which feature-based approaches are still the methods of choice. Some mentioned approaches focus on improving the matching result for extremely large viewpoint changes, but still do not reach the desired global georeferencing accuracy.

Our approach is based on previous work (Koch et al., 2016a; Zhuo et al., 2016), which have been proven to work for complex matching scenarios with multi-scale images. Compared with the state-of-the-art works mentioned above, our method is an advancement in following aspects:

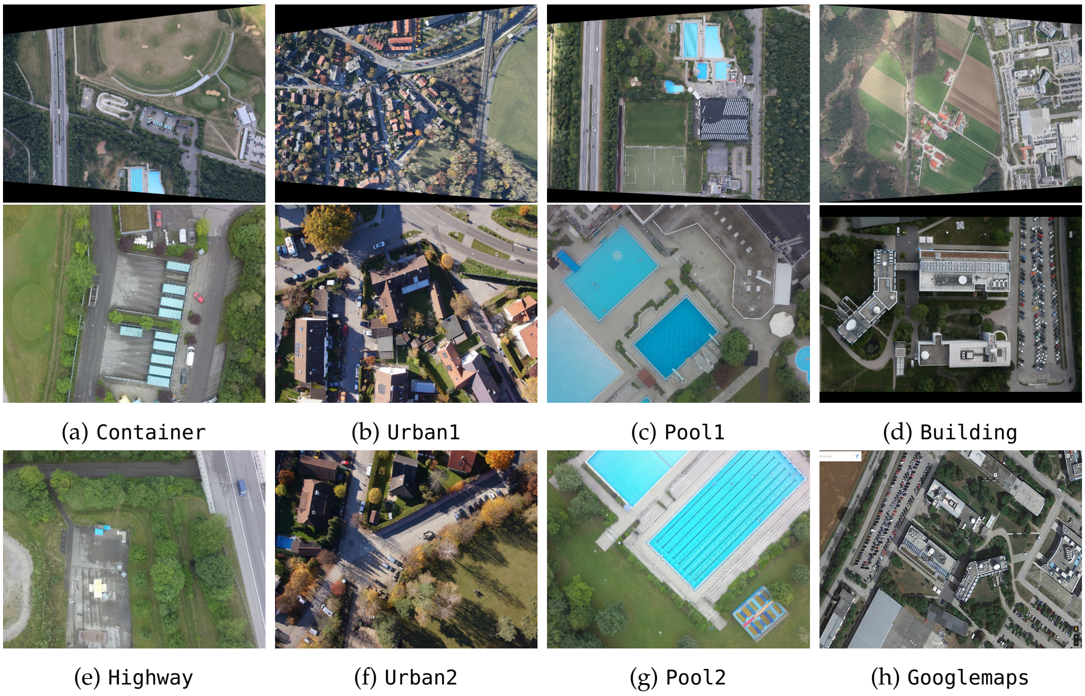
- To handle the large differences in scale and rotation between image pairs, we use a novel feature-matching approach which can overcome the challenge and robustly deliver abundant matches.
- Our method works for data of different scales, e.g., aerial images, aerial orthophotos and satellite images.
- Our method achieves not only decimeter-level co-registration accuracy, but also comparable absolute accuracy as that of the reference image, which are georeferenced in the conventional photogrammetric way.

A.3 MATCHING PERFORMANCE EVALUATION USING SIFT FEATURES

This section introduces different UAV and aerial image pairs and a comprehensive analysis of the matching performance using SIFT and ASIFT. Although one would expect that SIFT matching can successfully match the presented images, a robust and successful matching is not possible. In order to figure out why the popular SIFT matching method surprisingly fails, we analyze the influence of different SIFT parameters, such as octaves and levels, the ratio-test, but also image rotations. Experimental results demonstrate that the rotation invariance of SIFT is not as good as it has been considered to be and the deficiency in the rotation estimation of SIFT leads to non-optimal matching results. In addition to that, many correct matches are either not nearest neighbors in feature space or are rejected after applying the ratio-test.

A.3.1 *SIFT*

Among the state-of-the-art matching algorithms, SIFT has been proven to be scale and rotation invariant and outperform other local descriptors in various evaluations (Bekele et al., 2013; Dwarakanath et al., 2012; Heinly et al., 2012; Juan and Gwun, 2009). Besides, the ratio-test proposed by Lowe (2004) is widely applied to discard



(a) Container (b) Urban1 (c) Pool1 (d) Building

(e) Highway (f) Urban2 (g) Pool2 (h) Googlemaps

Figure A.2: Datasets used in this paper: Each column represents one (pre-processed) aerial reference image and two UAV target images. The UAV image in (d) should be matched to the aerial image (top right) and to a cropped part of a googlemaps image (h)

mismatches. In view of the substantial differences in scale and rotation of the UAV image and the aerial image, it makes sense to implement the SIFT matching algorithm (we use the OpenCV 3.0 implementation). This matching method is noted as "standard SIFT" in the following text.

The ratio-test discards mismatches by rejecting all potential matches with similar descriptors. It works well in most cases, however, applying the ratio-test in feature-based matching methods for images with repetitive structures often causes problems with similar descriptors. In this case, the distance ratio can be so high that these features would probably be defined as outliers. This can be critical especially when only a few correspondences remain after matching. To investigate how many correct matches are actually discarded by the ratio-test, we implemented SIFT matching and counted the correct matches before and after the ratio-test. Particularly, the distances of first two nearest neighbors are computed and compared with the threshold. Considering that the number of matches can be numerous and it is unrealistic to check every single match manually, we therefore computed the fundamental matrix between the two images with dozens of manually selected image correspondences, and then apply the epipolar constraint using the derived fundamental matrix to filter the raw matches. Afterwards, the filtered matches are again checked by manual inspection to ensure the purity of correct matches.

It needs to be pointed out that only a manually cropped part of the aerial image with almost the same image content of the UAV image was used for interest point detection, otherwise SIFT would fail to find correct matches for any dataset. This simplification of the matching problem is not feasible in practice and is only used for this analysis. The proposed method is able to match the original uncropped image pairs as this will be discussed in Section D.5.

		Levels							
		1	2	3	4	5	6	7	8
Octaves	1	12 / 50	15 / 61	15 / 64	17 / 74	17 / 84	11 / 91	17 / 78	14 / 91
	2	13 / 61	17 / 71	12 / 89	20 / 103	25 / 124	16 / 134	26 / 137	21 / 148
	3	13 / 63	17 / 76	13 / 93	22 / 108	26 / 131	17 / 142	27 / 148	22 / 153
	4	13 / 62	17 / 77	13 / 94	22 / 109	26 / 134	17 / 146	27 / 155	22 / 158
	5	13 / 62	17 / 77	13 / 93	22 / 110	26 / 136	17 / 148	27 / 157	22 / 159

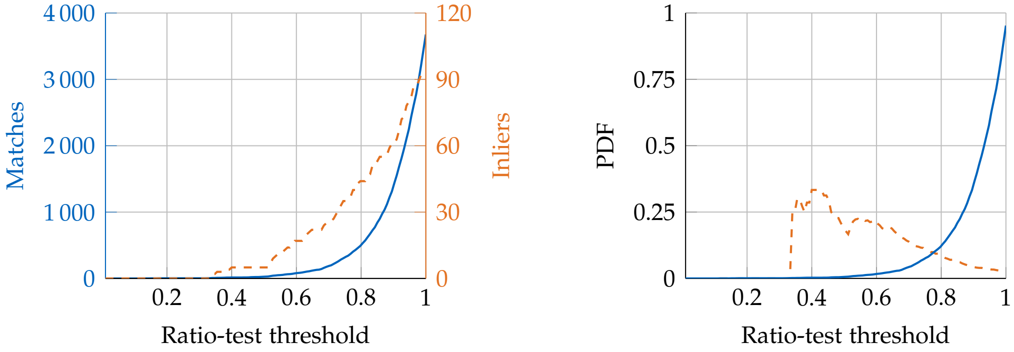
Table A.2: Analysis of SIFT performance with different octaves and levels for the `Container` dataset. Cells contain the number of correct matches (first number) from the set of remaining matches (second number) after applying the ratio-test with a fixed threshold of 0.75. Due to the scale adaption of the UAV image the number of keypoint detections saturates after two octaves. By increasing the levels more keypoints can be detected but the ratio of inliers decreases

To ensure the best matching result using the SIFT detector and descriptor we comprehensively tested different parameters. Specifically, we analyzed the effect of different ratio-test thresholds and different parameters of the SIFT detection, like the number of octaves and levels per octave. Other parameters were kept constant as they have only minor effect on the matching result. Concretely, we set the contrast threshold to 0.04, the edge threshold to 10 and the sigma of the Gaussian to 1.6. An extensive analysis was carried out for all of the datasets in Figure A.2, while only the results of the `Container` dataset is depicted. Nevertheless, we found similar results for all of our image pairs.

In a first step of our analysis, we study the effect of different numbers of octaves and levels in the SIFT detection step, while fixing the ratio-test threshold to a commonly used value of 0.75. The number of octaves is related to different image samplings, while the number of levels represent the number of scale spaces per octave and is therefore related to the amount of image blurring. Table A.2 lists the number of feasible correct matches from the set of remaining matches after applying the ratio-test for different values of octaves and levels. Due to the low image resolutions of the downsampled UAV image (664 × 885 pix) and cropped aerial image (971 × 665 pix), the number of keypoint detections saturates after two octaves. While increasing the number of levels per octave results in more matches surviving the ratio-test, the number of inliers stays constant at a very low number of around 20 matches.

According to this experimental result, we analyze different thresholds of the ratio-test in a next step while limiting the SIFT detector to three octaves and five levels. Like in the analysis above, we again count the number of remaining matches after the ratio-test and the number of inliers among them, as illustrated in Figure A.3a. A maximum number of around 100 correct matches can be found when only the first nearest neighbor is considered (equivalent to a threshold of 1). Comparing this number to the total number of around 4000 matches this is a very low ratio of inliers as can also be seen in Figure A.3b. Increasing the impact of the ratio-test (equivalent to lower values of the threshold), a lot of correct matches are rejected due to a high similarity to other keypoint descriptors, while the ratio of outliers is decreasing at the same time.

According to the results in Figure A.3b, the best ratio of inliers is suggested for threshold values between 0.3 and 0.5, but the absolute numbers of correct matches for these values is below ten and therefore not a reliable matching result. For our further analysis we choose a ratio-test threshold of 0.75, which is a good trade-off



(a) Absolute numbers of matches (blue, solid) and correct inliers (red, dashed)

(b) Probability density functions for correct (dashed) and incorrect (solid) matches

Figure A.3: Influence of different ratio-test thresholds for the `Container` dataset. (a) Number of remaining matches after applying ratio-test (solid) and number of correct matches among them (dashed). (b) Ratio of correct (dashed) and incorrect (solid) matches

Scenario	**Image fragment size (pix)**		**Keypoints**		**Correct matches**		
	Aerial	**UAV**	**Aerial**	**UAV**	**Nearest**	**Ratio-test**	**Nearest 100**
`Container`	971 × 665	664 × 885	3763	3682	81	27	690
`Highway`	617 × 908	571 × 762	2768	2560	46	22	521
`Urban1`	1197 × 1643	871 × 1307	10335	6266	47	27	304
`Urban2`	1199 × 1603	871 × 1307	9642	5757	293	176	1031
`Pool1`	838 × 1075	804 × 1071	5096	4202	87	47	451
`Pool2`	976 × 1074	799 × 1065	5788	4047	152	103	675
`Building`	1100 × 830	687 × 1030	4072	3270	76	39	498
`Googlemaps`	630 × 944	924 × 1668	3411	5963	45	21	565

Table A.3: Analysis of standard SIFT-matching on the proposed datasets in Figure A.2. Matching was performed on downsampled UAV images and cropped aerial images on the same image content of the UAV image. Keypoint detection was limited to 3 octaves and 5 levels and ratio-test threshold was set to 0.75. Results show number of feature points detected by the SIFT-detector, correct matches considering only first nearest neighbor, after applying the ratio-test and possible matches according to 100 nearest neighbors

between rejecting most of wrong matches and keeping a relatively high ratio of inliers.

Experimental results for the other datasets with these parameters are listed in Table A.3, which confirmed the difficulty of matching this kind of image pairs. Particularly in automatic registration systems for online geolocalization, it is crucial that the system is able to decide whether an image pair could be registered successfully or not. A high and reliable number of matches between 500 and 1000 is therefore indispensable for a trustable decision, compared to a rather low number below 50 like in our experiments, which could also satisfy random geometric transformations by chance.

However, the number of correct matches (using the same feature points and descriptors) can be significantly increased, if multiple nearest neighbors in feature space are considered as matching candidates. Figure A.4 shows the cumulative number of correct matches for the first 100 nearest neighbors for the `Container`

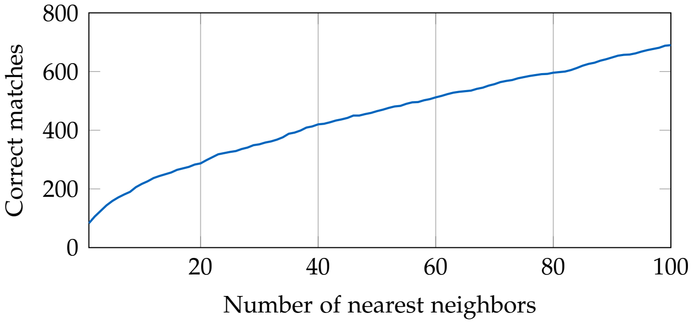


Figure A.4: Cumulative number of possible correct matches considering multiple nearest neighbors in the feature matching for the `Container` dataset.

dataset. The last column of Table A.3 lists the number of possible matches for the other datasets. This significant increase of correct matches for all datasets indicates that many corresponding keypoints in an image pair are not described perfectly by the SIFT descriptor, but can still be found among the first nearest neighbors in feature space.

A.3.2 *Influence of Rotation*

As shown in the matching results above, SIFT has unsatisfactory performance for matching UAV and aerial images. Considering the fact that the UAV and aerial images are both almost nadir view and the difference in scale has already been eliminated, the only observable difference is that the two images are not aligned in rotation. Therefore, the rotation invariance property of SIFT needs to be reconsidered and evaluated. To investigate into the problem, a series of experiments were carried out to test the influence of rotation. As listed in Table A.4, we compare the standard SIFT matching on the original unaligned images (denoted by 'Std. SIFT') from Table A.3 and on the aligned image (denoted by 'Std. SIFT Rotation aligned'); besides, instead of letting SIFT assign the orientation for each keypoint, we forced the orientation of all the detected key points in the aligned images manually to be a fixed value, here it was 0° for aligned images (denoted by 'Fixed-orientation'). The matching result was represented by the number of putative correspondences after ratio-test (denoted by 'Matches') and the correct matches among them (denoted by 'Inliers'). It is worth noting that the performance of matching between rotation-aligned images using standard SIFT does not get improved; however, the number of inliers increased substantially after we fixed the orientation of the keypoints. The experiment result shows that the rotation invariance of SIFT does not always work well, at least for the scenes in our datasets.

For further investigation into the influence of rotation, we also made a comparison with the ASIFT method, as Table A.5 shows. First, we compared the fixed-orientation SIFT with standard ASIFT on aligned images. As we achieved fewer correct matches for a tilt value of 4 at even higher computation cost, we, inspired by this finding, also fixed the orientation in ASIFT (denoted by 'Fixed-orientation') in the same way, and the matching performance get improved significantly. Comparing the results in column 2 and column 4, it can be seen that when the orientation is fixed, SIFT

Scenario	Inliers / Matches		
	Std. SIFT	Std. SIFT Rotation aligned	SIFT Rotation aligned Fixed-orientation
Container	27 / 320	22 / 349	30 / 306
Highway	22 / 204	26 / 263	52 / 277
Urban1	27 / 471	17 / 496	43 / 478
Urban2	103 / 635	179 / 677	267 / 734
Pool1	47 / 391	65 / 446	92 / 404
Pool2	103 / 635	179 / 677	267 / 734
Building	39 / 349	27 / 381	51 / 396
Googlemaps	21 / 535	21 / 509	35 / 394

Table A.4: Analysis of the influence of image-rotation on matching performace. Inliers and matches for downsampled UAV images and cropped aerial images, rotation-aligned UAV images and rotation-aligned UAV images with fixed orientation in the SIFT-detector

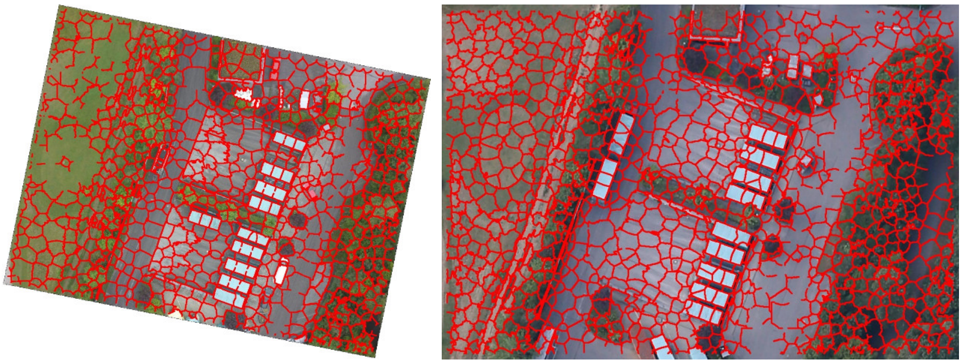
Scenario	Inliers / Matches		
	SIFT Rotation aligned Fixed-orientation	Std. ASIFT	ASIFT Rotation aligned Fixed-orientation
Container	30 / 306	25 / 281	46 / 283
Highway	52 / 227	56 / 249	70 / 237
Urban1	43 / 478	46 / 512	61 / 508
Urban2	229 / 829	254 / 1069	281 / 994
Pool1	92 / 404	73 / 346	109 / 404
Pool2	267 / 734	255 / 600	375 / 620
Building	51 / 394	45 / 382	78 / 424
Googlemaps	35 / 394	42 / 330	47 / 430

Table A.5: Comparison with ASIFT. Inliers and matches for pre-aligned images using standard SIFT with fixed orientation, ASIFT and pre-aligned images on ASIFT with fixed orientation

results in almost equivalent inliers than ASIFT, however, for a robust matching the number of inliers is still far from enough.

Based on the above findings, we summarize that the challenges of matching UAV imagery and airborne imagery stem mainly from the following aspects: inadequate matching candidates, ambiguous keypoint orientations and misuse of the ratio-test. To be more specific:

- The rotation invariance of SIFT does not work well when the images have large differences in scales and viewpoints. In standard SIFT, the dominant orientation is detected automatically. Instead, if we fix the orientations of SIFT keypoints, the number of correct matches increases significantly.
- When the image has repeated patterns, the local descriptors of the repeated structure can be so similar that the distance ratio between the nearest and second nearest neighbor is no more distinctive. As an important step in the standard matching pipeline, the ratio-test actually discards many correct matches and the remaining correspondences are not reliable. In contrast, considering



(a) UAV image

(b) Aerial Image

Figure A.5: Feature points highlighted in red, namely all the pixels at the boundaries of superpixels, after removing those feature points located at homogeneous areas for (a) the pre-aligned UAV image and (b) the aerial image of the `Container` dataset with 1000 SLIC superpixels

multiple nearest neighbors as matching hypotheses can help to increase the matching performance enormously.

A.4 PROPOSED IMAGE MATCHING METHOD

According to the reasons of the matching failure presented in Section A.3, the new matching approach is designed to eliminate each of the exposed bottlenecks. A new feature detection scheme increases the number of matchable keypoints which is necessary for a reliable matching result. To avoid loosing many correct matches which are not nearest neighbor in feature space or which are rejected by the ratio-test, we introduce a one-to-many matching scheme. To extract correct matches among them, a direct method using histogram voting is performed instead of the commonly used RANSAC scheme. An extension of this method can also handle unknown image rotations. In the end, the detected matches are used to estimate camera poses of the UAV images in the coordinate system of the reference images.

A.4.1 *Prerequisites*

The proposed method assumes that the scale difference between both images can be estimated and mostly eliminated in advance. This requirement can be generally fulfilled, as accurate positional information of aerial images is always available and UAV images are tagged with both GNSS and barometric altitude information. One of both sensors should deliver reliable data in any case.

Secondly, a pre-alignment with respect to the image rotation can be achieved using the on-board compass of the UAV. The next sections assume that a rough pre-alignment of the image pairs is feasible, but in case of no or only imprecise image heading information, Section A.4.5 presents an extension of the proposed method which allows to recover an unknown image rotation.

A.4.2 *Dense Feature Extraction*

The essential prerequisites of robust matching are sufficient and uniformly distributed features whose density should reflect the information content of the image. According to the results in Table A.4 and Table A.5, established keypoint detectors, such as in SIFT, do not always find a sufficient number of matchable features. To ensure a large number of inliers, a dense detection scheme is desired, but instead of using all pixels as potential feature points, only keypoints should be considered which are located along strong image gradients. This does not only reduce computational time but also rejects hardly matchable feature points at homogeneous areas with weak descriptors.

In view of the fact that image segmentation using SLIC (Simple Linear Iterative Clustering) (Achanta et al., 2012) can efficiently generate compact and highly uniform superpixels, whose boundaries mostly define strong variations in the intensities of the local neighborhood, like edges and corners, we therefore adopt all the pixels at the boundaries of superpixels as feature points. In practice, the number of desired superpixels can be specified according to the need for feature density and compactness. Since the relative scale difference of both images is known beforehand, the number and compactness of superpixels in both images are similar and therefore ensures the extraction of identical object boundaries. Figure A.5 highlights the feature points of a UAV and aerial image, namely all the pixels at the boundaries of superpixels, after removing those feature points located at homogeneous areas.

Afterwards a SIFT-descriptor for each detected feature point is computed. Since the UAV image is already aligned with the reference image, the scale space and feature orientation of SIFT-descriptors should be identically assigned for both images.

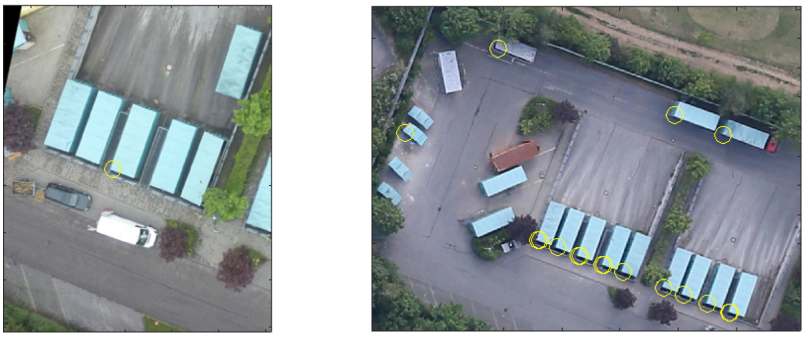
A.4.3 *One-to-Many Feature Matching*

In this phase, a feature descriptor in one image is matched with all other features in the other image using the euclidean distance calculation. In standard SIFT, only the first and second nearest neighbors are taken into account, so that many correct matches are actually discarded as presented in Table A.3. An example of ambiguous feature matching is demonstrated in Figure A.6. The correct feature point (left) would mainly be discarded for two reasons: first, the correct match may not be the first nearest neighbor in feature space; second, it may not pass the ratio-test due to the high similarity of the local descriptors.

To solve this problem, we propose a one-to-many matching scheme by taking the k-nearest neighbors as matching candidates to ensure that correct matches can be even found for corresponding keypoints which do not show nearest descriptors distances. Besides, the approximate nearest neighbor method (ANN) is applied to avoid the exhaustive search and to speed up the matching process. Although the idea of using a one-to-many matching scheme is not new, the next section proposes a new approach how to extract the correct matches among them.

A.4.4 *Geometric Match Verification with Histogram Voting*

It is pointed out in Section A.3 that the commonly used ratio-test in SIFT does not effectively determine whether a feature point is a correct match. As a substitute,



(a) Feature point in the UAV image

(b) Corresponding feature points in the aerial image

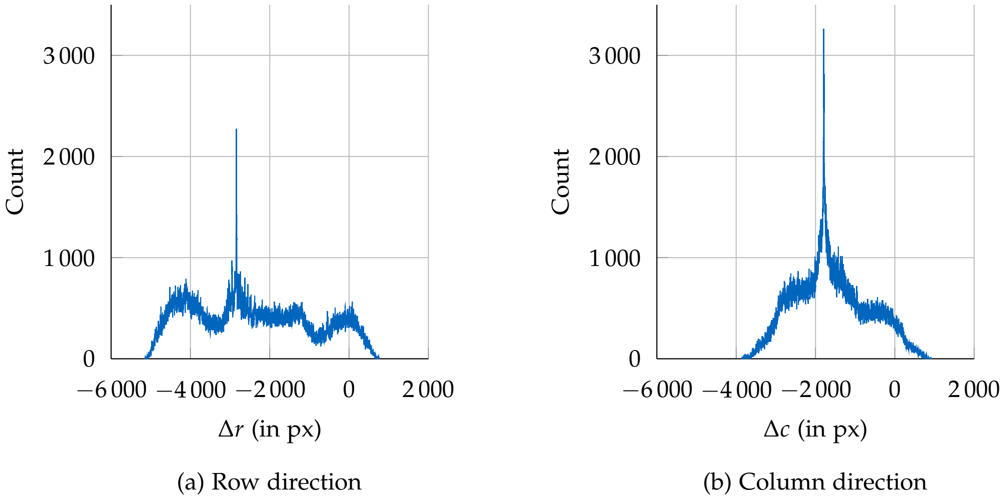
Figure A.6: Challenge of ambiguous feature matching. Feature points in the aerial image with the closest descriptor distances in image (b) to a feature point at the corner of a container in the UAV image (a). The correct match often can be found among a set of multiple nearest neighbors. These ambiguities need to be solved in order to extract the correct match

we use pixel-distances as a global geometric constraint to verify the matching hypotheses. The superpixel-based feature point extraction and one-to-many matching strategy result in a plethora of putative matches, which ensures a sufficient number of correct matches but also inevitably contains a massive number of mismatches. Postulating that the UAV and reference image both contain the same planar scene and the differences in their scales and rotations have already been eliminated, the transformation between the two aligned images can be simply approximated as a 2D-translation. Particularly, for each keypoint i, whose image coordinates are (x_u^i, y_u^i) in the UAV image and (x_r^i, y_r^i) in the reference image, and for each of its k matching hypothesis j $(j = 1 : k)$, whose image coordinates are (x_r^j, y_r^j) in the reference image, we calculate their coordinate differences $\Delta x^{i,j}$ and $\Delta y^{i,j}$ by $\Delta x^{i,j} = x_u^i - x_r^{i,j}$ and $\Delta y^{i,j} = y_u^i - y_r^{i,j}$. Correct matches are expected to satisfy the conditions $|T_x - \Delta x^{i,j}| \leq R \wedge |T_y - \Delta y^{i,j}| \leq R$, where R is a threshold related with the scene depth and T_x and T_y are the parameters of the unknown 2D translation. We can recover this translation by a simple histogram voting scheme. After computing $\Delta x^{i,j}$ and $\Delta y^{i,j}$ for all putative matches, distinctive peaks T_x and T_y in the both histograms are extracted.

Figure A.7 presents an example for this histogram voting regarding the `Container` scenario. While distances of wrong matches are randomly distributed, those of geometrically correct matches concentrate on or aggregate around a common value (T_x, T_y), thus shaping a distinct peak in the histogram. To allow for minor changes of image scene depth, we determine the matches located at close range to (T_x, T_y) as possibly correct matches, the distance threshold is denoted by R. The value of R is related to the change of scene depth as well as the accuracy of pre-alignment. A larger threshold R can compensate for these impacts and result in more matches, on the other hand, more outliers would also be introduced into the raw matches.

A.4.5 *Eliminating Differences in Image Rotation*

The scale difference between the UAV image and the reference image can be derived using either the on-board GNSS information or the barometric altitude sensor. In contrast, precise orientation-adaption fails for many UAVs due to inaccurate



(a) Row direction

(b) Column direction

Figure A.7: Geometric match verification of the `Container` scenario with histogram voting. Distribution of pixel distances for all putative matches according to the one-to-many matching in (a) row- and (b) column- direction. Distinct peaks represent unknown 2D-translation

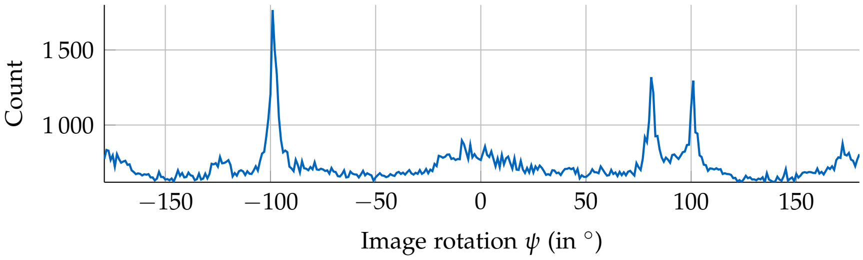


Figure A.8: Recovering the unknown image rotation in case of unavailable or inaccurate UAV IMU data. Extending proposed method by transforming UAV feature points with multiple rotation values before the histogram voting step. Figure shows the rotation histogram for the `Container` dataset. Maximum number of raw matches represents unknown image rotation

heading information provided by the low quality IMUs. Our assumption of correct matches follow a simple 2D translation fails in case of unaligned images. However, we can estimate the unknown image rotation by adapting the proposed matching approach with a rotation search scheme. Although Section A.3.2 shows, that fixing the orientation of the feature points in the SIFT descriptors results in a better matching performance if the images are pre-aligned, a sufficient number of correct matches can still be found for unaligned images with the keypoint orientation estimation of SIFT when using a denser feature detection like the one presented in Section A.4.2. After generating a set of putative one-to-many matches for unaligned images, the unknown image rotation is obtained by first dividing the rotation ψ equally into discrete rotation values $\psi^a = [-180, 180[$ deg. For each rotation ψ^a the feature points of the UAV images $pt_u^i = (x_u^i, y_u^i, 1)^T$ are rotated around the image center $pt_{u,rot}^{i,a} = M(p, \psi^a) \cdot pt_u^i$ with a transformation matrix $M(p, \psi^a) = [T(p)R(\psi^a)T(-p)]$, where $T(p)$ is a translation matrix with the coordinates of the image center p and R a rotation matrix with rotation angle ψ^a. Pixel distances are calculated according to $\Delta x_{rot}^{i,j,a} = x_{u,rot}^{i,a} - x_r^{i,j}$ and $\Delta y_{rot}^{i,j,a} = y_{u,rot}^{i,a} - y_r^{i,j}$ and histogram voting from Section A.4.4 is performed for each rotation. The maximum number of raw matches satisfying the threshold T_x^a and T_y^a is kept for all rotation values ψ^a. Figure A.8 shows the number of raw matches for different image rotations according to the `Container` dataset. The distinct peak at -104 deg represents the unknown image rotation.

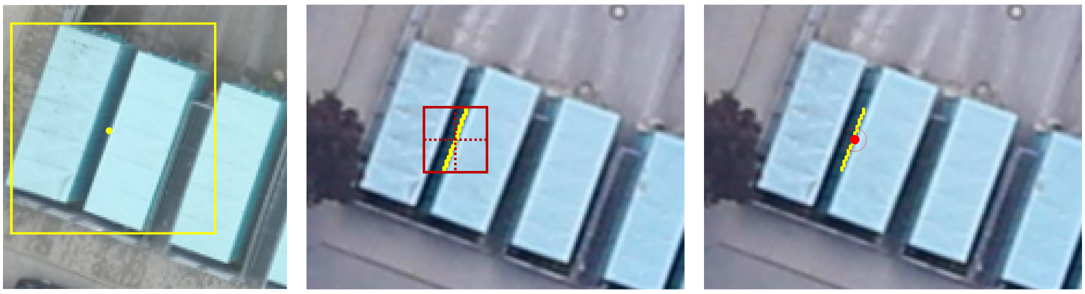
This method may be used for a full 360 deg search, however, the search range can be reduced in case of available inaccurate rotations from the on-board IMU. After recovering the unknown image rotation, further matches can be determined with fixed orientations according to the previous sections.

A.4.6 *Match Refinement*

After the geometric verification of the one-to-many matches, it is likely for some keypoints that they share multiple adjacent feature points in the other image as geometric correct matches. This is caused by the dense feature point extraction, which generates dense feature points especially along strong image edges. The distance threshold R allows multiple geometric correct matches for adjacent feature points for which the distance to T_x and T_y is below R. Figure A.9 illustrates these local ambiguities of the feature matches. One feature point in the UAV image in Figure A.9a corresponds to multiple geometrical correct matches in the aerial image in Figure A.9b . Even a successive RANSAC filtering step according to geometrical transformations will not truly solve these ambiguities, if Sampson distances or transfer errors of neighboring matches are below the filtering threshold. In order to ensure geometrical correct and unique one-to-one matches, a refinement step is applied for all geometrical correct matches by eliminating the ambiguities and optimizing the location of the feature points. The superpixel segmentation cannot guarantee exact locations of corresponding pixels in both images. The refinement consists of a NCC matching of a template in the local neighborhood of the UAV feature point (yellow rectangle in Figure A.9a). For all corresponding matching hypotheses (yellow dots in Figure A.9b), the corresponding patch is searched in a local search window around the feature points (red rectangle in Figure A.9b). The size of the search window for all aerial feature points can be set to the threshold R of the geometric verification. The NCC optimizes all matching hypotheses to the correct location, illustrated by the red dot in Figure A.9c. This method eliminates duplicate matches and refines feature point locations for inaccurate keypoints in a local neighborhood of the initial keypoints. These raw matches can now be used to estimate the fundamental matrix or homography in combination with RANSAC methods and to reject remaining outliers satisfying the geometric constraint. After computing the fundamental matrix, a guided matching method, as presented in Section 3, can be applied to find more matches if the threshold was chosen too small.

A.4.7 *Geo-registration of UAV Images*

As the UAV image and the reference image have overlapping areas, one 3D point in the object space could be visible both in the reference image and the UAV image. Such 3D points can be used as reference 3D points for geo-registration of UAV images. The prerequisite of the geo-registration is available georeferenced aerial image together with its heightmap, or one orthorectified mosaic with a high resolution DSM.



(a) UAV feature point and template size

(b) Aerial geometric inliers and size of the search window

(c) Refined aerial match (red) as shared optimized pixel location

Figure A.9: Refinement and duplicate elimination of geometric correct matches. (a) One feature point in the UAV image (yellow dot) and its template size (rectangle). (b) Corresponding geometric matches in the aerial image and search window for one match (red rectangle). (c) Refinement of all feature matches to the correct matching location (red dot)

- Match a UAV image U with the reference image R using the proposed matching method. Assume a feature point (x_r, y_r) in the reference image is matched to feature point (x_u, y_u) in the UAV images, this matching pair correspond to a 3D point $P(X, Y, Z)$ in the object space.
- If image R is an individual georeferenced aerial or satellite image, we assume its height map is available, which can be generated in the process of dense matching with neighboring images (d'Angelo and Reinartz, 2011). The height Z can be looked up in the height map and the planar coordinates X and Y can be calculated using the orientation parameters of R. If image R is an aerial orthophoto which is generated by an orthographic projection of the aerial image mosaic onto a high resolution DSM, the planar coordinates (X, Y) are namely the corresponding georeferenced coordinates of the pixel (x_r, y_r) in the orthophoto, and Z is namely the corresponding height at (X, Y) of the DSM.
- As the proposed matching method generates thousands of matches and each match results in a 3D point, those points can be used as reference 3D points to transform the UAV image to the same global coordinate system of the reference image. If there are UAV images sequences, a bundle adjustment can be performed to improve the global geo-registration accuracy.

A.5 EXPERIMENTS

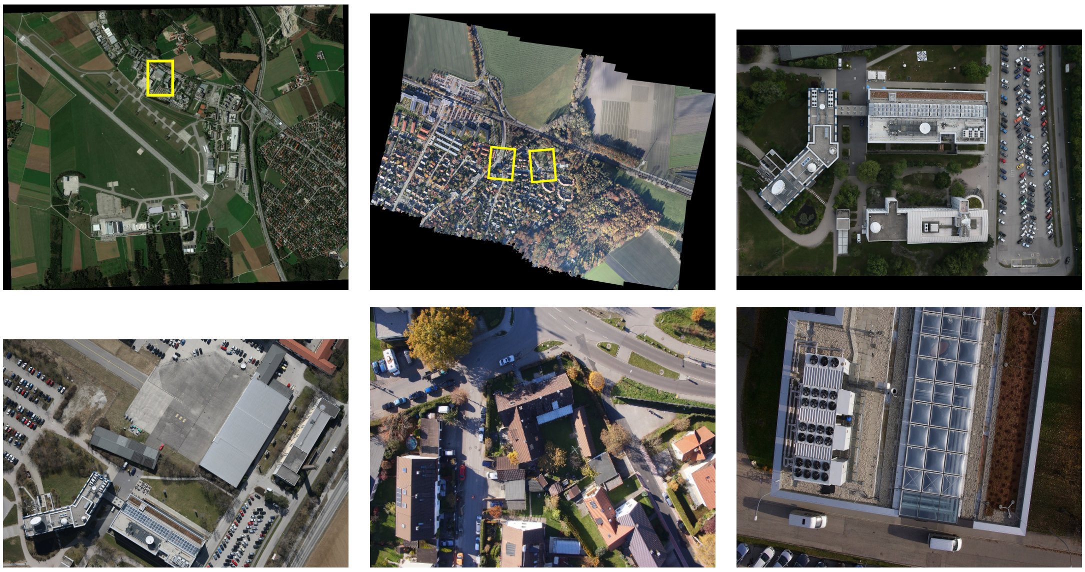
In order to verify the robustness and reliability of the proposed matching method, we compare the performance of our method with standard SIFT on different datasets. Furthermore, the generated matches are used for geo-registration and 3D reconstruction of the UAV images. Qualitative and quantitative analyses are presented to validate the accuracy of geo-registration, and on this basis, photogrammetric 3D products, such as orthophots, DSMs and merged points clouds are discussed.

A.5.1 *Data Acquisition*

Experiments were carried out based on offline flight data of four datasets: `Eichenau`, `Germering`, `EOC` and `WV2`. It is worth noting that for datasets `Eichenau`, `Germering`

Dataset	Reference Image				Target Image			
	Type / Date	Resolution (pix)	Height (m)	GSD (cm)	Type / Date	Resolution (pix)	Height (m)	GSD (cm)
Eichenau	AO 11/2015	9206 × 7357	600	20	UI 11/2015	573 × 794	100	1.8
Germering	AI 06/2014	5184 × 3902	700	9.4	UI 07/2014	823 × 996	100	2
EOC	AI 06/2014	5184 × 3902	340	4.6	UI 11/2014	1106 × 807	25-40	0.5-0.8
WV2	SI 2010	5292 × 6410	770, 000	46	AI 2015	497 × 332	350	4.4

Table A.6: Characteristics of the datasets used in the experiment. Target images are pre-aligned towards the reference image using GNSS/IMU data. AI: aerial imagery; AO: aerial orthophoto; SI: satellite imagery; UI: UAV imagery



(a) WV2

(b) Eichenau

(c) EOC

Figure A.10: Additional datasets for the experiment. Top: reference images. Bottom: target images. Overlapping areas are highlighted by yellow rectangles in the reference images

and EOC, which contains 72, 58 and 11 UAV images respectively, the whole UAV sequences were matched in an automatic manner. Showing the results for all image pairs is beyond the scope of this paper, so we focused on the same image pairs which were already introduced in Section 3. The Eichenau dataset contains two scenarios: Urban1 and Urban2. The UAV images were acquired with a Sony Nex-7 camera simultaneously with the reference aerial images on November 2nd, 2015. For both scenarios, we matched UAV images not only to aerial images but also to aerial orthophotos, which are generated by an orthographic projection onto a high resolution DSM (Hirschmuller, 2008; d'Angelo and Reinartz, 2011). The Germering dataset is comprised of four different scenarios: Container, Highway, Pool1 and Pool2. The reference aerial images of this dataset were captured on June 17th, 2014, whereas the UAV images were captured with a slight time delay on July 11th, 2014 with a GoPro Hero 3+ Black camera. The aerial images in the EOC dataset were acquired on June 16th, 2014 and the UAV images were captured on November 12th, 2014 with a Sony Nex-7 camera. In EOC dataset, all aerial images are almost nadir whereas the UAV images have both nadir views of the building roof and oblique

Scenario	Raw matches (SIFT)	Inliers F / Error (F)	Inliers H / Error (H)
Container	58	14 / 666.26	9 / 1767.55
Highway	49	15 / 1996.30	9 / 2210.20
Pool1	162	52 / 0.83	33 / 1.63
Pool2	107	18 / 618.54	10 / 1308.02
Eichenau1	287	45 / 19.11	48 / 3.63
Eichenau2	436	140 / 1.11	146 / 3.64
EOC	446	16 / 959.87	6 / 877.21
WV2	117	19 / 175.73	19 / 4.03
Building	553	16 / 595.06	11 / 317.59
Googlemaps	522	19 / 195.34	8 / 919.48

Table A.7: Results using standard SIFT: number of raw matches after applying SIFT for all scenarios. Inliers after estimating fundamental matrix (F) and homography (H) using RANSAC. Mean errors (in pixel) according to ground-truth F and H

views of the building façades. Only the nadir-view UAV images are matched with the aerial images, and the generated GCPs are used to geo-register the whole UAV image blocks including both nadir and oblique images. In addition, the nadir UAV images are also matched with a screenshot of Google Maps. In the WV2 dataset (Koch et al., 2016b), we match an aerial image from the EOC dataset with a WorldView-2 RGB satellite image of the year 2010 to validate the generalization ability of the proposed method and its robustness against large temporal changes. Besides, the datasets Eichenau, Germering and EOC are not significantly affected by temporal changes, as the vegetation periods are the same (except in EOC) and the appearances of buildings has not changed. All the aerial images were captured by a Canon EOS-1DX camera mounted on the DLR 4K sensor system (Kurz et al., 2014), which consists of two cameras with 15° sidewards looking angle and a FOV of 75° across. In data pre-processing, an orthographic projection of the aerial imagery was performed to generate nadir-view images. Figure A.2 and Figure A.10 illustrate all datasets used in the experiments, where the first row shows the reference images (pre-processed nadir-view aerial images and satellite image), and the other two rows are the corresponding target images (UAV and aerial images) to be matched. Detailed characteristics of the datasets are listed in Table A.6.

A.5.2 *Performance Test of Matching UAV Images with a Reference Image*

In order to validate the robustness and accuracy of the proposed method, we use the same image pairs presented in Section A.3, where the standard SIFT performed poorly in most of the cases. Different from the results in Table A.3, the matching is now performed with original aerial images other than the cropped images. As can be seen in Figure A.2, only a small portion of the aerial images is pictured in the UAV images. Thus, it is also tested if the matching benefits from our geometric constraints in the presence of large searching areas.

All image pairs are provided with rough information of positions and orientations from GNSS and IMU so that the images could be pre-aligned beforehand. Then the target images and the reference images were matched with the proposed matching method and standard SIFT. Specifically, 750 superpixels were segmented from the

Scenario	Raw matches (our)	Inliers F / Error (F)	Inliers H / Error (H)
Container	8264	4876 / 2.59	2835 / 7.01
Highway	1979	1184 / 2.79	1230 / 1.20
Pool1	6593	3599 / 1.87	2188 / 1.87
Pool2	14091	7555 / 2.01	4199 / 2.03
Eichenau1	4018	1850 / 4.35	1165 / 3.53
Eichenau2	5846	3204 / 1.09	3077 / 4.65
EOC	6834	3949 / 2.92	2586 / 3.18
WV2	15131	6290 / 2.22	6760 / 3.57
Building	9113	3526 / 3.15	1932 / 2.36
Googlemaps	15437	5120 / 3.42	3217 / 2.82

Table A.8: Results using proposed method: number of raw matches after applying our method for all scenarios. Inliers after estimating fundamental matrix (F) and homography (H) using RANSAC. Mean errors (in pixel) according to ground-truth F and H

UAV images, the threshold for the feature matching-distance was set to 0.2 as a trade-off between discarding apparent outliers and retaining enough matching hypotheses. 50 nearest neighbors were selected as matching candidates for the one-to-many matching and the distance threshold R for the geometric verification was set to 12 pixels. As for matching using SIFT, the threshold of ratio-test was set to 0.75.

In order to evaluate the matching accuracy, we created ground-truths of feature point correspondences for each dataset using manually selected and automatically detected matching correspondences. The quantitative results using standard SIFT and our proposed method are summarized in Tables A.7 and A.8, where Error (H) denotes the mean transfer error (the Euclidean distance between a point's true correspondence and the point mapped by the homography matrix H, which is estimated from matching correspondences) and Error (F) denotes the mean Sampson distance (the distance between a point to the corresponding epipolar line). Standard SIFT failed for almost all scenarios while the proposed method found abundant matches with much smaller errors.

Regarding matching accuracy, standard SIFT outperformed our method only at Pool1 scenario. As homography only considers transformation between two planes, those mismatches at areas with apparently different scene depths were discarded. The mean transfer error were only 2-3 pixels in most cases, corresponding to a ground distance of about 20-30 cm.

The matched feature points are marked in the UAV images (the second and third rows in Figure A.11). As a result of the superpixel segmentation, most matches are located at regions with rich textures and have apparently much higher density than SIFT-features. The projection transformations can then be estimated using these matches. The first row in Figure A.11 depicts the projected UAV images on the aerial images by the estimated homography.

A.5.3 *Evaluation of Geo-registration of UAV Images*

Following the proposed pipeline in Section A.4.7, plenty of 3D reference points were computed and then used as GCPs in a bundle block adjustment to geo-register the UAV images to the global coordinate frame.

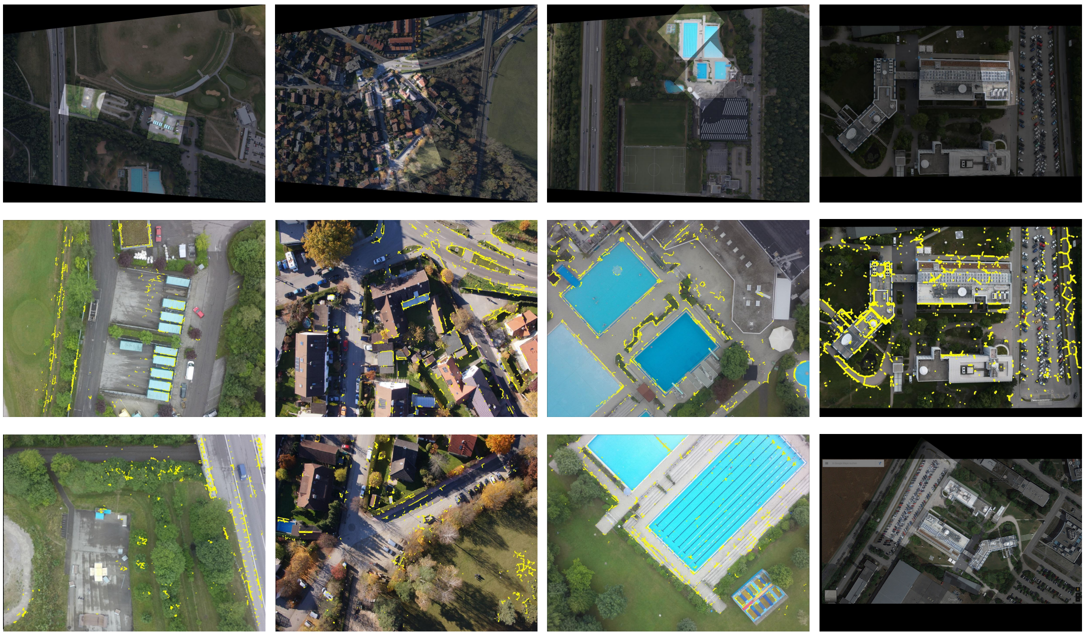


Figure A.11: Qualitative results of the proposed matching method according to the image pairs in Figure A.2. First row shows the overlapped UAV and aerial image pairs after applying an estimated homography calculated from our matches (also for the figure on the bottom right). Second and third row show the distribution of the geometrical correct matches in the UAV images (yellow dots)

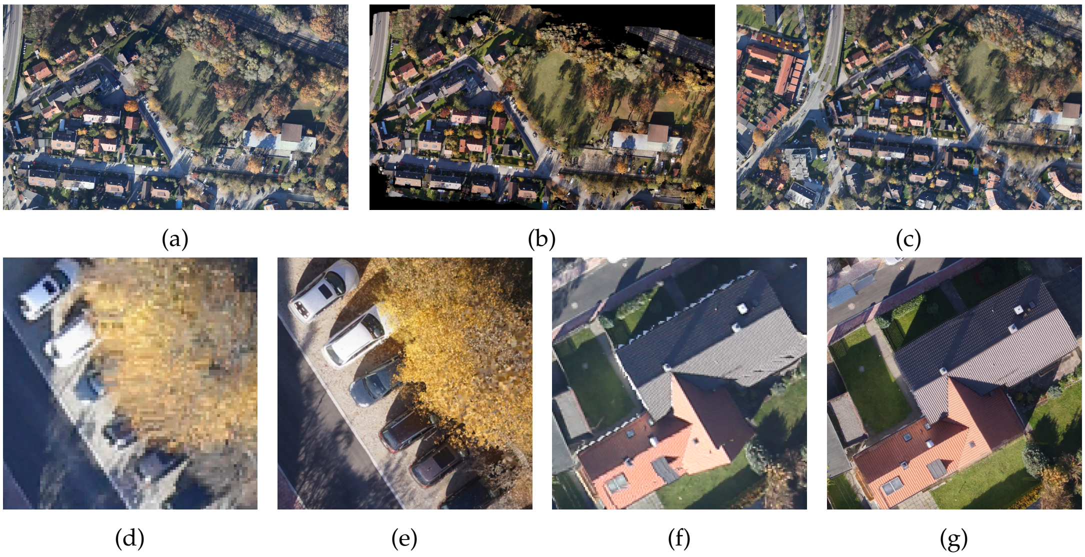


Figure A.12: Comparison of (a) aerial orthophoto with 20 cm GSD and (b) UAV orthophoto with 2 cm GSD of the `Eichenau` dataset. (c) 50% transparent overlap of both orthophotos; (d) and (e) compare cars and (f) and (g) show a roof on aerial and UAV orthophoto respectively

In order to verify the accuracy of geo-registration of UAV images, several evenly-distributed ground check points were selected across the survey area and their actual coordinates P_{rtk} were measured using a RTK GNSS receiver. Meanwhile, these ground check points were marked in all UAV images and their theoretical 3D coordinates P_{uav} were computed by triangulating the geo-registered UAV images. The column "$Error_{rtk}$" in Table A.9 and Table A.10 lists the errors $P_{uav} - P_{rtk}$ of `Eichenau`

Check Point	$\text{Error}_{\text{ref}}(\text{m})$			$\text{Error}_{\text{rtk}}(\text{m})$		
	Δx	Δy	Δz	Δx	Δy	Δz
1	0.04	-0.51	-0.21	-0.04	-0.39	-1.74
2	-0.05	-0.07	-0.15	-0.11	-0.40	-1.90
3	0.04	-0.41	-0.36	-0.10	-0.83	-2.04
4	-0.14	0.80	0.70	-0.35	-0.33	-1.91
5	-0.04	0.49	-0.17	-0.05	-0.21	-1.81
6	-0.03	0.12	-0.10	0.12	-0.36	-1.63

Table A.9: Errors of the coordinates of check points comparing to RTK GNSS measurements and the coordinates looked up in aerial orthophoto and DSM - `Eichenau` dataset

Check Point	$\text{Error}_{\text{ref}}(\text{m})$			$\text{Error}_{\text{rtk}}(\text{m})$		
	Δx	Δy	Δz	Δx	Δy	Δz
1	-0.06	-0.14	-0.38	0.34	-0.01	1.49
2	0.16	-0.67	0.37	0.43	-0.54	1.68
3	0.14	-0.02	0.46	0.56	0.16	1.76
4	0.11	-0.76	0.26	0.44	-0.76	1.71
5	0.19	-0.10	0.50	0.55	-0.06	0.75
6	-0.05	0.18	0.18	0.39	0.36	1.30
7	-0.08	0.41	-0.06	0.41	0.50	1.42

Table A.10: Errors of the coordinates of check points comparing to RTK GNSS measurements and the coordinates triangulated using aerial images - `Germering` dataset

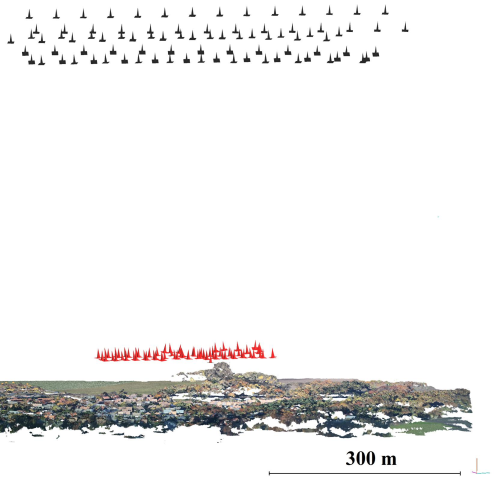


Figure A.13: Camera pose visualization for `Eichenau` dataset, showing camera poses (red) of the geo-registered UAV image blocks at 100 m altitude and the aerial image (black) blocks at 600 m altitude

and `Germering` datasets. The height errors in "$Error_{rtk}$" are around 2 meters, this is mainly caused by the systematic errors of the global digital elevation model like SRTM (Rabus et al., 2003), which was used as height reference during the processing of the reference images.

In order to validate accuracy of co-registration, the coordinates triangulated by geo-registered UAV images, P_{uav}, were compared with the identical points on the

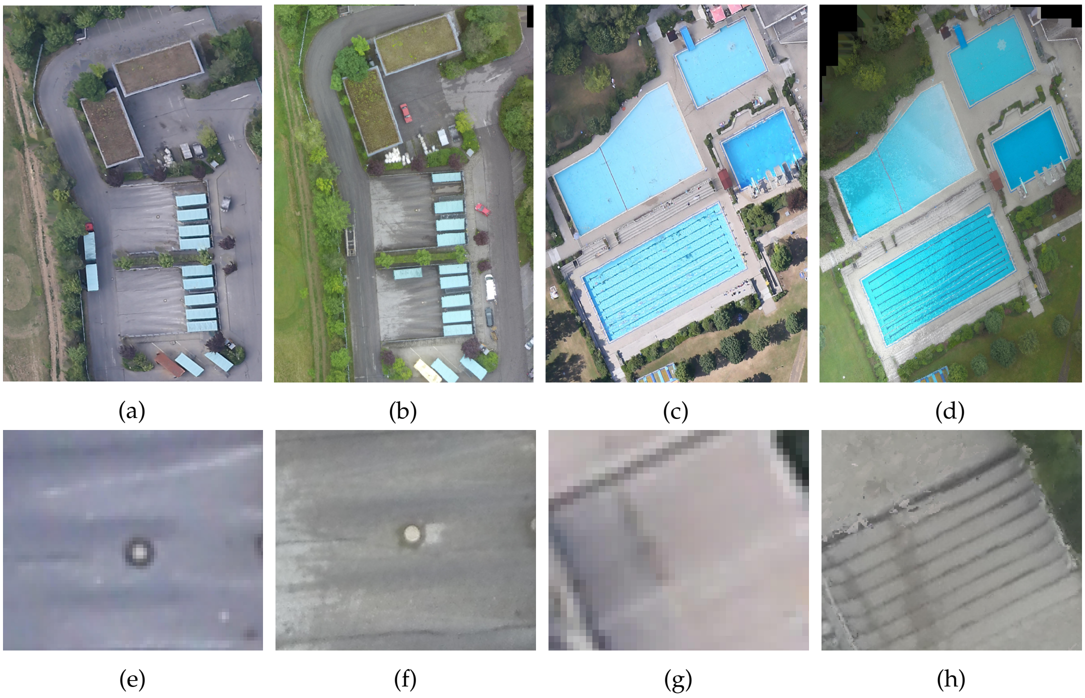
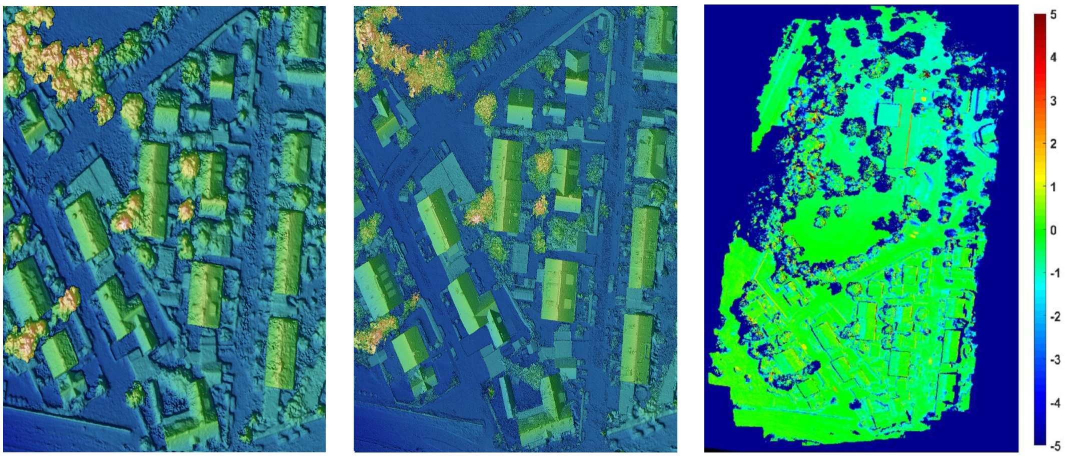


Figure A.14: Comparison of (a + c) aerial orthophotos with 20 cm GSD and (b + d) UAV orthophotos with 2 cm GSD of the `Germering` dataset; (e) and (f) compare a manhole and (g) and (h) staircases on aerial and UAV orthophoto respectively

reference image as well. In `Eichenau` dataset the reference image was an aerial orthophoto (with a high resolution DSM), so the corresponding coordinates P_{ref} were manually looked up in the orthophoto and DSM, as explained in Section A.4.7. In `Germering` dataset the reference image was an individual aerial image from a pre-georeferenced aerial images dataset, so the corresponding coordinates P_{ref} were triangulated using multiple pre-georeferenced aerial images from that dataset. The column "$Error_{ref}$" in Table A.9 and Table A.10 lists the error $P_{uav} - P_{ref}$.

Afterwards the orthophoto and DSM were reconstructed from the geo-registered UAV images using the software SURE (Rothermel et al., 2012). Figure A.12 illustrates the aerial orthophoto and the UAV orthophoto of `Eichenau` dataset. More specifically, (a) depicts the aerial orthophoto of the `Eichenau` dataset, whose resolution is 20 cm; (b) shows the UAV orthophoto of the `Eichenau` dataset, whose resolution is 2 cm. It is obvious that the UAV orthophoto has higher resolution and contains more details than the aerial orthophoto. (c) displays the UAV orthophoto overlapping on the aerial orthophoto with 50% transparency, it can be seen that the the two orthophotos are precisely aligned using the proposed geo-registration method. (d) and (e), (f) and (g) compare the appearance of corresponding objects on aerial orthophoto and UAV orthophoto, demonstrating that the UAV orthophoto contains richer textures than the aerial orthophoto. Figure A.13 illustrates the estimated camera poses as well as the reconstructed point cloud of the geo-registered UAV image blocks and the aerial image blocks. Despite the considerable scale difference, our matching approach still succeeds in an accurate registration. Similarly, Figure A.14 demonstrates the aerial orthophoto and UAV orthophoto of `Germering` dataset, whose resolutions are 20 cm and 2 cm respectively.

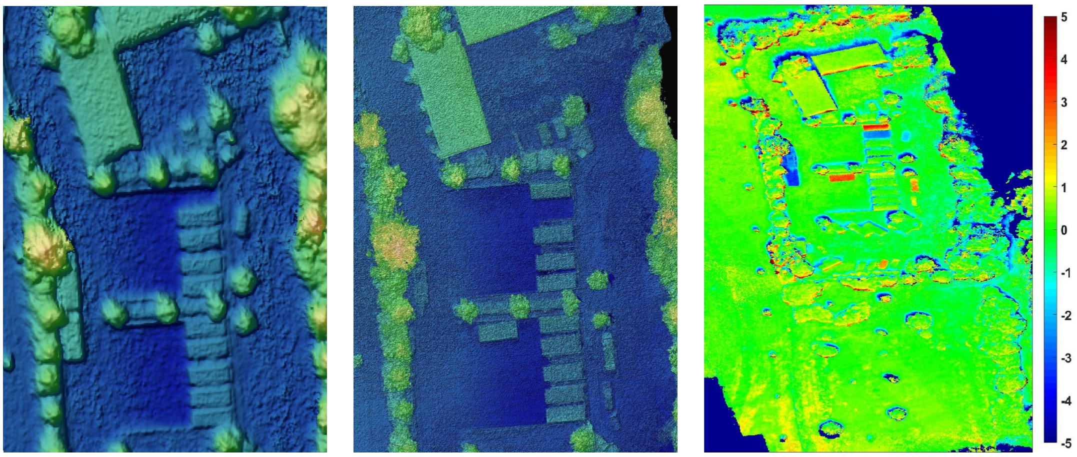


(a) Aerial DSM

(b) UAV DSM

(c) DSM differences

Figure A.15: Comparison of (a) aerial and (b) UAV DSM of `Eichenau` dataset. 20 cm GSD for aerial and 2 cm GSD for UAV DSM; (c) colormap illustrating the height differences between the two DSMs in meters



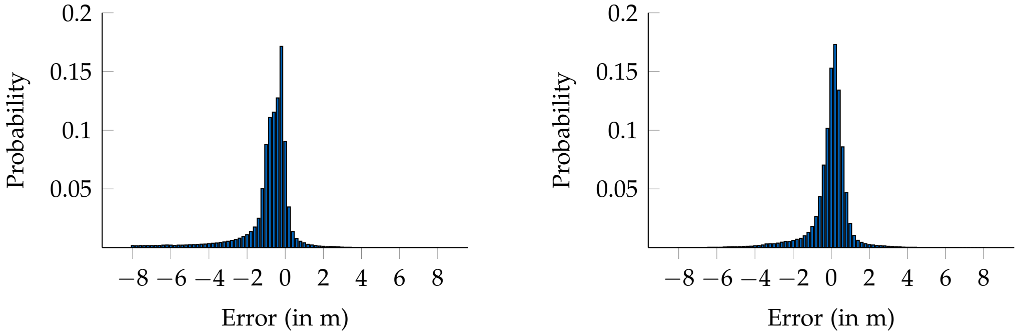
(a) Aerial DSM

(b) UAV DSM

(c) DSM differences

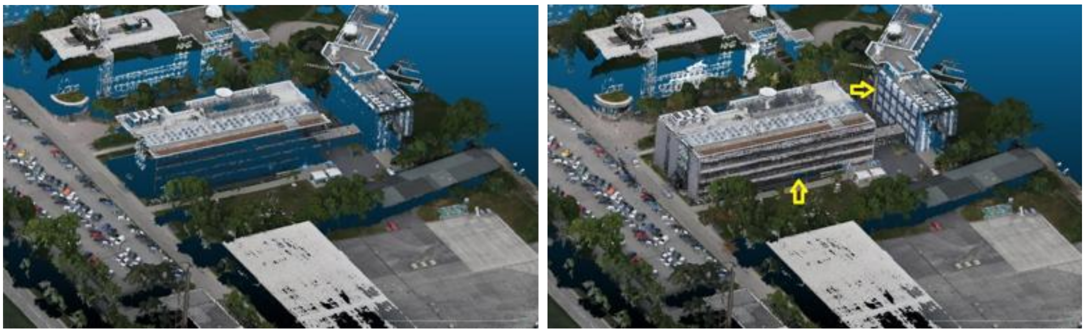
Figure A.16: Comparison of (a) aerial and (b) UAV DSM of `Germering` dataset. 20 cm GSD for aerial and 2 cm GSD for UAV DSM; (c) colormap illustrating the height differences between the two DSMs in meters

Figures A.15 and A.16 illustrate the aerial DSMs with 20 cm resolution and UAV DSMs with 2 cm resolution of `Eichenau` and `Germering` dataset respectively. The aerial DSM in (a) has blurred edge and inadequate details while the UAV DSM in (b) represent more refined details and sharper edges. Then the UAV DSM was resampled by bilinear interpolation to the same resolution of the aerial DSM and their height differences were calculated. (c) illustrates the colorized height differences ranging from -5 m to 5 m, and it is apparent that the errors are mostly smaller than 1 m. Note that the two red and one blue spots on the container site in Figure A.16(c) indicate movements of the containers due to different acquisition times of the captured images. In this sense, our matching method is able to cope with such temporal changes in scene. Figure A.17 shows the histograms of the height differences for both datasets.



(a) `Eichenau` (b) `Germering`

Figure A.17: Histograms of the height differences between the aligned DSMs generated from UAV and aerial images



(a) Aerial point cloud (b) Merged point cloud

Figure A.18: Comparison of the dense point clouds for (a) only aerial images and (b) additional registered nadir and oblique UAV images of the `EOC` dataset. The combination of aerial and UAV images can enrich 3D models for more details and add façades to buildings

A.5.4 *Application Scenario: Enriching 3D Building Models*

The `EOC` dataset represents an urban scene, demonstrating the benefits of a joint use of aerial and UAV imagery. Figure A.18a displays a dense georeferenced 3D point cloud generated solely from aerial images. Since the aerial images only contain nadir views of the scene, the reconstructed building façades are not complete, which is a typical problem for aerial photogrammetry.

We automatically geo-registered a sequence of nadir-view UAV images (see result for one image pair in Table A.8) to the aerial images. In addition, we also registered oblique UAV images facing the façades of the building to the already geo-registered UAV nadir views in a conventional photogrammetric way. Afterwards, a dense 3D point cloud was generated using all of the geo-registered UAV images, resulting in a complete reconstruction of the building with a much higher GSD than the aerial point cloud. The accurate geo-registration of the UAV images enables us to merge the UAV and aerial point cloud and leads to a comprehensive representation of the scene, as illustrated in Figure A.18b. It can be seen that the UAV point cloud is precisely aligned with the aerial point cloud. While the aerial point cloud covers a large area of the scene, the UAV point cloud contributes to information of the building façades (particularly at positions indicated by yellow arrows) and enriched details of the reconstructed building.

A.6 DISCUSSION

Our method achieves robust and accurate co-registration of images acquired from different acquisition platforms, thus opening up the possibility to integrate the information from multi-source images and achieve a more comprehensive understanding of the scene. Besides, repetitive image acquisition with manned aircrafts or satellites is quite expensive whereas it is convenient to perform with UAVs. The robust registration enables timely update of pre-existing remote sensing data using UAVs, which can also be applied in environment monitoring and change detection.

The main limitation of our method is that it only works for nadir or slightly tilted images. When a conspicuous height jump exists, the histogram may present multiple peaks, e.g., one representing matches on the ground-level and one matches on a higher level (like roofs). Therefore manual inspection is needed in this case. Moreover, it is difficult to determine the translation threshold R if the scene depth changes continuously in the image. As listed in the first column of Table A.8, there were remarkable fewer raw matches in the `Highway` scenario than in the other ones due to topographic changes. Also, those scenarios containing various scene depths (e.g. `Container` and `Eichenau`) resulted in wrong tilts when estimating the homography, leading to higher mean transfer errors (up to 7 pixels) compared to the scenarios with flat landscape.

A.7 CONCLUSION

This paper investigates into UAV geo-registration by matching UAV images with already georeferenced aerial imagery. On the basis of an extensive analysis why SIFT performs poorly for this kind of image pairs, a robust image matching approach is proposed to deliver a large number of reliable matching correspondences between the UAV and a reference image. The method is comprised of a novel feature detector, a one-to-many matching strategy and a global geometric constraint for outliers detection. The prerequisite of our proposed method is the availability of rough GNSS/IMU data of the UAV images to eliminate scale differences in the images and if possible to pre-align the images with the respect to the image rotation, although an extension of the method can handle unknown or imprecise image rotations.

Experimental results prove that our method outperforms SIFT/ASIFT in the aspects of quantity and accuracy of the detected matches. These matches are used to align UAV image blocks towards the reference images in a bundle block adjustment, which achieves a registration accuracy of $1-3$ GSD. A global accuracy evaluation of 3D points from geo-registered UAV images and terrestrial measurements from RTK GNSS show 0.5 m horizontal 1.5 m vertical deviations, which mainly stem from inaccurate georeferencing accuracy of the reference image.

REFERENCES

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., and Süsstrunk, S. (2012). "SLIC superpixels compared to state-of-the-art superpixel methods." IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 34(**11**), pp. 2274–2282.

Aicardi, I., Nex, F., Gerke, M., and Lingua, A. M. (2016). "An image-based approach for the co-registration of multi-temporal UAV image datasets." Remote Sensing 8(**9**), p. 779.

Alcantarilla, P. F., Bartoli, A., and Davison, A. J. (2012). "KAZE features." In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 214–227.

Apollonio, F., Ballabeni, A., Gaiani, M., and Remondino, F. (2014). "Evaluation of feature-based methods for automated network orientation." International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS) XL(**5**), pp. 47–54.

Bay, H., Ess, A., Tuytelaars, T., and Van Gool, L. (2008). "Speeded-up robust features (SURF)." Computer Vision and Image Understanding (CVIU) 110(**3**), pp. 346–359.

Bekele, D., Teutsch, M., and Schuchert, T. (2013). "Evaluation of binary keypoint descriptors." In: *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pp. 3652–3656.

Calonder, M., Lepetit, V., Strecha, C., and Fua, P. (2010). "Brief: Binary robust independent elementary features." In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 778–792.

Calonder, M., Lepetit, V., Ozuysal, M., Trzcinski, T., Strecha, C., and Fua, P. (2012). "BRIEF: computing a local binary descriptor very fast." IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 34(**7**), pp. 1281–1298.

Chiabrando, F., Lingua, A., and Piras, M. (2013). "Direct photogrammetry using UAV: tests and first results." International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS) XL(**1**), pp. 81–86.

Colomina, I. and Molina, P. (2014). "Unmanned aerial systems for photogrammetry and remote sensing: a review." ISPRS Journal of Photogrammetry and Remote Sensing 92, pp. 79–97.

Conte, G. and Doherty, P. (2009). "Vision-based unmanned aerial vehicle navigation using geo-referenced information." Journal on Advances in Signal Processing (EURASIP) 2009(**1**), 10:1–10:18.

Dwarakanath, D., Eichhorn, A., Halvorsen, P., and Griwodz, C. (2012). "Evaluating performance of feature extraction methods for practical 3D imaging systems." In: *Proceedings of the ACM Conference on Image and Vision Computing (CIVC)*, pp. 250–255.

Eisenbeiß, H. (2009). "UAV photogrammetry." PhD thesis. Zurich, Switzerland: Institute of Geodesy and Photogrammetry, ETH Zurich.

Fan, B., Du, Y., Zhu, L., and Tang, Y. (2010). "The registration of UAV down-looking aerial images to satellite images with image entropy and edges." In: *Proceedings of the International Conference on Intelligent Robotics and Applications (ICIRA)*. Springer, pp. 609–617.

Gerke, M. and Przybilla, H.-J. (2016). "Accuracy analysis of photogrammetric UAV image blocks: influence of onboard RTK-GNSS and cross flight patterns." PFG: Journal of Photogrammetry, Remote Sensing and Geoinformation Science 2016(**1**), pp. 17–30.

Heinly, J., Dunn, E., and Frahm, J.-M. (2012). "Comparative evaluation of binary features." In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 759–773.

Hirschmuller, H. (2008). "Stereo processing by semiglobal matching and mutual information." IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 30(**2**), pp. 328–341.

Jacobsen, K., Cramer, M., Ladstädter, R., Ressl, C., and Spreckels, V. (May 2010). "DGPF-Project: evaluation of digital photogrammetric camera systems geometric performance." PFG: Journal of Photogrammetry, Remote Sensing and Geoinformation Science 2010(**2**), pp. 83–97.

Juan, L. and Gwun, O. (2009). "A comparison of SIFT, PCA-SIFT and SURF." International Journal of Image Processing (IJIP) 3(**4**), pp. 143–152.

Koch, T., Zhuo, X., Reinartz, P., and Fraundorfer, F. (2016a). "A new paradigm for matching UAV-and aerial images." International Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS) III(**3**), pp. 83–90.

Koch, T., d'Angelo, P., Kurz, F., Fraundorfer, F., Reinartz, P., and Körner, M. (2016b). "The TUM-DLR multimodal earth observation evaluation benchmark." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR-WS)*, pp. 698–705.

Kurz, F, Meynberg, O, Rosenbaum, D, Türmer, S, Reinartz, P, and Schroeder, M (2012). "Low-cost optical camera system for disaster monitoring." International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS) XXXIX(**8**), pp. 33–37.

Kurz, F., Rosenbaum, D., Meynberg, O., Mattyus, G., and Reinartz, P. (2014). "Performance of a real-time sensor and processing system on a helicopter." International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS) XL-1, pp. 189–193.

Lin, T.-Y., Cui, Y., Belongie, S., and Hays, J. (2015). "Learning deep representations for ground-to-aerial geolocalization." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5007–5015.

Lowe, D. G. (2004). "Distinctive image features from scale-invariant keypoints." International Journal of Computer Vision (IJCV) 60(**2**), pp. 91–110.

Majdik, A. L., Verda, D., Albers-Schoenberg, Y., and Scaramuzza, D. (2015). "Air-ground matching: appearance-based GPS-denied urban localization of micro aerial vehicles." Journal of Field Robotics 32(**7**), pp. 1015–1039.

Nex, F. and Remondino, F. (2014). "UAV for 3D mapping applications: a review." Applied Geomatics 6(**1**), pp. 1–15.

Rabus, B., Eineder, M., Roth, A., and Bamler, R. (2003). "The shuttle radar topography mission-a new class of digital elevation models acquired by spaceborne radar." ISPRS Journal of Photogrammetry and Remote Sensing 57(**4**), pp. 241 –262.

Rosten, E., Porter, R., and Drummond, T. (2010). "Faster and better: A machine learning approach to corner detection." IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 32(**1**), pp. 105–119.

Rothermel, M., Wenzel, K., Fritsch, D., and Haala, N. (2012). "SURE: Photogrammetric surface reconstruction from imagery." In: *Proceedings of the Low Cost 3D Workshop (LC3D)*. Vol. 8.

Rublee, E., Rabaud, V., Konolige, K., and Bradski, G. (2011). "ORB: An efficient alternative to SIFT or SURF." In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2564–2571.

Shan, Q., Wu, C., Curless, B., Furukawa, Y., Hernandez, C., and Seitz, S. M. (2014). "Accurate geo-registration by ground-to-aerial image matching." In: *Proceedings of the IEEE International Conference on 3D Vision (3DV)*, pp. 525–532.

Verhoeven, G., Wieser, M., Briese, C., and Doneus, M. (2013). "Positioning in time and space: cost-effective exterior orientation for airborne archaeological photographs." International Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS) II(**5**), pp. 313–318.

Vexcel. *Vexcel UltraCam*. `http://www.vexcel-imaging.com/`. Accessed: 2017-02-10.

– *Vexcel UltraNav*. `http://www.vexcel-imaging.com/wp-content/uploads/2016/09/Brochure_UltraNav.pdf`. Accessed: 2017-02-10.

Xu, Y., Ou, J., He, H., Zhang, X., and Mills, J. (2016). "Mosaicking of Unmanned Aerial Vehicle Imagery in the Absence of Camera Poses." Remote Sensing 8(**3**), p. 204.

Yu, G. and Morel, J.-M. (2011). "ASIFT: an algorithm for fully affine invariant comparison." Image Processing On Line 1, pp. 11–38.

Zhao, H., Zhang, B., Wu, C., Zuo, Z., Chen, Z., and Bi, J. (2014). "Direct georeferencing of oblique and vertical imagery in different coordinate systems." ISPRS Journal of Photogrammetry and Remote Sensing 95, pp. 122 –133.

Zhuo, X, Cui, S, Kurz, F, and Reinartz, P (2016). "Fusion and classification of aerial images from MAVS and airplanes for local information enrichment." In: *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 3567–3570.

Zitová, B. and Flusser, J. (2003). "Image registration methods: a survey." Image and Vision Computing 21(**11**), pp. 977 –1000.

d'Angelo, P. and Reinartz, P. (2011). "Semiglobal matching results on the ISPRS stereo matching benchmark." International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS) XXXVII(**4**), pp. 79–84.

B

KOCH, T., KÖRNER M., FRAUNDORFER F. (2019) AUTOMATIC AND SEMANTICALLY-AWARE 3D UAV FLIGHT PLANNING FOR IMAGE-BASED 3D RECONSTRUCTION. REMOTE SENSING 11 (13), 1550.

This chapter represents a pre-print version of the published article with identical content. The original article appeared under `doi:10.3390/rs11131550`.

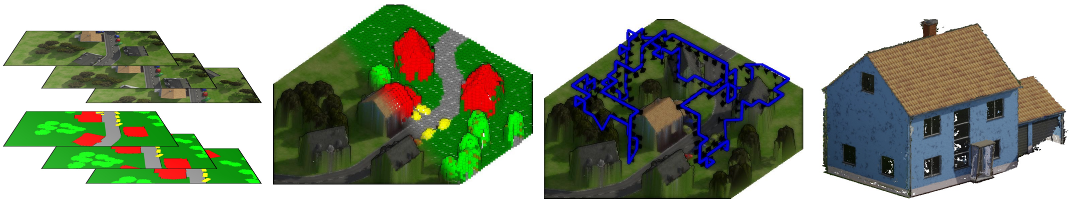
B.1 INTRODUCTION

Unmanned aerial vehicles (UAVs) have attracted significant attention in the field of 3D modeling, as they are capable of carrying high-resolution cameras combining advantages of both conventional airborne and terrestrial photogrammetry. The mobility and maneuverability of UAVs to freely move in three dimensions and simultaneously capture close-up images of an object with arbitrary viewing angles allow to generate high-resolution and photo-realistic 3D models with high accuracy by processing a series of overlapping images with current state-of-the-art *Structure from Motion (SfM)* and *Multi-View Stereo (MVS)* pipelines, such as Pix4D (Pix4Da), Bundler (Snavely et al., 2006), or Colmap (Schönberger and Frahm, 2016). These models are of high interest in various fields, such as the use of digitized building models for 3D city modeling (Vacanas et al., 2015), object inspection (Hallermann and Morgenthal, 2014), or cultural heritage documentation (Mostegel et al., 2017). However, the quality of resulting 3D models strongly relies on flight plans that satisfy the requirements of an image-based 3D modeling process which include the acquisition of multiple overlapping images, sufficient baselines between the camera viewpoints and the prevention of optical occlusions from surrounding obstacles. In terms of mapping mostly flat and spacious scenes, such as landscapes, flight planning can be easily executed in form of simple grid-like patterns or circular flights from the same altitude but can become exceedingly complex for densely built urban areas consisting of different kinds of human-made objects and vegetation. Planning a UAV trajectory in such areas involves considering the surrounding environment and keeping a safety distance toward any obstacle while ensuring that the entire object of interest is captured from close ranges and different perspectives.

The most common method to obtain aerial imagery in an automated fashion is to use an off-the-shelf flight planner, such as commercial flight planning software Pix4D Capture (Pix4Db), PrecisionHawk Precision Flight (Precisionhawk), DJI Flight Planner (DJI), or open-source based PixHawk ArduPilot (ArduPilot). These easy-to-use planners can generate simple polygons, regular grids, or circular trajectories, however, some prior knowledge of the scene height must be known in advance for designing a collision-free flight plan. For more complex scenes, such as urban areas, standard path planning methods are usually insufficient to generate high-quality 3D models, as we will show later. Therefore, UAV flights in such complex scenarios still require manual operation by experienced pilots in case standard flight planners are not feasible or do not guarantee a sufficient reconstruction quality. From a practical or even legal point of view, it may be even necessary to adapt the flight plan with respect to the semantics of the environment, especially in densely built areas. Restricted airspaces may be defined in regions that are prohibited to be accessed by the UAV or which should be avoided in case of an unexpected malfunction of the vehicle. These restricted areas could include other buildings, train rails, water bodies, parked cars, highways or other heavily frequented roads. Flying UAV in such environments is already challenging. If the resulting 3D model additionally demands certain photogrammetric properties, such as the desired *ground sampling distance (GSD)*, the acquisition of highly overlapping close-up images covering the entire object could become infeasible in the presence of restricted or prohibited airspaces.

General research on path planning for UAV mapping has already been initiated in recent years focusing on automation of the generation of optimal flight plans. Automated flight planning methods can be classified either as model-free and model-based methods. The former performs an *exploration* task in unknown environments by iteratively updating the model with new measurements via selecting the next best view from a current view. These models do not require prior knowledge of the scene but usually, they do not guarantee full coverage of the object. Methods of the latter class, on the other hand, rely on a coarse proxy model of the scene and refine the model by an optimal subsequent flight which is globally optimized. The targets of these *explore-and-exploit* approaches are manifold, such as maximizing the coverage of a target object (Hepp et al., 2018b; Roberts et al., 2017) or minimizing the acquisition time (Cheng et al., 2008) or energy consumption (Chakrabarty and Langelaan, 2009; Di Franco and Buttazzo, 2016). However, to the best of our knowledge, none of these works take into account the surrounding environment for generating safe UAV paths that additionally avoid or even restrict certain airspaces in the scene. With the tremendous advances in semantic image segmentation for aerial imagery by recent deep learning-based approaches (Zhu et al., 2017), accurate and consistent dense semantic maps can be generated, which extend the purely geometric 3D scene representation, helping to generate safe UAV flights under the consideration of the real environment. Since we want to adapt the flight path to the semantic properties of the scenery, an initial semantically-enriched proxy model of the entire scene is required, which leads us to employ a model-based approach. An inspiration of our path planning method was given by the works of Roberts et al. (2017) and Hepp et al. (2018), formulating the path planning problem as a graph-based optimization for maximizing the information gain obtained from a UAV trajectory with a set of heuristics representing 3D modeling image acquisition practices. With this paper, we build on these works by introducing more interpretable heuristics which directly influence user-specified requirements of the reconstruction quality, as well as optimizing for a minimum path length. In addition, we show how to incorporate semantic information to safe path planning. Figure B.1 illustrates the general idea of our path planning approach, consisting of a two-staged planning procedure, wherein a first nadir flight is used to generate a semantically-enriched proxy model of the entire environment which is further used to generate a set of viewpoint hypotheses in the free and accessible airspace. A discrete optimization among this camera graph is conducted to find a short and matchable path along the graph, which maximizes the reconstruction quality of the target object while considering restrictions on the airspace defined by the semantical cues.

Particularly, our contributions are as follows:



(a) UAV images and semantic maps

(b) Semantically-enriched initial model

(c) Semantic-aware 3D UAV path

(d) Final reconstructed 3D model

Figure B.1: Proposal of our UAV path planning methodology for generating 3D reconstructions of individual objects considering path restrictions based on semantics. A 3D proxy model is generated using a set of geo-referenced UAV images (a) from a simple flight above the environment which is further enriched by transferring 2D segmentation labels into 3D space, defining free, conditionally accessible and prohibited airspaces (b). A graph-based optimization estimates a collision-free trajectory for image acquisition viewpoints considering restricted airspaces while minimizing the respective path length (c). The acquired images of the traversed trajectory are suitable to generate high-resolution 3D reconstruction models (d)

(1) We propose a set of heuristics based on photogrammetric reconstruction parameters, leading to individual flight paths for arbitrary camera intrinsics that ensure the generation of 3D models in a user-specified resolution.

(2) We show how to exploit semantic segmentation of UAV imagery for extracting the target object and for generating a semantically-enriched initial 3D proxy model, which defines restricted and prohibited airspaces.

(3) We propose a model-based optimization scheme with respect to a semantic model that maximizes the object coverage while minimizing the corresponding path length and avoiding restricted airspaces.

(4) We propose a realistic synthetic 3D model suitable for a comprehensive evaluation of urban flight planning, including a highly detailed building model embedded in a realistic and interchangeable scenery.

B.2 RELATED WORK

The rapid development of UAVs and sensors has contributed significantly to their popularity in many industries nowadays, such as urban mapping, object inspection, precision agriculture, and surveying tasks. Equipped with high-resolution cameras and the utilization of most recent SfM and MVS methods on image sequences, 3D models of the environment can be generated in a much greater level of detail compared to conventional manned aircraft. However, the quality of such reconstructions highly depends on the camera network configuration during the acquisition process. An exhaustive amount of work addressed the problem of selecting the best views from a large amount of different views hypotheses (Furukawa and Hernández, 2015; Furukawa et al., 2010; Goesele et al., 2007; Rumpler et al., 2011; Snavely et al., 2006). These works point out the crucial parameters which affect the reconstruction quality, such as parallax angles and baselines between views, as well as their observation angles and distances toward the object's surface and propose meaningful heuristics to model the reconstruction quality from different camera constellations.

An integration of these parameters is already used for automating the image acquisition process for large-scale areas (ArduPilot; DJI; Pix4Da; Precisionhawk), allowing to plan UAV flights as simple geometric patterns, such as regular grids or circular flights with respect to the desired GSD. These off-the-shelf planners are sufficient in case of spacious and flat terrains without obstacles (Nex and Remondino, 2014), but are not suitable for use in uneven, densely built or heavily vegetated environments. Since no 3D model of the environment is taken into consideration, these trajectories either do not cover every part of the object of interest due to occlusions by obstacles or may even cause an accident with an adjacent obstacle in the environment.

More advanced path planning approaches aim to automatically map objects in either completely unknown environments or based on a very coarse prior model of the environment. Methods of the first group solve an *exploration* task by iteratively selecting the most promising view to refine the explored model based on a current view with new measurements. This incremental scene modeling and viewpoint planning is commonly known as *next best view (NBV)* planning, which is already a long-standing part of research in the field of Robotics. The methods alternately fuse incoming measurements from a new viewpoint into the reconstruction of the scene and estimate novel viewpoints in order to increase the information about the object. Classical sensors for these measurements include laser scanners (Kriegel et al., 2015), RGB-D sensors (Fan et al., 2016; Heng et al., 2011; Hepp et al., 2018a; Loianno et al., 2015; Meng et al., 2017; Michael et al., 2012; Sturm et al., 2013) and cameras (Border et al., 2018; Dunn and Frahm, 2009; Kumar Ramakrishnan and Grauman, 2018; Mendez et al., 2017; Palazzolo and Stachniss, 2018; Stumberg et al., 2016). Such methods are usually hard to implement utilizing cameras as selected sensors, as the generation of depth maps, which is necessary to derive new 3D information, requires significant onboard processing power or at least a wireless connection to the ground-station for data transmission, in order to merge incoming measurements with the current model. Additionally, selecting next best views in accordance to MVS requirements—in particular, maintaining sufficient baselines and parallax angles of adjacent views—on the fly is a challenging task since the actual mapped free airspace might be very limited.

In contrast to model-free *exploration* methods that focus on autonomy and real-time capability in unknown environments, model-based path planning algorithms rely on an available proxy model of the environment and focus on estimating a subsequent optimal path to maximize the coverage and accuracy of the object globally (Hepp et al., 2018b; Hoppe et al., 2012; Jing et al., 2016; Peng and Isler, 2019; Roberts et al., 2017; Smith et al., 2018). In contrary to active modeling, these *explore-and-exploit* methods do not receive any feedback from the acquired images during the *exploitation* flight, which demands high attention to the applied heuristics being used for generating the refinement path. The global optimization of coverage and accuracy, on the other hand, usually leads to larger completeness and smoother trajectories compared to model-free methods. Recent work has proposed to extend this procedure by iteratively refining the model from several subsequent flights, taking into account the remaining model uncertainty between each flight (Huang et al., 2018; Peng and Isler, 2019). Furthermore, the execution of the optimized path is easy and fast for any kind of UAV by simply navigating alongside the optimized waypoints. The prior model can either be based on an existing map with height information (Jing et al., 2016) or is generated by photogrammetric reconstructions from a preceding manual flight at a safe altitude or via standard flight planning methods (*e.g.*, regular grids or circular trajectories) (Hepp et al., 2018b; Roberts et al., 2017) and is usually expressed by a set of discrete 3D points in a voxel space (Alsadik et al., 2013; Hepp et al., 2018b; Roberts et al., 2017; Smith et al., 2018) or by volumetric surfaces, such as triangulated meshes (Bircher et al., 2016; Hoppe et al., 2012; Jing et al., 2016; Peng and Isler, 2019). In order to define appropriate views for the optimized trajectory, camera viewpoint hypotheses are either regularly sampled in the free 3D airspace (Roberts et al., 2017; Smith et al., 2018) resulting in 3D camera graphs, or are sparsely sampled in a 2D view manifold (Peng and Isler, 2019) or in skeleton sets (Snavely et al., 2008) around the object. Subsequently, an optimization is defined in order to find a connected subset of these viewpoint hypotheses to define a suitable path through the camera graph. Alternatively, the locations of the of regularly sampled viewpoint candidates can be continuously refined during the optimization (Hepp et al., 2018b). As a means of assessing the suitability of camera viewpoints for the reconstruction, hand-crafted heuristics are usually defined considering the necessities for a successful SfM and MVS workflow. These include multi-view requirements (Alsadik et al., 2013; Hoppe et al., 2012; Smith et al., 2018), ground resolution (Bircher et al., 2016; Hoppe et al., 2012), 3D uncertainty (Mostegel et al., 2016) and the coverage of the object (Hepp et al., 2018b; Roberts et al., 2017; Smith et al., 2018). Instead of using hand-crafted heuristics, several works used machine learning methods to learn heuristics that allow predicting the confidence in the output of a MVS without executing it (Devrim Kaba et al., 2017; Hepp et al., 2018a; Mostegel et al., 2016).

Recently, efficient methodologies formulate the view planning problem as a discrete optimization task and exploit submodularity in the optimization process, standing for fast and reliable convergence, even for a large number of viewpoint hypotheses (Hepp et al., 2018b; Roberts et al., 2017). The main advantage of this idea is to jointly assess additional information gain of individual viewpoints for arbitrary viewpoint constellations in a global manner. This allows formulating the path planning task as an orienteering problem, which can be solved with simple greedy algorithms, by optimizing a path which collects as many information gains as possible for a specific path length. The results presented in previous work reveal notable trajectories for generating high-fidelity image-based 3D reconstructions. However, setting a suitable path length in the optimization may require expert knowledge and highly affects the trajectory estimation, since, due to the purely additive nature of orienteering problem, adding additional views will never decrease the objective function. This might lead to abundant redundant views for overestimated path lengths and incomplete reconstructions for underestimated path lengths. Although the presented heuristics follow best practices for MVS requirements, they do not respect user-specific demands on the resulting 3D model, such as the number of views and observations angles of the object surface or a required model resolution using arbitrary cameras. Additionally, prior work so far solely considers purely geometric cues for flight planning of both small-scale and large-scale areas. With the vast progress in semantic segmentation using deep learning-based approaches, the applicability of neural networks for semantic segmentation of aerial and UAV imagery was demonstrated in several works (Chen et al., 2018; Kaiser et al., 2017; Marmanis et al., 2016).

In this paper, we show how to incorporate semantic cues into UAV flight planning for generating safe trajectories for real-world 3D mapping applications, which allow to define inadmissible airspaces above user-defined object types. Additionally, we

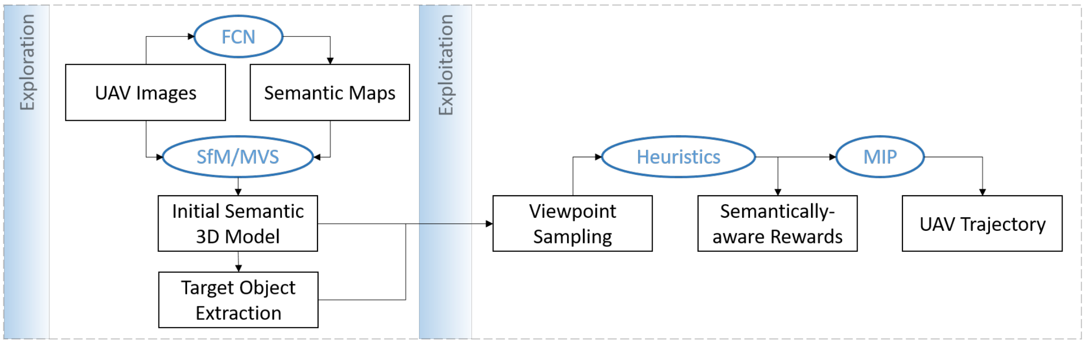


Figure B.2: Overview of the proposed workflow of our UAV path planning approach. Based on a exploration flight at a safe altitude, the captured images are segmented and fused to an initial semantic 3D model. After selection of the target object, numerous camera viewpoints are sampled and assessed according to their eligibility for the reconstruction process, while the semantic information of the environment assigns restricted or prohibited airspaces for the UAV. Finally, a discrete graph-based optimization estimates the optimal semantically-aware trajectory which ensures a high-quality 3D model of the target object.

propose a set of heuristics for SfM and MVS image acquisition used in the optimization allowing for the maintenance of a pre-defined model resolution for the entire targeted object. Although the preferred task of photogrammetric 3D modeling is to maximize the reconstruction quality rather than minimizing the path length, we integrate a penalization for lengthy paths without a significant drop in the reconstruction quality.

B.3 PROPOSED FLIGHT PLANNING PIPELINE

Our flight planning methodology follows a two-staged *explore-and-exploit* approach, consisting of two subsequent flights, where a first safe *exploration* flight is used to generate an initial proxy model of the environment which is further refined by an optimized *exploitation* path in terms of full coverage, high-resolution and accuracy of the object to be reconstructed. Latter additionally respects restrictions of the airspace derived from semantic cues to avoid hazardousness and prohibited areas and to elude collisions with the surrounding environment. Our work is inspired by the works of Roberts et al. (2017) and Hepp et al. (2018) in the matter of estimating a closed trajectory from numerous viewpoint hypotheses by exploiting submodularity in the optimization procedure. An overview of our complete workflow is depicted in Figure B.2. First, the acquired images of the exploration flight are processed to generate a semantically-enriched coarse proxy model which defines free and occupied airspace. The semantic cues help to extract the object of interest and, based on the proxy model, a set of viewpoint hypotheses is generated and evaluated according to their eligibility for reconstructing the target object with respect to our heuristics used for MVS image acquisition. Adjacent viewpoints are evaluated according to their matchability and connected to a camera graph. Finally, an *exploitation* flight is optimized by finding a closed and short path among the camera graph which maximizes the reconstruction quality and avoids prohibited and minimizes hazardousness airspaces defined by the semantics of the proxy model. Summarizing the objectives of the path planning problem, the following requirements need to be fulfilled by our methodology:

1. *Coverage*: every point on the object surface has to be visible in at least two images to be able to triangulate its position in 3D space from the images.

2. *Safety*: the estimated trajectory has to avoid collisions with obstacles and has to be aware of the semantics of the surrounding environment in terms of restricted and prohibited airspaces.

3. *Path length*: the estimated trajectory should be as short as possible and avoid redundant views, as several images taken from similar camera poses introduce local uncertainties in depth estimation by glancing intersections.

4. *Heuristics*: The estimated trajectory should facilitate complete reconstruction of the target object considering photogrammetric reconstruction criteria, such as GSD, observation angles, number of views, and sufficient overlap between adjacent views.

5. *Quality assessment*: the path planning method should return an approximation of the expected reconstruction quality before the execution of the flight, in order to adjust the path or plan another subsequent path.

The following sections provide a detailed description of the proposed methodology, starting with the outline of the path planning problem and the definition of the optimization objective in Section B.3.1. Details on the generation of the semantically-enriched proxy model from a set of nadir images and the extraction of the target object from the proxy model are provided in Section B.3.2. Section B.3.3 describes the generation of numerous viewpoint hypotheses, which are assessed with respect to our proposed heuristics explained in Section B.3.4. Finally, Section B.3.5 presents the semantic-aware optimization.

B.3.1 *Notation and Definition of the Path Planning Problem*

The objective of our path planning problem is to find a feasible UAV trajectory to acquire images of a target object such that the final 3D reconstruction model is of high quality. The object of interest is expressed as a sparse set of discrete surface points $\boldsymbol{s}_{j=1\ldots J} = \left(\boldsymbol{x}_j, \boldsymbol{\eta}_j\right) \in \mathcal{S}$, comprised of 3D locations $\boldsymbol{x}_j \in \mathbb{R}^3$ and normal vectors $\boldsymbol{\eta}_j \in \mathbb{R}^3$ on the tangent plane of the object. We consider a discrete optimization scheme and represent our camera viewpoint hypotheses as an undirected weighted graph $G = (\mathcal{P}, \mathcal{E})$, composed of a set $\mathcal{P}$ of nodes as camera poses $\boldsymbol{p}_{i\ldots I} = (\boldsymbol{c}_i, \boldsymbol{r}_i) \in \mathcal{P}$ consisting of 3D locations $\boldsymbol{c}_i \in \mathbb{R}^3$ and camera orientations $\boldsymbol{r}_i \in \mathbb{R}^3$ defined as roll, pitch and yaw angles. Adjacent and matchable viewpoints in the graph are connected through a set of edges $\mathcal{E} = \left\{\boldsymbol{e}_k = \left(\boldsymbol{p}_i, \boldsymbol{p}_j\right)\right\}$ with associated weights $\mathcal{W} = \left\{\boldsymbol{w}_k = \left(w_k^{\text{eucl}}, w_k^{\text{sem}}\right)\right\}$, representing a Euclidean distance $w_k^{\text{eucl}} \in \mathbb{R}$ and a semantic label cost $w_k^{\text{sem}} \in \mathbb{R}$. We define a feasible trajectory $\mathcal{T} = \{\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_n\} \subset \mathcal{P}$

as a subset of connected camera poses in the camera graph G. The goal of the path planning problem is to find an optimal trajectory

$$\begin{aligned} \mathcal{T}^* &= \arg\max_{\mathcal{T}} R(\mathcal{T}) \\ \text{subject to } &\sum_{e \in \mathcal{E}} w^{\text{eucl}} \rightarrow \min, \\ &\sum_{e \in \mathcal{E}} w^{\text{sem}} < L^{\text{sem}} \end{aligned} \tag{B.1}$$

that maximizes the reconstructability $R : \mathcal{P} \rightarrow \mathbb{R}$ of the target object $\mathcal{S}$, while minimizing the corresponding path length and restricting the path not to exceed an accumulated label cost limit L^{sem}. The reconstructability $R(\mathcal{T}) = \sum_{\mathcal{T}} I(\boldsymbol{p}(\mathcal{T}), \mathcal{S})$ obtained from a trajectory $\mathcal{T}$ is defined as the accumulated information reward $I(\boldsymbol{p}(\mathcal{T}), \mathcal{S})$ of all camera poses $\boldsymbol{p}(\mathcal{T})$ of that trajectory. The computation of rewards I requires a set of heuristics, approximating the impact of an arbitrary camera pose $\boldsymbol{p}$ for the reconstruction quality of the object surface $\mathcal{S}$. Besides rating of the camera poses regarding the distance toward the object surface and the incidence angles of camera rays, the proposed heuristics also address the assessment of camera configurations of adjacent camera poses with respect to a successful multi-view stereo matching.

B.3.2 *Semantically-Enriched Initial 3D Model*

Given a series of nadir or oblique images encompassing the object of interest and its surrounding environment, a coarse proxy model of the entire scene is generated by processing the initial images with current state-of-the-art SfM and MVS pipelines, such as Pix4D (Pix4Da), Colmap (Schönberger and Frahm, 2016), or Bundler (Snavely et al., 2006). The initial flight can be realized either by a manual flight at a safe altitude or via commonly used predefined flight planning systems resulting in grid-like or circular patterns, which is feasible in most sceneries. In order to compute the subsequent trajectory in the same reference frame as the initial flight, we incorporate GNSS coordinates of the UAV or utilize *ground control point (GCP)* to the bundle adjustment. The model only requires a low resolution and can exhibit gaps in the reconstruction, such as missing façades, but should cover a large amount of the surrounding environment, which determines accessible and occupied air space for the viewpoint planning. The model itself can be either expressed as a dense point cloud with low point sampling density or by regularly sampled 3D points from the faces of a triangulated mesh. The initial proxy model generation can be computed fast even with off-the-shelf mobile computers. Alternatively, a coarse 3D model can be already generated on-board the UAV during the *exploration* flight (Wendel et al., 2012).

At the same time, a pixel-wise dense semantic segmentation of the images is conducted using a *fully convolutional network (FCN)* (Long et al., 2015). An adjustment of the number of classes required for our task (building, lawn, tree, street, car, others) and the utilization of a diverse set of available and manually annotated UAV and aerial nadir images from different altitudes and various scenes was carried out for training the network. Available training data from (Semantic Drone Dataset) and (ISPRS Potsdam) was extended with manually annotated UAV images from

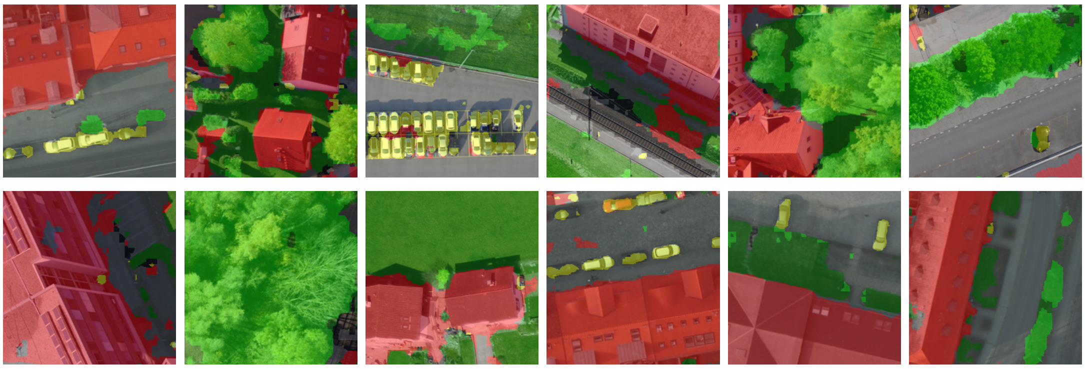


Figure B.3: Example of semantic segmentation results on our validation images with a fine-tuned FCN model (Long et al., 2015) trained on UAV and aerial images. Visualization is color coded for buildings (🟥), streets (◼), low vegetation (🟩), high vegetation (🟩), cars (🟨), and others (⬛)

different scenes to achieve a total amount of 3069 images split into 60% training and 40% validation images. The quantitative evaluation after refining the pre-trained model for 50 epochs yield a global accuracy of 0.81 and a mean Intersection over Union (IoU) score of 0.52, indicating a reasonable segmentation performance for our task. We infer every single UAV image used for the initial 3D reconstruction to the segmentation network, in order to facilitate the redundancy of overlapping areas for reducing labeling uncertainty. To propagate the 2D semantic labels, we make use of the visibility information obtained from the 3D modeling process and back-project every single 3D point into every image in which it is visible and compute the point label by majority voting. Despite the rather small receptive field of the FCN-8s providing merely coarse segmentation boundaries as shown in Figure B.3, an adequate semantic enrichment of the 3D model can be achieved for a relatively large grid spacing of adjacent viewpoints (3–4 m in our experiments) by exploiting the redundancy of overlapping images.

Since the initial 3D model is coarsely geo-referenced, it is possible to refine the segmentation results of the 3D scene for hardly distinguishable objects of the same semantic class with the use of *open street map (OSM)* information. For instance, the distinction of various types of roads, which can hardly be determined by 2D semantic segmentation methodologies, could be a crucial requirement for generating safe UAV paths. Heavily frequented road sections (*e.g.*, parking lots, highways and trunk roads) should be highly avoided, while restrictions on side roads and driveways could be less strict. The already segmented road sections of the initial 3D model can therefore be extended with subtypes by automatically inferring the classes from OSM to the labeled 3D points. Since OSM provides numerous and detailed map features, this procedure can be extended for various land cover classes and facilitates user-defined restrictions, such as the differentiation of residential and industrial buildings.

Given an approximate semantically-enriched 3D model of the environment, the target object to be finally reconstructed needs to be identified, extracted and completed in a semi-automatic manner. As the initial model could be incomplete during the reconstruction process and the usage of nadir-views results in gaps in the model, such as missing façades and other unseen object details, the target model needs to be completed to ensure camera poses pointing toward these missing details. With the assumption of simplified building models, we identify and extract the target object by a simple 3D region growing approach exploiting the semantic labels of the 3D points. A user input of one corresponding 3D point belonging to the object to be reconstructed serves as the seed for the region growing process. After isolation of the target object, we equally sample surface points $\boldsymbol{s} = (\boldsymbol{x}, \boldsymbol{\eta})$ of the object outline to the ground level and compute 3D point normals required for the proposed heuristics.

B.3.3 *Camera Viewpoint Hypotheses Generation*

The goal of the trajectory planning is to define a set of viewpoints allowing the triangulation of as many 3D points of the target object surface as possible according to photogrammetric necessities for a successful reconstruction. A large amount of evenly distributed viewpoint candidates $\boldsymbol{c}$ is sampled in the free airspace inside a bounding box around the extracted object, excluding camera viewpoints which are closer to any surrounding obstacle than a predefined safety buffer. This safety buffer can be adapted according to the corresponding semantic labels of the environment in order to increase the distance toward hazardous objects, such as trees, which often lack in completeness for photogrammetric reconstructions. For each viewpoint candidate, we also store a vector containing the semantic labels of all proxy 3D points located below the camera viewpoints.

Besides the location of camera viewpoints, orientations $\boldsymbol{r}$ need to be assigned pointing toward the target object while avoiding occlusions with obstacles. Although the subsequent reconstruction process favors fronto-parallel views toward the target surface to ensure a high-quality reconstruction, adjacent viewpoints also require smooth transitions with high overlap. Since viewpoint orientations pointing toward the closest surface point results in fronto-parallel views, the matchability at edges of the object might be insufficient due to large orientation changes. On the other hand, viewpoint orientations which always point toward the center of the object result in large overlap but slanted views toward the object surface in case of elongated or other complex object structures. In comparison to other approaches, which either assign orientations pointing toward the center of the object (Cheng et al., 2008) or include the orientation estimation in the optimization (Hepp et al., 2018b; Roberts et al., 2017), we perform a visibility assessment of each viewpoint to identify 3D surface points which are visible from each specific viewpoint location considering the surrounding environment. A fast visibility computation approach (Katz et al., 2007) is utilized and visibilities for all viewpoints and surface points are stored in an indicator matrix $\boldsymbol{U} \in \mathbb{R}^{I \times J}$. In particular, a look-at-vector $\boldsymbol{n}_i \in \mathbb{R}^3$ for each viewpoint $\boldsymbol{c}_i$ is computed and directed toward the weighted mean of all visible 3D points $\mathcal{S}_{c_i} \subset \mathcal{S}$ from the corresponding 3D location of the viewpoint. In order to prioritize object points that are closer to the camera, $\boldsymbol{n}_i$ is further weighted by the distance toward all visible 3D points. We begin with computing weighting coefficients τ_j for each visible surface point $\boldsymbol{x}_j \in \mathcal{S}_{c_i}$ from a camera view $\boldsymbol{c}_i$ utilizing the normalized distances from all visible surface points toward the camera location by

$$\tau_j = 1 - \sqrt[k]{\frac{\|\boldsymbol{x}_j - \boldsymbol{c}_i\| - \min\left(\{\|\boldsymbol{x} - \boldsymbol{c}_i\|\}\right)}{\max\left(\{\|\boldsymbol{x} - \boldsymbol{c}_i\|\}\right) - \min\left(\{\|\boldsymbol{x} - \boldsymbol{c}_i\|\}\right)}}, \tag{B.2}$$

where k controls the strength of favoring closer surface points toward the camera viewpoints. The weighting coefficients $\left\{\tau_j \in \mathbb{R} : 0 \leq \tau_j \leq 1\right\}$ reflect the influence of each surface point based on its distance to the camera viewpoint, resulting in large values for closer surface points and decreasing values for farther points. In combination with the normalized direction vectors toward each visible surface point, the weighted look-at vector $\boldsymbol{n}_i$ is computed by

$$\boldsymbol{n}_i = \frac{1}{\sum_j \tau_j} \sum_{x_j \in \mathcal{S}_{c_i}} \tau_j \cdot \frac{\boldsymbol{x}_j - \boldsymbol{c}_i}{\|\boldsymbol{x}_j - \boldsymbol{c}_i\|}. \tag{B.3}$$

We found that this simple procedure results in suitable viewpoint orientations applicable for different object outlines, avoids occluded views and results in almost fronto-parallel views with smooth transitions at object boundaries allowing a large image overlap needed for a successful image registration. Additionally, the orientation estimation does not have to be included in the optimization, which would increase the complexity of the optimization. Finally, the look-at vectors $\boldsymbol{n}_i$ are converted into pose orientations $\boldsymbol{r}_i$, composed of three Euler angles $\varphi_i = 0$, $\theta_i = \sin^{-1}(-n_{i,y})$ and $\psi_i = \tan^{-1}(\frac{n_{i,x}}{n_{i,z}})$ representing roll, pitch and yaw angles, whereas roll angles are fixed to zero, as we assume axis aligned camera views. After updating the visibility matrix $\boldsymbol{U}$ with respect to the camera intrinsics and assigned orientations, theoretical overlaps between views are computed. Nodes of adjacent camera viewpoints which satisfy a specific overlap constraint (*e.g.*, 75%), are connected via edges $\boldsymbol{e}$ in the graph G, comprised of the Euclidean distance w^{eucl} between the corresponding nodes and the semantic label costs w^{sem}, defined as the mean distribution of assigned labels of ground points between both nodes.

B.3.4 *Path Planning Heuristics*

In terms of the optimization defined in Equation B.1, the abundant viewpoint hypotheses $\boldsymbol{c}_i$ have to be assessed with respect to their eligibility for reconstructing the object. Following best practices on image acquisition for photogrammetric 3D reconstruction, a set of heuristics is defined which reflect the requirements of the subsequent steps of image registration and dense matching. There is an extensive amount of relevant literature on the principles of photogrammetric 3D modeling (Förstner and Wrobel, 2016; Hartley and Zisserman, 2003; Luhmann et al., 2013) pointing out decisive aspects for achieving high-quality reconstructions from a set of images:

1) *Distance*: the distance between camera viewpoints and object surface defines the resulting model resolution and depends on the desired point density and the camera intrinsics.

2) *Observation angle*: shallow observation angles between the camera views and surface normals are favored in MVS approaches.

3) *Multiple views*: every part of the scene has to be observed from at least two views from different perspectives with sufficient overlap between the views. The identification of corresponding points in overlapping images is the requirement for robustly estimating camera poses and for triangulating 3D object points.

4) *Parallax angle*: shallow parallax angles increase the triangulation error and therefore affect the model quality, while too large angles decrease the matchability between the views due to a lack of image similarity between the views

which could result in a failure of the image registration step or in gaps in the reconstructed 3D model.

The heuristics are used to predict the eligibility of the viewpoints for the reconstruction and ensures that the target object can be sufficiently reconstructed using the estimated viewpoints from the trajectory. Requirements 1) and 2) can be formulated independently for all viewpoints, while 3) and 4) depend on a pairwise or even multi-view assessment. We define information rewards

$$I(\boldsymbol{p}_i, \mathcal{S}_{\boldsymbol{p}_i}) = \sum_{\boldsymbol{x}_j \in \mathcal{S}_{\boldsymbol{p}_i}} I_{\mathrm{d}}(\boldsymbol{c}_i, \boldsymbol{x}_j) I_{\mathrm{a}}(\boldsymbol{n}_i, \boldsymbol{\eta}_j) \tag{B.4}$$

for all viewpoints combining requirements 1) and 2) as a distance-based and observation angle-based reward $I_{\mathrm{d}}(\boldsymbol{c}_i, \boldsymbol{x}_j)$ and $I_{\mathrm{a}}(\boldsymbol{n}_i, \boldsymbol{\eta}_j)$.

B.3.4.1 *Distance*

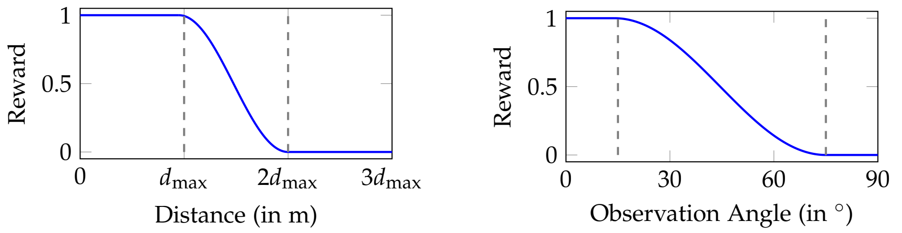
The resolution of the reconstruction depends on the camera intrinsics and the acquisition distances toward the object surface and is usually defined as the GSD or point density after the dense matching reconstruction step. High-resolution models are of high interest for modeling, monitoring and inspecting objects and can be realized by the use of high-resolution cameras or capturing close-up views of the object. Since the goal of the path planning is to provide an equal point density for every part of the object, regardless of its shape and height, we define a maximum distance threshold $d_{\max}$ between a viewpoint and the observed surface points in order to achieve a user-specified model resolution. The maximum tolerable distance to obtain the required GSD also depends on the camera intrinsics and is given by $d_{\max} = \frac{\mathrm{GSD} \cdot f}{\text{pixel size}}$ with a focal length f. We define a smooth symmetrical function $I_{\mathrm{d}}(d)$ for the distance $d = \|\boldsymbol{c}_i - \boldsymbol{x}_j\|$ between a camera viewpoint $\boldsymbol{p}_i$ and an object surface point $\boldsymbol{x}_j$, that assigns maximum reward for a distance less than $d_{\max}$ and decreasing returns up to a multitude of $d_{\max}$. The distance-based reward function

$$I_{\mathrm{d}}(d) = \begin{cases} 1, & \text{if } d < d_{\max}, \\ 0, & \text{if } d > 2d_{\max}, \\ \frac{1}{2}\left(1 - \cos\left(\frac{d\pi}{d_{\max}}\right)\right), & \text{otherwise} \end{cases} \tag{B.5}$$

returns no rewards for distances larger than twice of $d_{\max}$. A visualization of $I_{\mathrm{d}}(d)$ is shown in Figure B.4a.

B.3.4.2 *Observation Angle*

Besides the importance of distances between camera viewpoints and surface points, the observation angles of the camera rays toward the surface normals are also of particular relevance for the quality of the reconstruction. It is commonly known that fronto-parallel views toward a planar surface result in a higher reconstruction quality, due to minor distortions of the objects appearance in the image, which leads to a more robust and reliable matching result (Furukawa and Hernández, 2015). Although viewpoint orientations are already computed and favoring fronto-parallel views, the abundance of viewpoint hypotheses still have to be evaluated according



(a) Distance reward function

(b) Observation angle reward function

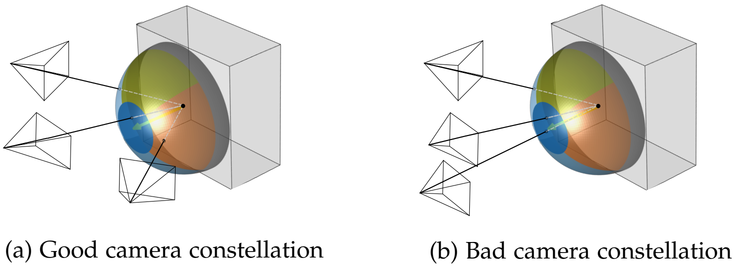
Figure B.4: Heuristics for individual viewpoint candidates considering distances and observation angles toward an object point. Left: Rewards considering the distance between viewpoint and object points regarding the maximum distance d required to achieve a user-specified GSD. Right: rewards based on observation angles defining maximum rewards for fronto-parallel views (here: up to 15°) and zero rewards for more than 75°

to their observation angles. Hence, we adapt our reward function for observation angles $\alpha = \cos^{-1}\left(\boldsymbol{n}_i^\top \cdot \boldsymbol{\eta}_j\right)$ and define two thresholds $\alpha_{\min}$ and $\alpha_{\max}$, where the first is used for maximum rewards for low observation angles and second represents the maximum tolerable observation angle for returning rewards. We define the observation angle-based reward function

$$I_{\mathrm{a}}(\alpha) = \begin{cases} 1, & \text{if } \alpha < \alpha_{\min}, \\ 0, & \text{if } \alpha > \alpha_{\max}, \\ \frac{1}{2}\left(1 + \cos\left(\frac{\pi(\alpha - \alpha_{\min})}{\alpha_{\max} - \alpha_{\min}}\right)\right), & \text{otherwise.} \end{cases} \tag{B.6}$$

Since our experiments focus on the reconstruction of buildings, we follow the proposal of Furukawa and Hernández (2015) pointing out that observation angles up to 15° yield best reconstruction results for planar surfaces, such as building façades. This suggestion is in accordance with the extensive study about the impacts of the acquisition geometry for dense matching algorithms by Wenzel et al. (2013). Therefore we set $\alpha_{\min} = 15°$, while the upper threshold—indicating a failure of MVS algorithms due to large object distortions in the image— was empirically determined to $\alpha_{\max} = 75°$ and approved by the study in (Wenzel et al., 2013). A visualization of the reward function for these thresholds is depicted in Figure B.4b. Note that these values are optimal for the reconstruction of objects mainly composed of flat surfaces, while more complex objects with curved or tilted surfaces would require stricter thresholds.

During the computation of observation angles, we also store observation directions due to the requirement of large parallax angles, as stated in requirement (4). The impact of different parallax angles for the reconstruction quality has already been largely investigated in several works (Förstner and Wrobel, 2016; Kraus, 2011; Wenzel et al., 2013). In particular, a hemisphere is constructed for each surface point $\boldsymbol{x}_j$ directed along its corresponding normal vector $\boldsymbol{\eta}_j$ and discretized into six distinct segments in order to distinguish between different observation directions. A visualization of the hemispheres for potential camera constellations is shown in Figure B.5. Aside from a segment for frontal views occupying an area of a unit circle with an opening angle of $\alpha_{\min}$ and a segment for discarded observation angles above $\alpha_{\max}$, the remaining segments occupy equal areas on the surface of the hemisphere.


(a) Good camera constellation

(b) Bad camera constellation

Figure B.5: Observation angle segments for two camera constellations. A hemisphere is generated along the objects point normal vector and divided into different segments. Each camera ray intersects the hemisphere in a specific segment. Good camera constellations intersect the hemisphere in different segments (a), while camera pairs with ego-motion and small baselines intersect in the same segment (b)

Each viewpoint ray toward an object surface point intersects the hemisphere in a specific segment $\text{seg}(\boldsymbol{p}_i, \boldsymbol{s}_j) \in \{0, 1, 2, 3, 4\}$, which is stored for all visible viewpoint and surface point pairs. We exploit this result in order to find suitable viewpoints intersecting the hemispheres in as many segments as possible and avoiding similar intersection segments, leading to shallow parallax angles and glancing intersections.

B.3.5 *Submodular Trajectory Optimization*

Recent works have shown that the task of path planning for MVS image acquisition can be efficiently addressed by employing submodularity to the candidate view selection (Hepp et al., 2018b; Roberts et al., 2017) enabling approximation guarantees on the solution using greedy methods. With respect to our notation in Section B.3.1, submodularity is a property of a set function $f : 2^{|\mathcal{P}|} \rightarrow \mathbb{R}$ that assigns each subset $\mathcal{T} \subseteq \mathcal{P}$ a value $f(\mathcal{T})$. $f(\cdot)$ is submodular if for every $\mathcal{T}_1 \subseteq \mathcal{T}_2 \subseteq \mathcal{P}$ and an element $\boldsymbol{p} \in \mathcal{P} \setminus \mathcal{T}_2$ it holds that $\Delta(\boldsymbol{p}|\mathcal{T}_1) \geq \Delta(\boldsymbol{p}|\mathcal{T}_2)$. An equivalent and more commonly used definition of submodularity for $\mathcal{T}_1, \mathcal{T}_2 \subseteq \mathcal{P}$ is given by $f(\mathcal{T}_1 \cup \mathcal{T}_2) + f(\mathcal{T}_1 \cap \mathcal{T}_2) \leq f(\mathcal{T}_1) + f(\mathcal{T}_2)$. In other words, submodularity implies that adding an element to a small subset results in large rewards while adding the same element to a larger subset leads to diminishing returns. Speaking of our path planning problem, as we increase more viewpoint candidates to our trajectory, the marginal benefit of adding another viewpoint with large overlap to the set decreases. Adding the same viewpoint to a smaller set with limited coverage, on the other hand, leads to larger rewards. This property hinders explicit modeling of stereo-matching, as already pointed out by Hepp et al. (2018). Adding a viewpoint to a smaller subset $\mathcal{T}_1$ which does not allow a stereo matching yields less reward (zero, as it is not matchable) than adding it to a larger set $\mathcal{T}_2$ to which it is matchable and therefore it violates the submodularity condition. For that reason, a submodular function $f(\cdot)$ has to be defined which approximates stereo matching in terms of contributions from single views for 3D modeling. This requires $f(\cdot)$ to be both monotone and non-decreasing stated as monotonicity, which means that adding more elements to the set cannot decrease its value. The marginal gain of a viewpoint candidate $\boldsymbol{p}$ toward a trajectory $\mathcal{T}$ is given by $\Delta(\boldsymbol{p}|\mathcal{T}) := f(\mathcal{T} \cup \boldsymbol{p}) - f(\mathcal{T})$. It has been shown that a simple greedy algorithm can be considered for providing a solution of the NP-hard maximization of submodular functions with a reasonable approximation guarantee (Krause and

Golovin, 2014). Similar to Hepp et al. (2018), we constrain our submodular objective function

$$f\left(\boldsymbol{s}_j, \mathcal{T}\right) = \min\left(1, \sum_{\boldsymbol{p}_i \in \mathcal{T}} \frac{1}{v} I(\boldsymbol{p}_i, \boldsymbol{s}_j)\right) \tag{B.7}$$

to limit the maximum reward for each surface point to 1, where v reduces the obtained reward from a single view in order to enforce at least v different views capturing the same surface point $\boldsymbol{s}_j$. Since this objective function is both monotone and non-decreasing, we can transform the individual information rewards $I(\boldsymbol{p}_i, \mathcal{S}_{\boldsymbol{p}_i})$ from Equation B.4 for all viewpoint candidates to tightly additive information rewards I_i^{add} utilizing a simple greedy algorithm given in Algorithm 1. Note that the submodular function in Equation B.7 on its own does not explicitly incorporate stereo matching, as it only considers single contributions based on the distance and observation angles from single viewpoints toward the object surface. However, a stereo matching approximation is firstly given by the matchability graph, ensuring paths along the graph for which viewpoints exhibit large overlap toward preceding viewpoints. Secondly, the greedy algorithm incorporates the observation angle segments by penalizing information rewards for camera viewpoints which intersect already seen surface points in the same observation angle segments. This helps to decrease the additive information rewards for cameras with only little parallax angles and therefore avoids ego-motions in the optimized path which are obstructive for stereo matching.

The greedy method iteratively computes the marginal rewards of each viewpoint for the current reconstructability of each surface point and adds the viewpoint with the highest additive information reward I_i^{add} toward the output set. After each iteration, the reconstructability of all surface points is updated according to the previously selected viewpoint rewards. The marginal reward of remaining viewpoints with similar intersection segments of already considered viewpoints is reduced and therefore these are less likely to be chosen in the next iteration. This procedure is repeated until the marginal rewards of all viewpoints have been considered and assigned to the output set. After executing the greedy method, each viewpoint candidate $\boldsymbol{p}_i$ is coupled with a marginal information reward I_i^{add} representing its value for the reconstructability of the object. Roberts et al. (2017) presented an efficient way to transform additive rewards into a standard additive orienteering problem, formulated as a *mixed-integer programming (MIP)* problem, which can be solved with off-the-shelf solvers.

An orienteering problem can be considered as a combination of a traveling salesman problem and knapsack problem. In other words, the optimization needs to find a closed path that maximizes the collected rewards under a time or travel budget constraint. However, the choice of a suitable travel budget is hard to predict for some scenes and the optimization will almost always fulfill the full path constraint due to the pure additive nature of the rewards which always increases the full coverage of the model. Given an overestimated path length L^{eucl}, a similar amount of total rewards can be obtained with a shorter trajectory by penalizing lengthy paths with a regularization factor λ. With respect to the semantical restriction on the airspace,

Algorithm 1 The greedy method for maximizing a monotone submodular function

```
function GREEDY($\mathcal{P}, \mathcal{S}, I(\cdot)$)
    $I \leftarrow \forall \boldsymbol{p} \in \mathcal{P}$ : compute $I(\boldsymbol{p}, \mathcal{S})$        ▷ Compute individual rewards for all viewpoints
    Seg $\leftarrow \forall \boldsymbol{p} \in \mathcal{P}$ : compute seg$(\boldsymbol{p}, \mathcal{S})$        ▷ Compute intersection segments for all viewpoints
    $R \leftarrow \emptyset$        ▷ Initialize reconstructability of object $\mathcal{S}$
    $H \leftarrow \emptyset$        ▷ Initialize observation directions of object $\mathcal{S}$
    for $m \leftarrow 0$ to $|I|$ do
        $i^{\text{add}} \leftarrow \arg\max_{i \in I} f(R \cup i) - f(R) - |H \cap \text{Seg}_i|$
        $R \leftarrow R \cup i^{\text{add}}$
        $H \leftarrow H \cup \text{Seg}_{i^{\text{add}}}$
        $I \leftarrow I \setminus \{i^{\text{add}}\}$
    end for
    return $R, i^{\text{add}}$
end function
```

the optimized trajectory must not exceed a user-defined path length L^{sem} above restricted objects. Summarized, the optimization objective can be formulated as

$$
\begin{aligned}
\mathcal{T}^* = \arg\max_{\mathcal{T}} & \sum_{\boldsymbol{p}_i \in \mathcal{T}} I_i^{\text{add}} - \lambda \sum_{\boldsymbol{e}_k \in \mathcal{E}} w_k^{\text{eucl}} \\
\text{subject to} & \sum_{\boldsymbol{e}_k \in \mathcal{E}} w_k^{\text{eucl}} < L^{\text{eucl}}, \\
& \sum_{\boldsymbol{e}_k \in \mathcal{E}} w_k^{\text{sem}} < L^{\text{sem}},
\end{aligned}
\tag{B.8}
$$

where I_i^{add} defines the additive rewards of the nodes along a path $\mathcal{T}$ with traversed Euclidean distances $\sum_{\boldsymbol{e}_k \in \mathcal{E}} w_k^{\text{eucl}}$ and traversed distances above semantical restricted airspaces $\sum_{\boldsymbol{e}_k \in \mathcal{E}} w_k^{\text{sem}}$. The regularization forces to reduce the maximum path length L^{eucl} for similar optimization results in shorter paths. The second constraint allows the optimization to select nodes in restricted but not prohibited airspaces but, however, encourages to find the most efficient and shortest path through these conditionally accessible airspaces.

B.4 EXPERIMENTS

We evaluated the proposed path planning approach both qualitatively and quantitatively with a series of different experiments using synthetic and real-world data. To provide a more profound analysis of the influence of semantic restrictions, the following evaluation consists of two components. First, we need to evaluate the reconstruction results using our pipeline without semantic constraints in order to validate the general path planning itself. Secondly, comparing these baseline results with the reconstruction results of using paths which consider the semantic constraints. In the optimal case, paths which follow the semantic restrictions should return similar reconstruction results but avoid flyovers of certain objects. Since the complete pipeline from image acquisition to the final 3D model consists of several different tasks, including SfM and MVS, the reconstruction results are highly influenced by the performance of these algorithms. For this reason, we decided to use a state-

Table B.1: Statistics of the datasets used in our experiments.

Dataset	Data Type	Extent of Building (in m)	Extent of Scene (in m)	Nr. of Nodes	Grid Spacing (in m)	Required GSD (in cm)
House	Synthetic	16 × 8 × 12	50 × 50 × 30	2643	3	2.0
Silo	Real	25 × 22 × 25	93 × 85 × 30	2328	4	2.0
Farm	Real	60 × 16 × 9	110 × 65 × 30	1716	5	1.5

of-the-art 3D reconstruction pipeline for all experiments with the same settings to analyze the differences in the 3D models as a result of the different acquisition plans. Since Pix4D (Pix4Da) is a well-known photogrammetric mapping software which integrates state-of-the-art processing steps for both SfM and MVS steps and therefore is often used for processing UAV images, we decided to choose this software for our experiments. However, the results after processing the images with comparable software (*e.g.*, Colmap (Schönberger and Frahm, 2016)) do not substantially differ from the results of Pix4D. For the sake of simplicity we therefore only report the results using Pix4D. In order to investigate whether adjacent image viewpoints can be successfully matched, only temporally neighboring images were matched instead of an exhaustive image matching strategy.

Finding suitable data for a comprehensive analysis is hard to realize, as minor modifications of the parameters can end up with different trajectories, for which all of them need to be executed in individual flights. Moreover, comparing the reconstruction results lack the availability of ground truth data on a large scale. For this reason, we generated a synthetic dataset, composed of various objects arranged to a realistic and interchangeable scenery, which allows comparing the reconstruction results derived from different trajectories with exact ground truth. Additionally, we also show the real-world applicability with two real sceneries consisting of different building shapes with a diverse complexity of the surroundings.

B.4.1 *Synthetic Scene*

We introduce a new customized synthetic scene (the synthetic scene is freely available at `https://www.bgu.tum.de/lmf/synbuil/` which was generated with the open-source computer graphics software Blender (Blender). The main object of interest is a conventional living house located at the center of the scene, consisting of a balcony as overhang, an inset doorway and an adjacent garage. The buildings façades are textured with dirt allowing for a dense reconstruction without severe gaps from homogeneous areas. The roof consists of individual 3-dimensional roof tiles allowing for investigations of detailed structures. The building is surrounded by obstacles, such as trees and adjacent buildings placed beside a main road crossing the building. Additionally, a couple of cars are located on the roadside and in front of the building, which are later used to further restrict the airspace. Figure B.6 shows an overview of the synthetic scene while properties of the scene are listed in Table B.1.

The 3D proxy model, which was considered as input for all methods which were investigated, was created from rendered RGB images of 10 nadir-directed viewpoints at a safe altitude of 70 m encompassing the whole scenery. All images were rendered with a resolution of 750 × 500 px for a virtual camera with a sensor size of $22.2 \times 14.6\,\text{mm}^2$ and a focal length of 30 mm. Since our semantic segmentation

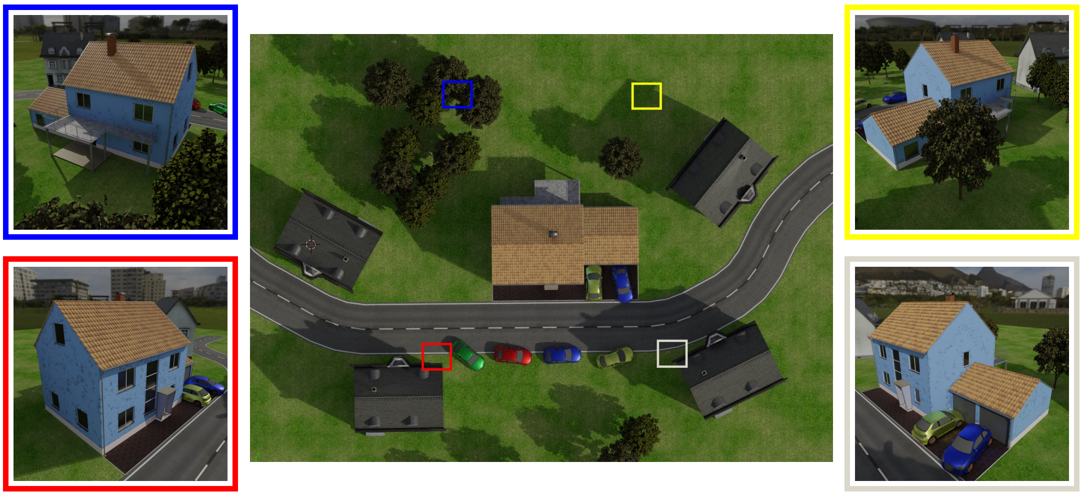


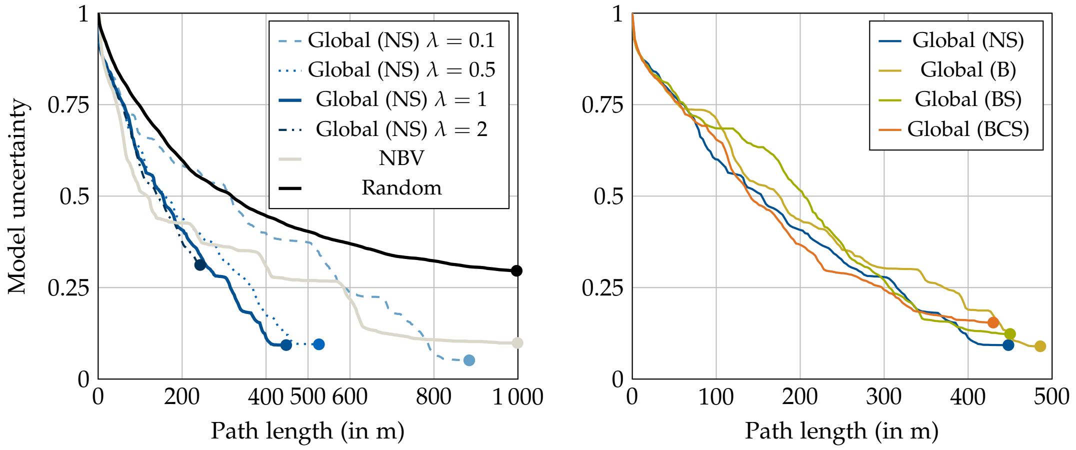
Figure B.6: Overview of our synthetic scene used in our experiments (mid). Sample views of the building for the highlighted areas are shown in the left and right column.

network was trained on real UAV images, the generalization on synthetic data is rather poor. For that reason, we additionally rendered semantic maps for all nadir views directly from Blender. The 3D proxy model was derived by feeding the rendered RGB images into Pix4D for generating a dense 3D point cloud, which was further enriched with the semantic maps following the strategy in Section B.3.2. Evenly distributed viewpoint hypotheses were sampled in the free airspace from a regular 3D grid with a spacing of 3 m, while keeping a safe distance of 3 m toward all obstacles. The camera orientations for all viewpoints were assigned with the strategy explained in Section B.3.3.

B.4.1.1 *Optimization Evaluation*

An analysis of the overall performance of the proposed methodology, as well as the influence of the path length regularization term was conducted with the introduced synthetic scene. A series of estimated trajectories with different regularization parameters were compared toward both automated and manual baseline trajectories. Since the objective function in Equation B.1, in combination with the proposed heuristics, serves as a measure of the expected certainty of the object's reconstructability R, individual results for the reconstructability can be derived for arbitrary subsets $\mathcal{T}$ of the camera graph without the need of acquiring and processing the images. A study concerning the influence of the regularization for jointly optimizing the reconstructability and the path length was conducted by first optimizing a solely geometrical trajectory without semantic constraints with different regularization parameters λ and an overestimated path length $L_{\mathrm{eucl}} = 1000\,\mathrm{m}$.

Figure B.7a depicts the expected model uncertainty with respect to the path length for different values of λ. As expected, low values (*e.g.*, $\lambda = 0.1$) increase the optimized path lengths but also yield a higher degree of certainty, while large values ($\lambda = 2$) result in shorter paths but reduced certainties of the reconstructability. A reasonable compromise of short path lengths and high model certainty can be realized for regularization parameters in the range of $\lambda = 1$. Compared to the optimized path with $\lambda = 0.1$, a minor loss of 4.2% of the model certainty is recognizable for $\lambda = 1$ whereas the path length has been reduced by half.



(a) Comparison of the reconstructability for different optimization approaches as a function of path length

(b) Comparison of semantically-aware optimization for $\lambda = 1$

Figure B.7: Comparison of different optimization methods in terms of the expected model uncertainty for different path lengths assuming the same objective function. The effects of various regularization parameters are shown in blue and the performances of baseline approaches are depicted in gray and black (a). Note that $\lambda = 1$ leads to a balanced trade-off between short path lengths and high model certainty. Comparison of the semantically-aware global optimization with $\lambda = 1$ for different restrictions on the airspace (b)

We compare the global optimization against two other automated path planning baselines, specifically a random trajectory and an online-capable NBV approach for which both make use of the same camera graph, heuristics, and objective function. Regarding the random trajectory, subsequent views are randomly sampled from the camera graph, while the shortest paths between the selected nodes in our graph are computed until a total path length of 1000 m is reached. Each visited node between two sampled nodes is considered as an acquisition viewpoint. This procedure was repeated for 50 times and the averaged model uncertainty for all obtained trajectories are shown in Figure B.7a. Due to the random sampling, highly redundant views from similar positions above the building and only a few views capturing the buildings façades result in a larger degree of model uncertainty compared to the globally optimized trajectories. The path planning strategy of the NBV method starts—similar to the global method—from the viewpoint with maximum reward and greedily selects the next best view from the neighboring nodes according to their marginal rewards. Again, this strategy was repeated until a path length of 1000 m was reached. Comparing toward to the global optimization, NBV rapidly decreases the model uncertainty, as it traverses along the largest gradients of the marginal rewards. However, due to the local search strategy and the highly non-linear nature of the objective function, the NBV approach can get stuck in a local minimum in already seen areas which results in diminishing marginal rewards. This characteristic property is clearly evident in the plateaus of Figure 4.9a. Summarizing, the NBV method can be effective for fast exploration of the object due to the gradient-based optimization, but, however, does not guarantee to recover all local details of the object. The global approach, on the other hand, exploits submodularity which contributes to the selection of suitable viewpoints covering all parts of the object, while the global optimization refines all viewpoints of the trajectory simultaneously, leading to less redundant acquisition views. It is evident, that the global approach is superior toward the baselines in terms of shorter flight paths and higher model certainty. In addition, the globally optimized trajectories have also been proven to be superior in terms of the quality of the generated 3D models, as presented in Section B.4.1.3.

B.4.1.2 *Semantically-Aware Optimization Evaluation*

Following our study in Section B.4.1.1, a regularization parameter of $\lambda = 1$ allows for a reasonable trade-off between short path lengths and high model certainty and is therefore kept for further experiments investigating the semantic constraints of the airspace. Since path optimizations in Section B.4.1.1 only consider purely geometric constraints resulting in a collision-free and matchable viewpoint path in the camera graph, we additionally restrict and prohibit certain airspaces according to the semantics of the underlying proxy model. Depending on the application, restrictions can be defined in two ways: a hard restriction eliminates nodes and their corresponding edges above a certain semantic cue in the camera graph, while soft restrictions limit the path length to a maximum tolerable distance L^{sem} above specified semantic cues. The latter is realized by the secondary condition in Equation B.1.

Precisely, we optimized three semantically constrained trajectories with the following restrictions:

- *No semantics (NS)*: this path from Section B.4.1.1 serves as a baseline and only considers geometric constraints.
- *Building (B)*: hard restriction for airspaces above other buildings than the target building.
- *Building & Street (BS)*: in addition to (B), airspaces above streets are softly restricted to maximum path length of $L = 12\,\text{m}$, approximately twice the width of a regular street.
- *Building & Car & Street (BCS)*: In addition to (BS), hard restrictions above cars are imposed.

The semantic constraints affect the generation of the camera graph, resulting in a limited number of accessible nodes (hard restrictions) and only conditionally accessible nodes (soft restrictions). Statistics of the affected nodes and edges for the synthetic scene are listed in Table B.2. The optimization for the semantically-aware path plans was conducted in the same fashion as in Section B.4.1.1, except for the additional side constraint for the soft restrictions above streets. A comparison of the optimized paths with respect to the path length and model uncertainty is shown in Figure B.7b. From this figure it can be seen that, despite further restrictions in the airspace, only slight losses in the model certainty have to be expected from the optimized paths, indicating that satisfactory reconstruction results can be achieved from these restricted trajectories with a similar path length.

As follows from the visualization of the optimized paths in Figure B.8, the increase of restrictions on the airspace has a substantial influence on the estimated path along the camera graph, yet yielding reasonable trajectories encompassing the entire building while avoiding prohibited objects. It is worth noting that the soft constraint on streets for (BS) and (BCS) result in trajectories which simply cross the street twice in a direct way at suitable locations. In terms of flight safety, these trajectories are by far more desirable than the unconstrained path, since risky long-term periods above

Table B.2: Effects of semantic constraints on the graph generation for the synthetic scene. The free airspace is further restricted for various semantical constraints affecting the number of accessible and conditionally accessible nodes

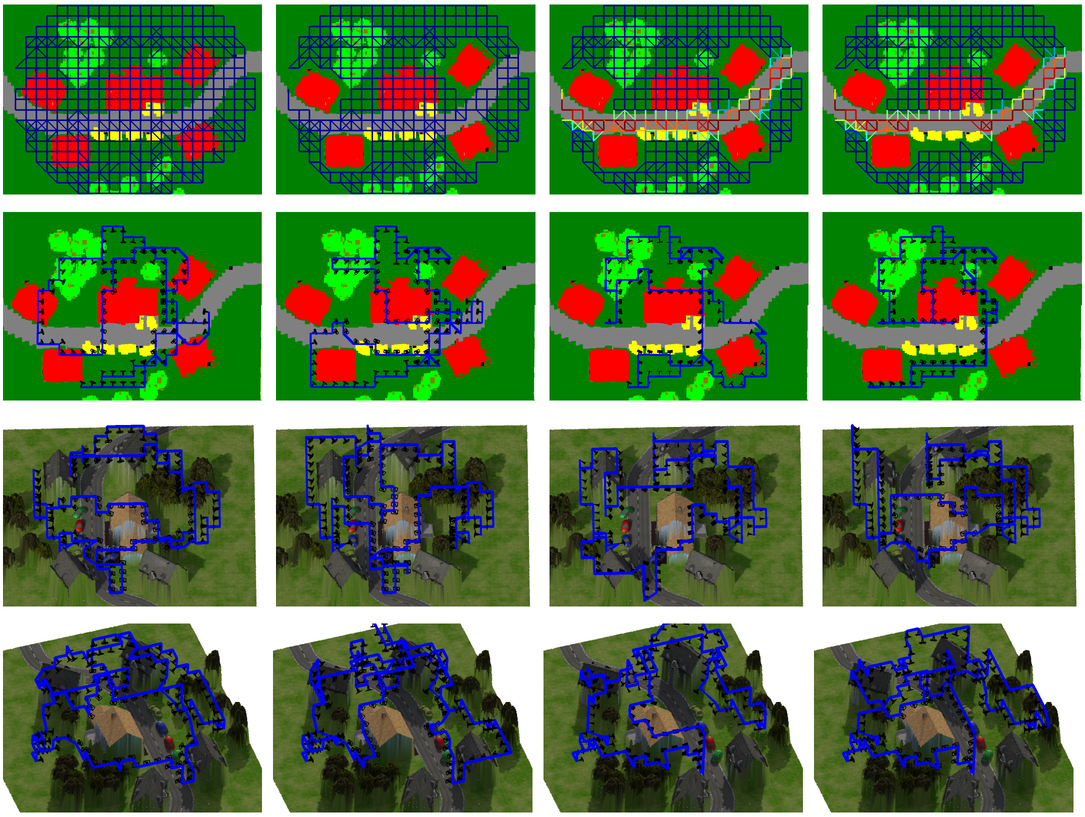
Constraint	Nodes		Edges	
	Free	Restricted	Free	Restricted
No semantics (NS)	2643	0	13,634	0
Building (B)	2333 (88%)	0	11,836 (87%)	0
Building & Street (BS)	2333 (88%)	459 (20%)	11,836 (87%)	2555 (21%)
Building & Car & Street (BCS)	2208 (83%)	388 (17%)	11,084 (81%)	2164 (20%)

hazardousness roads are mostly avoided. As outlined in Figure B.9, the validity of the semantical restrictions can also be expressed as histograms of traversed semantic labels of the proxy model for the optimized paths. While viewpoints above streets are favored for (NS) and (B), they are highly avoided for (BS) and (BCS).

Since the heuristics were already computed for all potential viewpoints, the expected reconstruction quality can be assessed for only subsets obtained by the estimated trajectories. This allows for investigations whether the photogrammetric requirements are met for each surface point by analyzing the observation distances and observations angles between surface points and selected viewpoints, as well as their multi-view configuration. Distributions of the individual photogrammetrical properties of each surface point for different semantically-aware trajectories are shown in Figure B.10. It is apparent that around 75% of the surface points were mapped from at least three different perspectives according to the observation direction segments when considering the non-semantically restricted path (NS). Changes in the semantic-based restrictions on the free airspace only affected 4.4% of the surface points for the utmost restriction on the airspace (BCS). Regarding the observation angles, up to 64% of the surface points were seen within 15° observation angle and 85% with less than 30°, indicating the compliance of fronto-parallel views. Similar to the multi-view assessment, further restrictions on the airspace had only a minor effect on the observation angles. The maximum distance for achieving a GSD below 2 cm with the virtual camera is $d_{\max} = 20\,\mathrm{m}$ which was met for 70% of the surface points. It is worth mentioning, that for an increasing restriction on the airspace even closer views were selected. Reason for this finding is that viewpoints with an optimal distance toward the object could be restricted and eluded to closer views for gaining at least an equal amount of rewards instead of more distant views with fewer rewards.

B.4.1.3 *Reconstruction Performance*

The use of the synthetic model allows for a revealing quantitative and qualitative evaluation of the reconstruction quality from arbitrary viewpoints. RGB images from the viewpoints of the globally optimized paths and baseline paths were rendered in Blender and subsequently processed in Pix4D, including the registration of the images and the generation of a densified point cloud, which was further assessed with respect to the ground truth model. Following the evaluation protocol of related works (Hepp et al., 2018b; Knapitsch et al., 2017; Roberts et al., 2017; Smith et al., 2018), the quality of the reconstructed point clouds can be quantitatively assessed by comparing them toward the ground truth model using the quantities of



(a) Global (NS) (b) Global (B) (c) Global (BS) (d) Global (BCS)

Figure B.8: Visualization of the optimized paths for different semantical restrictions on the airspace for the synthetic scene. The first row shows a nadir view of the entire camera graph as accessible and traversable UAV viewpoints. Color-coded edges represent associated semantical costs w_k^{sem} for the corresponding restrictions. The second row visualizes the optimized camera paths together with the acquisition viewpoints as black camera symbols. Different perspectives with the RGB proxy model are shown in the third and fourth row

precision, completeness, and F-score. Precision quantifies how many reconstructed points are located close to the ground truth model with a distance equal or less than an investigated threshold d. Completeness is defined as vice versa and quantifies how many ground truth points are located in an equal or less distance toward the reconstructed points than d. We analyzed the results for two different thresholds $d_1 = 5\,\text{cm}$ and $d_2 = 10\,\text{cm}$. Furthermore, an assessment of the point density, which is required to be consistent along the entire object surface, was conducted by computing geometrical distances between neighboring reconstructed points.

Additionally, we compared the optimized paths against commonly used flight planning baselines, precisely we generated circular flights at two different altitudes and radii (30 m altitude with 30 m radius and 20 m altitude with 37 m radius) with oblique views pointing toward the center of the building. A quantitative evaluation regarding the reconstruction errors and point density error are listed in Table B.3 and a visualization of the spatial occurrences of these errors are shown in Figure B.11. While circular baseline paths revealed unsatisfying reconstruction results in terms of a low point density and gaps in the reconstruction due to occlusions from overhangs of the roof and balcony, the unconstrained global optimization (NS) yielded best reconstruction quality for all investigated errors. The distance-based heuristics led

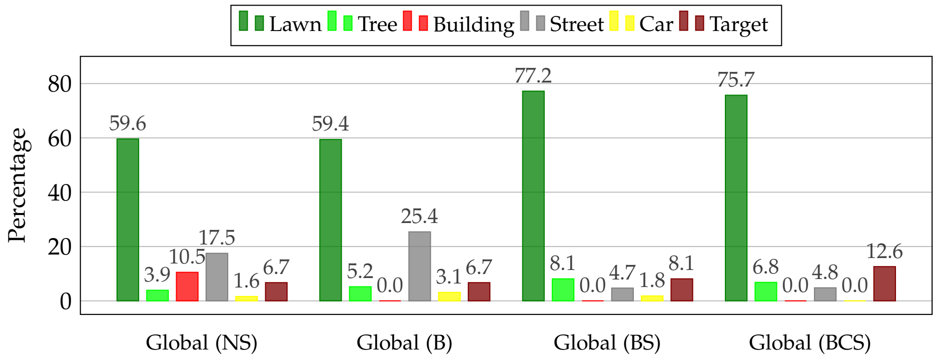
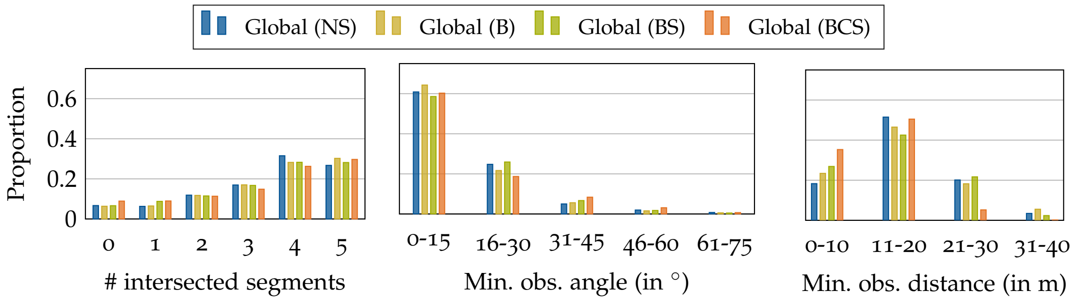


Figure B.9: Proportions of traversed ground labels for different semantically constrained flight paths



(a) Multi-view assessment

(b) Observation angle

(c) Distance

Figure B.10: Evaluation of heuristics for optimized semantically-aware trajectories per surface point *s*. Number of intersected surface hemispheres (**a**), minimum observed observation angles (**b**) and minimum observation distance (**c**)

to close-up views, resulting in a high global point density of more than 97% for all reconstructed points of the building. Comparing the completeness error, lower circular flights yielded less optical occlusions, which, however, is limited by the surrounding environment. Paths considering the proxy model generally performed better in terms of completeness, since low altitude viewpoints can be selected from the free airspace, however globally optimized paths revealed significantly better completeness, especially for occluded areas. The last few percentage points are generally hard to achieve since the building consists of different materials, such as windows, which are generally difficult to reconstruct. It is worth noting that, according to Section B.4.1.1, all globally optimized paths did not exceed a path length of 490 m acquiring a maximum amount of 162 images for (B), while both random and NBV paths were limited to 1000 m resulting in 321 and 323 viewpoints, respectively. The visualizations in Figure B.11 reveal local inaccuracies for the random and NBV paths, whereas all globally optimized paths show decent results for all parts of the building. The most difficult part concerns the façade beneath the balcony and the occluded façade of the garage caused by the single tree, whereby former resulted from a low contrast of the weakly textured and illuminated façade and latter from a hardly observable area.

Comparing the results of different semantic restrictions on the airspace, only a minor decrease in terms of completeness is notable, which matches the expected model uncertainty in Figure 4.9b. Regarding the precision of the reconstruction—a quality measure according to the noise of the reconstruction depending on the

Table B.3: Quantitative evaluation of the reconstruction results for the synthetic scene obtained from different path planning methods. We report the point density as the percentage of reconstructed points that have a shorter distance towards their nearest neighbor than the demanded GSD = 2 cm, as well as one and a half times the distance ($1.5 \cdot \mathrm{GSD} = 3\,\mathrm{cm}$). The reconstruction errors are stated for $d_1 = 5\,\mathrm{cm}$ and $d_2 = 10\,\mathrm{cm}$. The proposed globally optimized paths are superior to the baseline methods, while featuring a shorter path. The severely limited free airspaces due to different semantic restrictions lead only to a slight drop in the reconstruction quality

Method	Images	Density (%) ↑		Precision (%) ↑		Completeness (%) ↑		F-Score (%) ↑	
		GSD	1.5·GSD	d_1	d_2	d_1	d_2	d_1	d_2
Circle 30 m	100	46.9	73.3	88.8	96.3	79.2	91.8	83.7	94.0
Circle 20 m	100	29.7	60.3	89.7	95.8	84.0	93.9	86.7	94.8
Random	321	94.1	98.9	96.4	98.6	83.3	91.0	89.4	94.6
Greedy NBV	323	96.9	99.8	96.8	98.7	86.5	92.7	91.4	95.6
Global (NS)	148	97.6	99.9	96.7	98.9	91.1	95.7	93.8	97.2
Global (B)	162	97.3	99.8	96.2	98.7	88.3	95.5	92.1	97.1
Global (BC)	148	97.6	99.8	96.4	98.7	89.4	94.8	92.8	96.7
Global (BCS)	152	97.3	99.8	96.5	98.8	87.7	95.1	91.9	96.9

camera constellations—it can be noted that all paths considering viewpoints from our camera graph achieved comparable good values, which proves the suitability of the proposed viewpoint generation process in Section B.3.3.

B.4.2 *Real-World Performance*

We show the real-world applicability of our methodology by planning and executing safe flight paths for high-fidelity reconstructions of two buildings. Our experimental site consists of a *Silo* and a *Farm* building, which define the objects to be finally reconstructed from our estimated flight paths with a user-specified GSD for the entire object surface w.r.t. known camera intrinsics. The buildings differ in their shapes, while the surrounding environment features hazardous obstacles—such as high vegetation, buildings, cars and a trunk road—which were considered during the flight planning. An overview of the real-world scenes are depicted in Figure B.12 and statistics of the scene extent are shown in Table D.2. We evaluated and qualitatively compared the reconstructed models generated with acquired images from the estimated trajectories against regular baseline flight paths prepared in accordance with established flight planning practices. A DJI Mavic Pro 2 was used for both experiments, equipped with a 12 Mpx Hasselblad camera with a focal length of 24 mm. The parameters of the camera intrinsics were included in our heuristic computation. The estimated flight plans were finally executed by uploading the waypoints to the UAV, followed by an autonomous acquisition flight without human intervention. Similar to the reconstruction process in Section B.4.1.3, the acquired images were processed in Pix4D for generating a dense point cloud and a triangulated mesh, which served as our final reconstruction model.

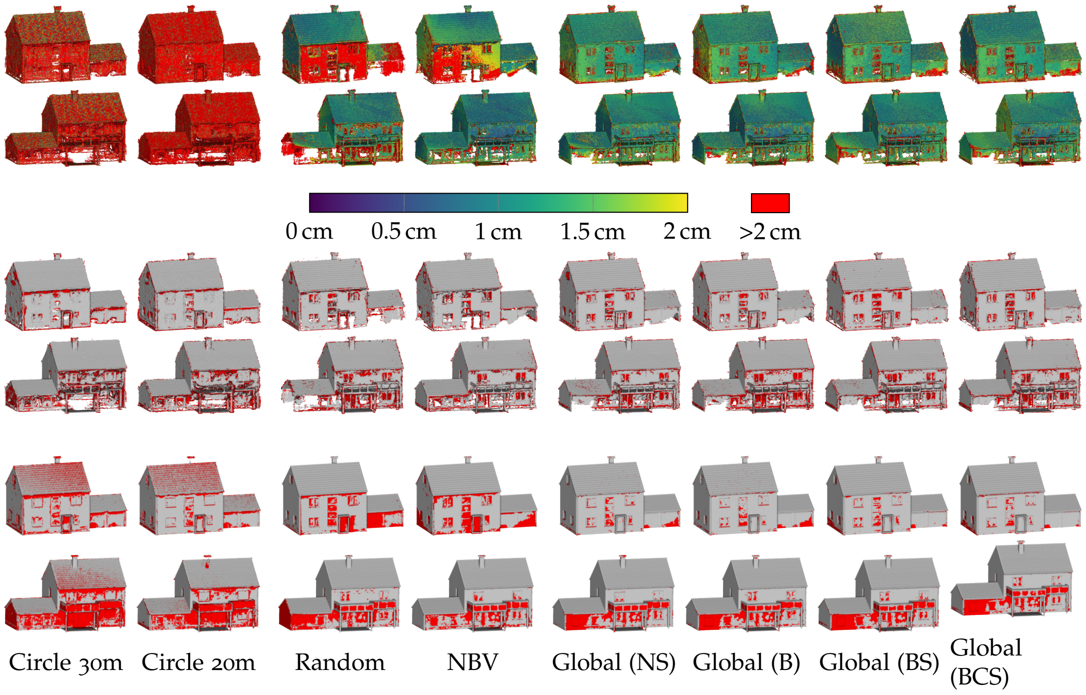
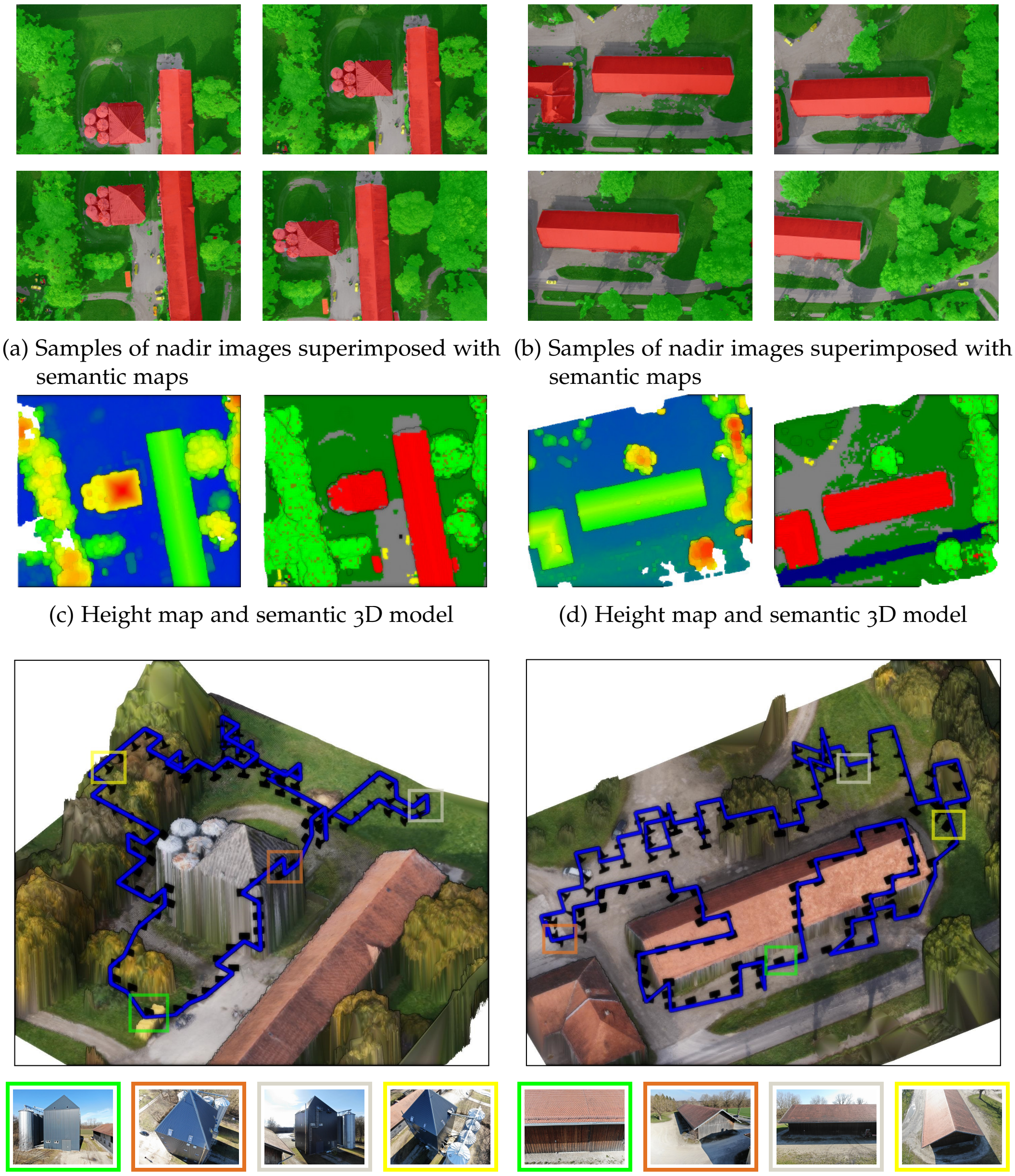


Figure B.11: Qualitative comparison of the reconstruction results on a dense point cloud for the synthetic scene using different methodologies (columns). The first two rows show the point density as colored distances towards adjacent points on the front and rear side of the building, while red points indicate distances above the required GSD of 2 cm. The reconstruction errors of precision and completeness for $d = 5$ cm are visualized in rows three and four, and rows five and six, respectively, wherein red points indicate erroneous areas

B.4.2.1 *Silo*

The first object of interest is a high-rise granary, which features large planar façades. In order to generate a high fidelity reconstruction, the flight path required low flight altitudes for capturing frontal images of the façades as well as glimpses of an occluded façade by contiguous pipes. The surroundings of the granary impeded the execution of a simple circular low altitude flight by high vegetation and another building. We generated an optimized path, avoiding the high-grown trees and restricting fly-overs above the adjacent building.

The initial model was generated from eight nadir images acquired in a grid-like pattern at 100 m altitude encompassing the entire surrounding area. Since the reconstruction of branches and leaves is often incomplete due to their small size and the disturbance of wind affect the consistency of matches across multiple images, we sampled evenly distributed points from a coarse mesh, filling up gaps in the reconstruction model. After inferring the images into the FCN model, we assigned each point in the model with a semantic label leading to the semantic initial 3D model used for our trajectory planning. Visualizations of the semantic input images and the respective proxy model are shown in Figures B.12a and B.12c. A total amount of 2328 viewpoint hypotheses in the accessible airspace was sampled with a grid spacing of 4 m, while keeping a safety buffer of 10 m toward high vegetation and 5 m toward other obstacles. The viewpoints were evaluated in terms of a required GSD of 2.0 cm (for half image resolution) and connected in the camera graph when



(a) Samples of nadir images superimposed with semantic maps

(b) Samples of nadir images superimposed with semantic maps

(c) Height map and semantic 3D model

(d) Height map and semantic 3D model

(e) Acquisition flight path and sample images

(f) Acquisition flight path and sample images

Figure B.12: Real world experiments for the *Silo* scene (left) and *Farm* scene (right). A set of segmented nadir images (a, b) is used to generate a semantically enriched 3D proxy model of the entire scene (c, d). The final trajectories (blue lines) and discrete image acquisition viewpoints (black cameras) are visualized in (e, f) including sample images for the highlighted viewpoints. The restricted areas include adjacent buildings for the *Silo* scene and adjacent buildings, as well as trunk roads for the *Farm* scene

a mandatory overlap of adjacent views of at least 75% was met. The optimization was conducted with $\lambda = 1$, yielding to a trajectory with a path length of 405 m with 98 different views. A visualization of the optimized trajectory and its viewpoints is shown in Figure B.12e. The path features both oblique views covering the roof of the silo and close-up fronto-parallel views of the façades, while it avoids the surrounding trees and passes through the narrow gap between the two buildings

Table B.4: Comparing automatic and semantically-aware trajectories toward established baseline trajectories for the real world experiments including the number of acquired images, the number of acquisition viewpoints above restricted areas, the path length, and the average distance and standard deviation of adjacent 3D points after generating a dense 3D point cloud from the acquired images. More details about the generation of the baseline trajectories are given in Sections B.4.2.1 and B.4.2.2

Dataset	**Baseline**				**Optimized**			
	Images	**Restricted Viewpoints**	**Path Length (m)**	**Density (cm)**	**Images**	**Restricted Viewpoints**	**Path Length (m)**	**Density (cm)**
Silo	90	24	184	2.2 ± 1.2	98	0	405	2.0 ± 0.9
Farm	89	23	732	2.3 ± 0.8	131	0	677	0.9 ± 0.4

without crossing the adjacent building. We compared the reconstruction results using the acquired images of the optimized trajectory against a baseline of a circular flight at a safe altitude of 40 m with 90 acquired images pointing toward the center of the silo. The reconstructed models of both paths are shown in Figure B.13, while Table B.4 compares the trajectories and reconstruction results for the baseline and optimized trajectory. It is evident that our optimized path recovers a higher amount of details than the baseline path, as well as a more complete model, even for hardly observable parts of the silo, such as the highly occluded façade and the façade towards the restricted airspace above the adjacent building. The triangulated mesh exhibits planar façades but still preservers local details, such as sharp edges and almost completely reconstructed pipes with a high level of detail. The visualization of the closest distances for the reconstructed points shows that the desired GSD was achieved for almost every part of the silo, except for the occluded façades. The point density of the baseline model, on the other hand, decreases toward the ground part of the building, due to a fixed flight altitude. Moreover, the baseline path was not able to recover the occluded façade, exhibits distortions in the planar façades, and lost the preservation of local details.

B.4.2.2 *Farm*

The second object is an elongated farm building of low height and with large overhangs from the roof toward the buildings façades. In order to recover the entire building, it is, therefore, necessary to capture images from very low altitudes facing the buildings façades. However, the surrounding environment, as shown in Figures B.12b and B.12d, impeded established flight planning due to high-grown adjacent trees, buildings, and a crossing trunk road. In particular, the latter should be avoided to be overflown, especially at very low altitudes. For that reason, the semantic proxy model was further enriched by the use of OSM data by converting already as road labeled 3D points into restricted areas. Similar to the *Silo* scene, further restrictions were imposed on flights above other buildings.

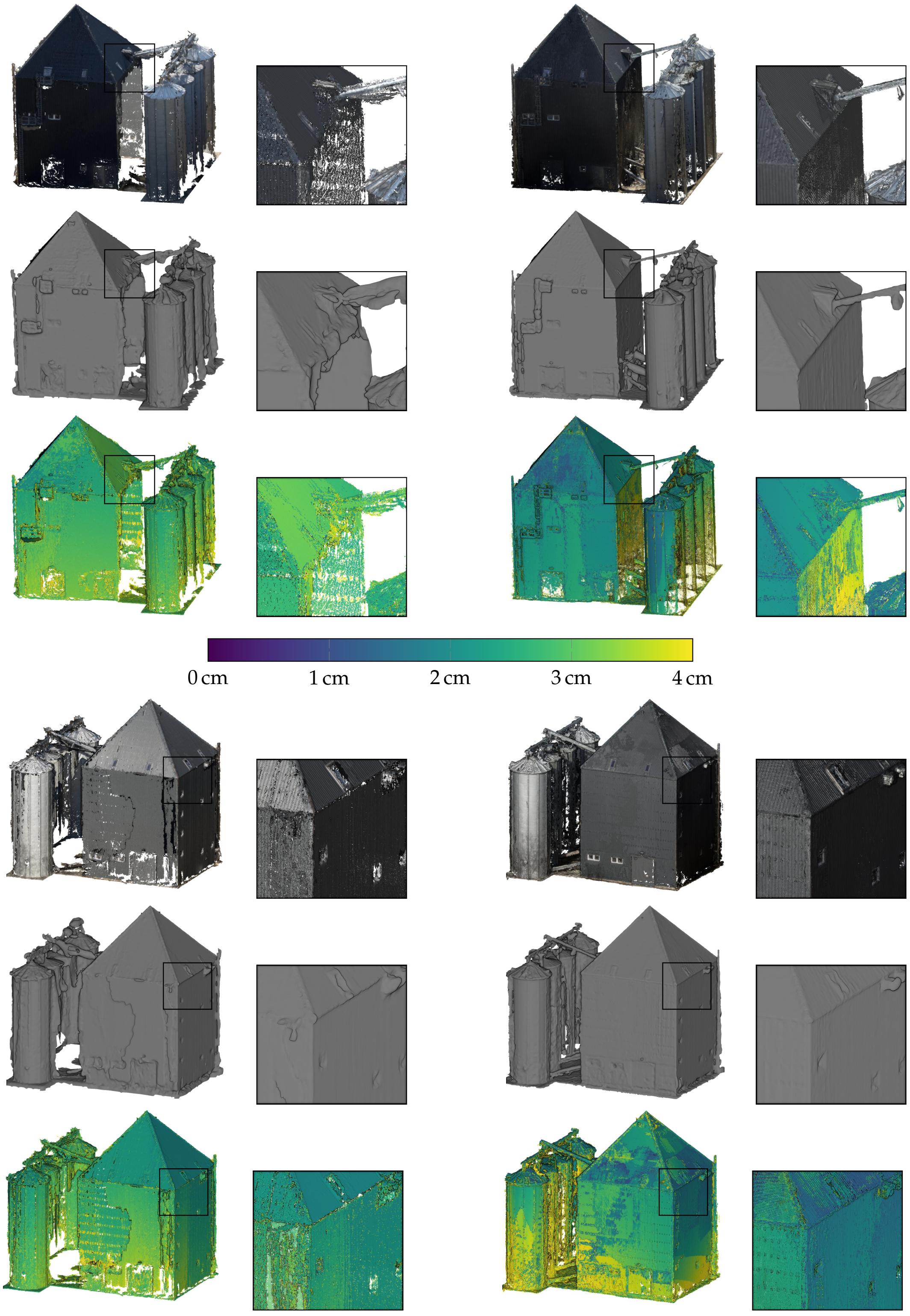
The parameters for the optimization were set in the same way as in Section B.4.2.1, yielding a trajectory with a path length of 677 m and 131 unique image acquisition viewpoints, as shown in Figure B.12f. The trajectory strictly follows the boundary toward the trunk road, avoids the adjacent building, and evades the single tree in front of the buildings façade. Besides oblique images covering the roof of the building, the buildings façades were captured from fronto-parallel viewpoints at very low altitudes up to 5 m. A comparison of the reconstruction result from the optimized trajectory was conducted against a baseline of a grid-like acquisition pattern at 40 m with 89 images pointing toward the center of the building. Table B.4 summarizes both trajectory statistics and quantitative results of the obtained 3D models from the baseline and optimized trajectories, while visualizations of the 3D models are shown in Figure B.14. Due to the large overhangs of the building's roof and the high altitude of the baseline trajectory, the façades are mostly occluded and therefore hardly reconstructed. In contrast, the reconstruction of the buildings façades in the model of the optimized trajectory is vastly improved, with the exception of a partial gap at one side, caused by a technical malfunction of the gimbal of the UAV for some images. It is worth noting, that the optimized model features a similar point density for both the roof and the façades of the building in the range of the required GSD. Comparing the point densities of both models, 95.2%, 99.7% and 99.9% of the reconstructed 3D points derived from the optimized trajectory have an equal or less distance toward adjacent neighboring points for different distance thresholds ($d_1 = 1.50\,\text{cm}$, $d_2 = 2.25\,\text{cm}$, $d_3 = 3.00\,\text{cm}$), while the baseline only achieved 17.1%, 37.1% and 91.8%, respectively.

B.5 CONCLUSIONS

We proposed a semantically-aware 3D UAV path planning pipeline for acquiring images to generate high-fidelity 3D models. Our framework is based on a semantically-enriched proxy model of the environment from a set of safely acquired images, which is used to restrict and prohibit accessible airspaces for the UAV, allowing for safe acquisition paths in complex and densely built environments. An optimized subsequent refinement path allows for acquiring a sequence of close-up images with respect to a user-defined model resolution and fulfills the requirements of SfM and MVS image acquisition, considering the surrounding geometric and semantic environment. We proposed a set of meaningful heuristics and exploit submodularity for formulating the path planning problem as a discrete graph-based optimization. The optimization follows an orienteering problem and maximizes the reconstructability of the object while minimizing the corresponding path length. Additionally, it includes the avoidance of prohibited airspaces and respects conditionally restricted airspaces, such as traversing highly frequented roads.

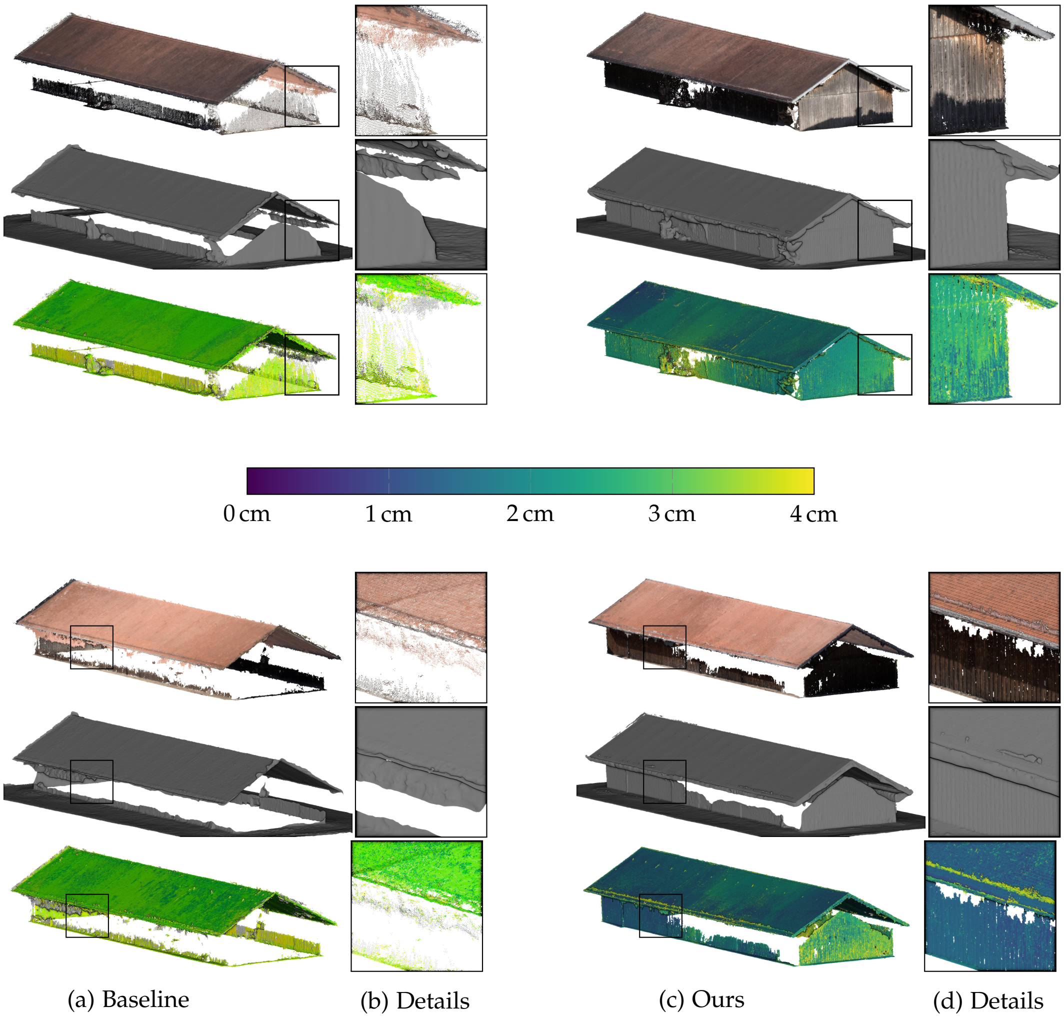
Experiments on synthetic and real-world scenes have demonstrated the applicability of our proposed method requiring only minimal human interaction for complicated scenes, for which established flight plans yield insufficient reconstruction results and highly experienced pilots are demanded for manual operation of the vehicle. We have shown that the optimized trajectories are safe in terms of user-specified restrictions and prohibitions on the accessible airspace but are still capable of generating high-fidelity reconstruction models with respect to the desired model resolution. The model-based approach and the proposed heuristics furthermore allow for retrieving information about the expected reconstruction quality before the actual execution. This allows for further adaptations of the flight path or even the localization of suitable image acquisition viewpoints on the ground level for capturing images of hardly observable parts of the object with a hand-held camera. It is worth noting, that the proposed framework is not limited to building reconstruction tasks, but will work for any 3D object of interest.

The discrete nature of the regularly sampled viewpoints leads to a multitude of images, which are necessary for the registration of adjacent views but do not enhance MVS processing, for which similar results could be achieved with only a subset of the acquired images. This limits the potential placement of viewpoints and leads to over- and undersampled areas. A more flexible viewpoint placement strategy could lead to even more sophisticated viewpoints with a reduced number of hypotheses, thus a reduction of the optimization complexity. Furthermore, it would be conceivable to include additional costs to the optimization for the gimbal motion needed between adjacent viewpoint perspectives in order to minimize the required gimbal operations for the entire flight. Although the optimization would account for estimating a single trajectory for reconstructing several isolated target objects at the same time, the viewpoint orientations are currently assigned toward a single target object. However, an extension of the optimization could incorporate multiple orientations for each camera viewpoint toward several target objects. Hepp et al. (2018) and Roberts et al. (2017) have shown that viewpoint orientations can be integrated in the optimization as well, however, the complexity of the optimization exceedingly increases. A reduction to only few meaningful orientation hypotheses for each viewpoint would be favorable for the optimization, which selects the best perspective for each viewpoint for maximizing the total reconstructabililty of all target objects. Besides leveraging semantics for restricting the airspace, an extension of the trajectory optimization could include respecting the material of individual object parts, such as windows, roofs, and façades, which require customized acquisition requirements. In terms of safe automated flight planning, further research should include keeping the UAV in sight with the pilot at any time during the autonomous acquisition flight, by, for instance, constraining the UAV trajectory optimization to the visible airspace of the pilot's path.



(a) Baseline (b) Details (c) Ours (d) Details

Figure B.13: Qualitative reconstruction results for the *Silo* scene using a baseline UAV path (a) and our optimized path (c). Rows show the densified point cloud (top), a triangulated mesh (mid) and the point density (bottom). Two different viewpoints are visualized which are separated by the colormap of the point density, showing the closest distances between adjacent 3D points



(a) Baseline (b) Details (c) Ours (d) Details

Figure B.14: Qualitative reconstruction results for the *Farm* scene using a baseline UAV path (a) and our optimized path (c). Rows show the densified point cloud (top), a triangulated mesh (mid) and the point density (bottom). Two different viewpoints are visualized which are separated by the colormap of the point density, showing the closest distances between adjacent 3D points

REFERENCES

Alsadik, B., Gerke, M., and Vosselman, G. (2013). "Automated camera network design for 3D modeling of cultural heritage objects." Journal of Cultural Heritage 14(**6**), pp. 515–526.

ArduPilot. *ArduPilot: Mission Planner*. http://ardupilot.org/planner/. Accessed: 2019-05-28.

Bircher, A., Kamel, M., Alexis, K., Burri, M., Oettershagen, P., Omari, S., Mantel, T., and Siegwart, R. (2016). "Three-dimensional coverage path planning via viewpoint resampling and tour optimization for aerial robots." Autonomous Robots 40(**6**), pp. 1059–1078.

Blender. *Blender: a 3D modelling and rendering package*. Blender Foundation. Blender Institute, Amsterdam. URL: http://www.blender.org.

Border, R., Gammell, J. D., and Newman, P. (2018). "Surface Edge Explorer (SEE): planning next best views directly from 3D observations." In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1–8.

Chakrabarty, A. and Langelaan, J. (2009). "Energy maps for long-range path planning for small-and micro-UAVs." In: *Proceedings of AIAA Guidance, Navigation, and Control Conference (GNC)*, p. 6113.

Chen, K., Fu, K., Yan, M., Gao, X., Sun, X., and Wei, X. (2018). "Semantic segmentation of aerial images with shuffling convolutional neural networks." IEEE Geoscience and Remote Sensing Letters 15(**2**), pp. 173–177.

Cheng, P., Keller, J., and Kumar, V. (2008). "Time-optimal UAV trajectory planning for 3D urban structure coverage." In: *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2750–2757.

DJI. *DJI: Flight Planner*. https://www.djiflightplanner.com/. Accessed: 2019-05-28.

Devrim Kaba, M., Gokhan Uzunbas, M., and Nam Lim, S. (2017). "A reinforcement learning approach to the view planning problem." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6933–6941.

Di Franco, C. and Buttazzo, G. (2016). "Coverage path planning for UAVs photogrammetry with energy and resolution constraints." Journal of Intelligent & Robotic Systems 83(**3-4**), pp. 445–462.

Dunn, E. and Frahm, J.-M. (2009). "Next Best View Planning for Active Model Improvement." In: *Proceedings of the British Machine Vision Conference (BMVC)*, pp. 1–11.

Fan, X., Zhang, L., Brown, B., and Rusinkiewicz, S. (2016). "Automated view and path planning for scalable multi-object 3D scanning." ACM Transactions on Graphics (TOG) 35(**6**), p. 239.

Förstner, W. and Wrobel, B. P. (2016). *Photogrammetric computer vision*. Springer.

Furukawa, Y. and Hernández, C. (2015). "Multi-view stereo: a tutorial." Foundations and Trends in Computer Graphics and Vision 9(**1-2**), pp. 1–148.

Furukawa, Y., Curless, B., Seitz, S. M., and Szeliski, R. (2010). "Towards internet-scale multi-view stereo." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1434–1441.

Goesele, M., Snavely, N., Curless, B., Hoppe, H., and Seitz, S. M. (2007). "Multi-view stereo for community photo collections." In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1–8.

Hallermann, N. and Morgenthal, G. (2014). "Visual inspection strategies for large bridges using unmanned aerial vehicles (UAV)." In: *Proceedings of the International Conference on Bridge Maintenance, Safety and Management (IABMAS)*, pp. 661–667.

Hartley, R. and Zisserman, A. (2003). *Multiple view geometry in computer vision*. Cambridge University Press.

Heng, L., Lee, G. H., Fraundorfer, F., and Pollefeys, M. (2011). "Real-time photo-realistic 3D mapping for micro aerial vehicles." In: *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4012–4019.

Hepp, B., Dey, D., Sinha, S. N., Kapoor, A., Joshi, N., and Hilliges, O. (2018a). "Learn-to-Score: efficient 3D scene exploration by predicting view utility." In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 437–452.

Hepp, B., Nießner, M., and Hilliges, O. (2018b). "Plan3D: viewpoint and trajectory optimization for aerial multi-view stereo reconstruction." ACM Transactions on Graphics (TOG) 38(**1**), p. 4.

Hoppe, C., Wendel, A., Zollmann, S., Pirker, K., Irschara, A., Bischof, H., and Kluckner, S. (2012). "Photogrammetric camera network design for micro aerial vehicles." In: *Proceedings of the Computer Vision Winter Workshop (CVWW)*. Vol. 8, pp. 1–3.

Huang, R., Zou, D., Vaughan, R., and Tan, P. (2018). "Active image-based modeling with a toy drone." In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1–8.

ISPRS Potsdam. *ISPRS 2D Semantic Labelling Contest - Potsdam*. http://www2.isprs.org/commissions/comm3/wg4/2d-sem-label-potsdam.html. Accessed: 2019-05-28.

Jing, W., Polden, J., Tao, P. Y., Lin, W., and Shimada, K. (2016). "View planning for 3D shape reconstruction of buildings with unmanned aerial vehicles." In: *Proceedings of the IEEE International Conference on Control, Automation, Robotics and Vision (ICARCV)*, pp. 1–6.

Kaiser, P., Wegner, J. D., Lucchi, A., Jaggi, M., Hofmann, T., and Schindler, K. (2017). "Learning aerial image segmentation from online maps." IEEE Transactions on Geoscience and Remote Sensing 55(**11**), pp. 6054–6068.

Katz, S., Tal, A., and Basri, R. (2007). "Direct Visibility of Point Sets." ACM Transactions on Graphics (TOG) 26(**3**), p. 24.

Knapitsch, A., Park, J., Zhou, Q.-Y., and Koltun, V. (2017). "Tanks and Temples: benchmarking large-scale scene reconstruction." ACM Transactions on Graphics (TOG) 36(**4**), p. 78.

Kraus, K. (2011). *Photogrammetry: geometry from images and laser scans*. Walter de Gruyter.

Krause, A. and Golovin, D. (2014). *Submodular Function Maximization.*

Kriegel, S., Rink, C., Bodenmüller, T., and Suppa, M. (2015). "Efficient next-best-scan planning for autonomous 3D surface reconstruction of unknown objects." Journal of Real-Time Image Processing 10(**4**), pp. 611–631.

Kumar Ramakrishnan, S. and Grauman, K. (2018). "Sidekick policy learning for active visual exploration." In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 413–430.

Loianno, G., Thomas, J., and Kumar, V. (2015). "Cooperative localization and mapping of MAVs using RGB-D sensors." In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4021–4028.

Long, J., Shelhamer, E., and Darrell, T. (2015). "Fully convolutional networks for semantic segmentation." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440.

Luhmann, T., Robson, S., Kyle, S., and Boehm, J. (2013). *Close-range photogrammetry and 3D imaging*. Walter de Gruyter.

Marmanis, D., Wegner, J. D., Galliani, S., Schindler, K., Datcu, M., and Stilla, U. (2016). "Semantic segmentation of aerial images with an ensemble of CNNs." International Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS) III(**3**), pp. 473–480.

Mendez, O., Hadfield, S., Pugeault, N., and Bowden, R. (2017). "Taking the scenic route to 3D: optimising reconstruction from moving cameras." In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Vol. 3, pp. 4687–4695.

Meng, Z., Qin, H., Chen, Z., Chen, X., Sun, H., Lin, F., and Ang Jr, M. H. (2017). "A two-stage optimized next-view planning framework for 3-D unknown environment exploration, and structural reconstruction." IEEE Robotics and Automation Letters 2(**3**), pp. 1680–1687.

Michael, N., Shen, S., Mohta, K., Kumar, V., Nagatani, K., Okada, Y., Kiribayashi, S., Otake, K., Yoshida, K., Ohno, K., et al. (2012). "Collaborative mapping of an earthquake damaged building via ground and aerial robots." Journal of Field Robotics 29(**5**), pp. 832–841.

Mostegel, C., Rumpler, M., Fraundorfer, F., and Bischof, H. (2016). "UAV-based autonomous image acquisition with multi-view stereo quality assurance by confidence prediction." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR-WS)*, pp. 1–10.

Mostegel, C., Prettenthaler, R., Fraundorfer, F., and Bischof, H. (2017). "Scalable surface reconstruction from point clouds with extreme scale and density diversity." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 904–913.

Nex, F. and Remondino, F. (2014). "UAV for 3D mapping applications: a review." Applied Geomatics 6(**1**), pp. 1–15.

Palazzolo, E. and Stachniss, C. (2018). "Effective exploration for MAVs based on the expected information gain." Drones 2(**1**), p. 9.

Peng, C. and Isler, V. (2019). "Adaptive view planning for aerial 3D reconstruction." In: *Proceedings of the International Conference on Robotics and Automation (ICRA)*, pp. 2981–2987.

Pix4Da. *Pix4D: Professional Photogrammetry and Drone Mapping Software*. http://www.pix4d.com/. Accessed: 2019-04-22.

Pix4Db. *Pix4D: Capture*. https://pix4d.com/product/pix4dcapture/. Accessed: 2019-05-28.

Precisionhawk. *Precisionhawk: Precision Flight*. https://www.precisionhawk.com/precisionflight/. Accessed: 2019-05-28.

Roberts, M., Dey, D., Truong, A., Sinha, S., Shah, S., Kapoor, A., Hanrahan, P., and Joshi, N. (2017). "Submodular trajectory optimization for aerial 3D scanning." In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 5324–5333.

Rumpler, M., Irschara, A., and Bischof, H. (2011). "Multi-view stereo: redundancy benefits for 3D reconstruction." In: *Proceedings of the Workshop of the Austrian Association for Pattern Recognition (AAPR)*.

Schönberger, J. L. and Frahm, J.-M. (2016). "Structure-from-motion revisited." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4104–4113.

Semantic Drone Dataset. *TUG ICG: Semantic Drone Dataset*. http://dronedataset.icg.tugraz.at. Accessed: 2019-05-28.

Smith, N., Moehrle, N., Goesele, M., and Heidrich, W. (2018). "Aerial path planning for urban scene reconstruction: a continuous optimization method and benchmark." In: *Proceedings of the ACM SIGGRAPH Conference and Exhibition on Computer Graphics and Interactive Techniques*, p. 183.

Snavely, N., Seitz, S. M., and Szeliski, R. (2006). "Photo Tourism: exploring photo collections in 3D." ACM Transactions on Graphics (TOG) 25(**3**), pp. 835–846.

– (2008). "Skeletal graphs for efficient structure from motion." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8.

Stumberg, L. von, Usenko, V., Engel, J., Stückler, J., and Cremers, D. (2016). "Autonomous exploration with a low-cost quadrocopter using semi-dense monocular SLAM." arXiv preprint arXiv:1609.07835.

Sturm, J., Bylow, E., Kerl, C., Kahl, F., and Cremer, D. (2013). "Dense tracking and mapping with a quadrocopter." International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS) XL(**1**), pp. 395–400.

Vacanas, Y., Themistocleous, K., Agapiou, A., and Hadjimitsis, D. (2015). "Building information modelling (BIM) and unmanned aerial vehicle (UAV) technologies in infrastructure construction project management and delay and disruption analysis." In: *Proceedings of the International Conference on Remote Sensing and Geoinformation of the Environment (RSCy)*. International Society for Optics and Photonics.

Wendel, A., Maurer, M., Graber, G., Pock, T., and Bischof, H. (2012). "Dense reconstruction on-the-fly." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1450–1457.

Wenzel, K., Rothermel, M., Fritsch, D., and Haala, N. (2013). "Image acquisition and model selection for multi-view stereo." International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS) XL(**5**), pp. 251–258.

Zhu, X. X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., and Fraundorfer, F. (2017). "Deep learning in remote sensing: a comprehensive review and list of resources." IEEE Geoscience and Remote Sensing Magazine 5(**4**), pp. 8–36.

C

KOCH, T., KÖRNER M., FRAUNDORFER F. (2016) AUTOMATIC ALIGNMENT OF INDOOR AND OUTDOOR BUILDING MODELS USING 3D LINE SEGMENTS. IN PROCEEDINGS OF THE IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION WORKSHOPS, PP. 689–697.

This chapter represents a pre-print version of the published article with identical content. The original article appeared under doi:10.1109/CVPRW.2016.91.

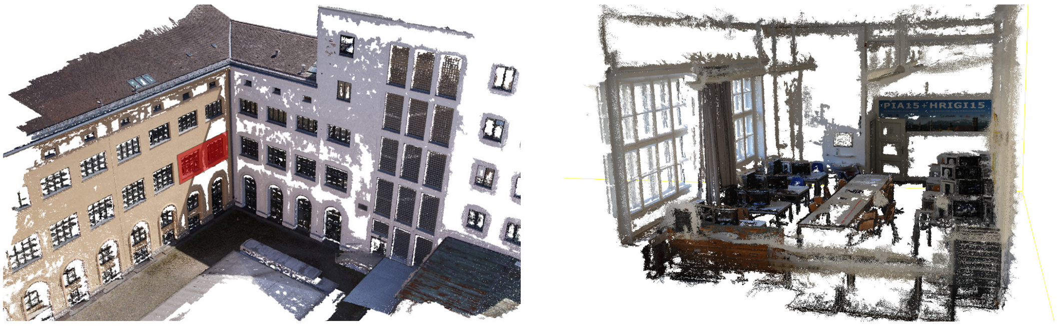
C.1 INTRODUCTION

A cheap and fast way for generating building models is to obtain 3D information from image sequences. Typically, 3D reconstruction pipelines like *Structure-from-Motion* (*SfM*) followed by *Multi-View Stereo* (*MVS*), and meshing are used for small and large scale reconstructions. With the increasing research on image-based indoor modeling in the recent past, an integration of indoor and outdoor models of the same building is consequently the next step. For instance, Figure C.1 shows the reconstruction of our computer-lab which should be connected to the outdoor façade of the building.

When trying to fit a model of the building interior into an existing outdoor model, typically there are no visual correspondences for the alignment using tie points. Therefore, manual work is needed, like using CAD models or floor plans. An automated way providing the true or at least the most probable locations in the outdoor model assumes to reduce human interaction. Since performing a complete reconstruction using continuous image sequences capturing the entire scene by moving from the outside into the inside of the building is either inaccurate caused by drifts or even unfeasible by the lack of matchable features in most cases, an approach using individual reconstructions is desirable. This also allows for matching models generated from image sequences acquired at different points in time.

The most challenging task in matching indoor and outdoor models is to find structures that appear in both image sets but do not describe physically the same part of the scene. To achieve an alignment of these models, topological structures must be found which can be seen from both inside and outside the building, like windows and doors. Identifying and detecting these objects could be done by semantic image segmentation (scene parsing (Brust et al., 2015)) or point cloud analysis (Martinovic et al., 2015). Exploiting the fact that window and door frames can be characterized by dominant and co-planar edges, we propose a method employing 3D line segments.

The contribution of this work is (i) a novel framework for aligning individual image-based 3D reconstructions by (ii) using 3D lines for detecting and matching shared geometric structures in different 3D models.



(a) Building exterior

(b) Building Interior

Figure C.1: Dense point clouds of (a) a building façade from images captured by an UAV and (b) inside our computer lab. The true location of the lab is indicated by the red polygon in (a)

C.2 RELATED WORK

Although the field of 3D reconstruction, scene interpretation, and modelling of man-made objects like buildings is a well-known and widely studied research topic, there is, to our best knowledge, only little research investigating the question how to automatically align 3D indoor and outdoor models reconstructed by individual image sequences. However, the existing demand of integrating multiple image-based reconstruction models can be demonstrated by the example of the very recent *Chillon Project* (Strecha et al., 2014), which aimed to fully reconstruct the interior and the exterior of a complex castle in Switzerland. Due to different camera models and acquisition modes (terrestrial and aerial), a fully automatic reconstruction process is not possible. Instead, multiple sub-models were generated and projected in the same reference coordinate system afterwards in a rather manual way by using Ground-Control-Points or selecting tie points in the images by hand. Although the result shows an impressive reconstruction of a complex architectural object, it also demonstrates the extensive manual interaction which is still needed to connect multiple sub-models.

Cohen et al. (2015) propose a method for merging multiple SfM reconstruction models of a single building which can not be merged due to occlusions or insufficient visual overlap. The approach exploits symmetries and repetitive structures of building façades, as well as semantic reasoning to find reasonable connection points of adjacent models and use them for stitching the models.

In our scenario, we face a similar problem of having no visual overlap when trying to stitch indoor and outdoor models. However, in place of finding connection points, which do not exist in the separated models anyway, we try to find shared geometrical structures that appear in both models like window frames and doors. These shapes can be expressed as edge maps and matched to find suitable connections. When trying to match similar shapes in edge images, *chamfer matching* (Barrow et al., 1977) is widely used, especially in presence of clutter and incompleteness. In our approach, we make use of 3D lines to generate such edge maps which are finally tested for suitable correspondences using chamfer matching.

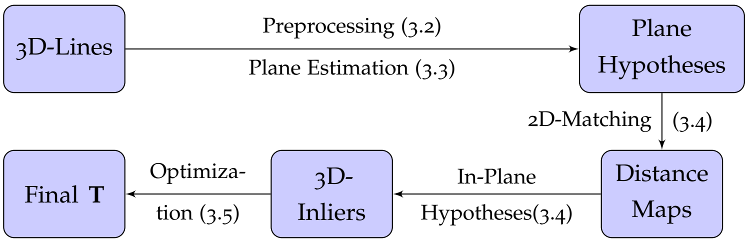


Figure C.2: Workflow of the proposed method for aligning building interior and exterior. See denoted chapters for details

C.3 SYSTEM PIPELINE

This section describes the pipeline of our proposed model matching approach, as illustrated in Figure C.2. After giving an overview about the basic concept of the method, a detailed description of the individual parts is provided.

C.3.1 *Overview*

Given two sets of 3D line segments $L_1 = \left\{l_1^1, ..., l_1^n\right\}$ and $L_2 = \left\{l_2^1, ..., l_2^m\right\}$, the overall goal is to find a transformation $\mathbf{T} = (\mathbf{R}, \mathbf{t}, s)$ to align L_1 to L_2, where **t**, **R** and s define the parameters of a 3D similarity transformation as a 3D translation vector, a 3×3 rotation matrix and a scale. Each segment l is defined by its two endpoints. After identifying $i = 1, .., k$ corresponding line segments in L_1 and L_2, the parameters of **T** can be estimated by

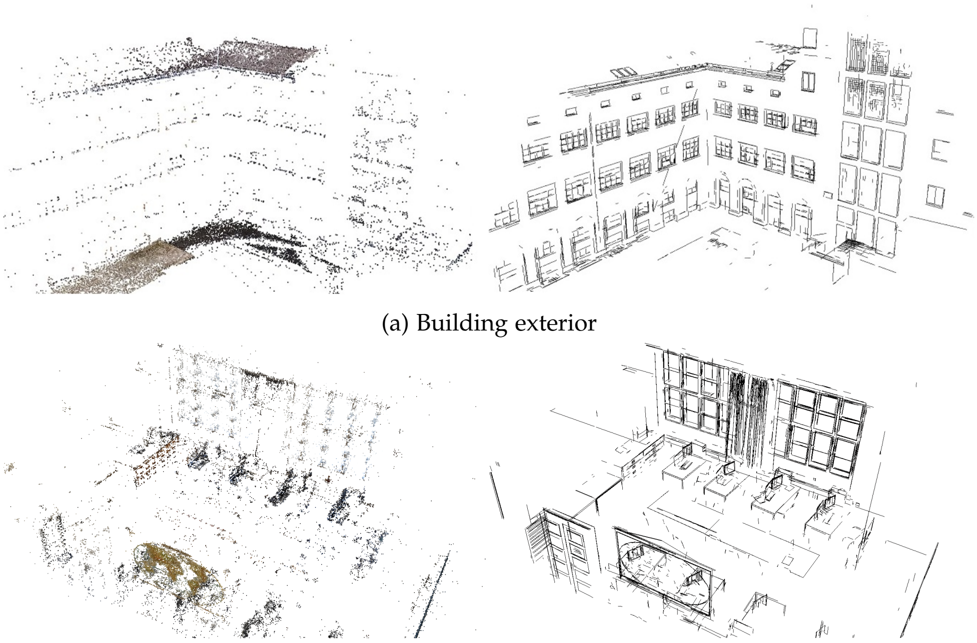
$$\mathbf{T} = \arg\min_{T} \sum_{i=1}^{k} d\left(l_2^i, \pi\left(l_1^i, \hat{\mathbf{T}}\right)\right), \tag{C.1}$$

where $\pi\left(l, \hat{\mathbf{T}}\right)$ projects a line segment l with $\hat{\mathbf{T}}$, and $d\left(l_2, l_1\right)$ computes the length of the perpendicular of two 3D line segments extended to infinity.

As only a small subset out of several thousand pairs of 3D line segments in $L_1 \times L_2$ are expected to be correct 3D line matches, an exhaustive matching scheme is not applicable. Instead, the matching problem is reduced to 2D by defining multiple plane hypotheses in both models, projecting 3D lines onto these planes, and performing 2D binary matching. From the resulting distance maps, local minimums can be extracted which indicate potentially matching locations of the indoor model. After coarse alignment and identifying 3D line correspondences, a refinement of **T** is applied in 3D by minimizing Equation C.1.

C.3.2 *3D Line Generation*

In a first step, for interior and exterior models, 3D line segments have to be generated from a set of overlapping images. This is realized by initially computing image orientations using, *e.g.*, classical SfM pipelines, like *VSfM* (Wu), *Pix4D* (Pix4Da), or *Bundler* (Snavely et al., 2006). As the following line segment reconstruction step assumes images to be undistorted, radial distortion in the images should be removed in advance or modeled within the SfM process. Further, both models need to be approximately equally scaled. This can be achieved by fixing the scale in the SfM process by including one known real-world distance, the usage of GPS information,



(a) Building exterior

(b) Building Interior

Figure C.3: Sparse point cloud (left) and corresponding 3D line segments (right) for building exterior (a) and interior (b) of the *Office* dataset

or a calibrated stereo camera configuration. Although the building interior often consists of poorly textured walls — which translates into problems during image matching due to the low number of matchable feature points — a feasible number of feature points for the pose estimation process should be found in most cases. Subsequently, the computed camera orientations and undistorted images are used to generate 3D line segments following the *Line3D* method proposed by Hofer et al. (2015). Figure C.3 shows a comparison of the sparse point cloud obtained from the SfM process and the 3D line segment reconstruction of the building in Figure C.1. It can be clearly seen that the derived sparse point clouds do not contain information in low textured areas, while reconstructed feature points at the façade and window frames only populate on corners and junctions. A detection of shared structures in both models based on the point cloud seems to be unfeasible. MVS approaches help to increase the density of the point cloud, but still perform bad in poorly textured areas like walls or windows, as exemplary shown in Figure C.1. Additionally, the enormous number of obtained 3D points handicap an efficient analysis of the scene structure. However, the reconstructed 3D line segments contain much more geometric information of the scene, particularly in terms of interpreting façades and windows. Additionally, analyzing 3D lines can be done far more efficient by the drastically lower number of lines compared to the densified point cloud, as noted in Table C.1. The alignment of both models by matching corresponding 3D line segments of window frames seems reasonable. We do not assume prior information of the building structure, but expect that window frames can be dissembled to orthogonal and co-planar 3D lines. This allows us to first define possible window plane hypotheses and then to reduce the matching problem from 3D to 2D.

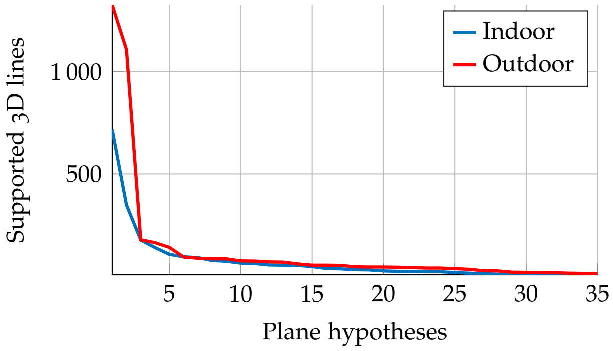


Figure C.4: Number of supported 3D line segments for the 35 most dominant 3D planes in the *Office* dataset

C.3.3 *Window Plane Hypotheses Generation*

This section describes the generation of possible window plane hypotheses which are further used to apply 2D matching and find corresponding 3D line segments in both models.

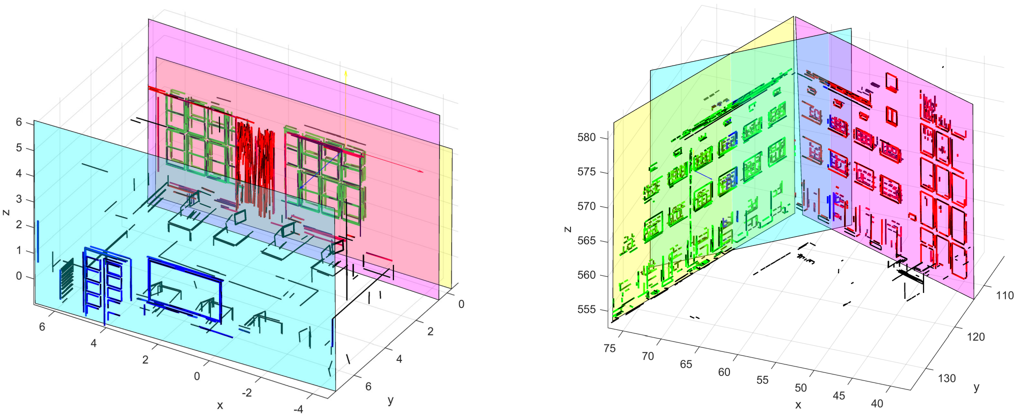
Vertical Alignment Like many man-made constructions, the interior and exterior of buildings mostly consist of planar horizontal and vertical surfaces. This allows us for making use of the Manhattan-world assumption and first identify dominant orthogonal orientations by computing orientation histograms of the 3D lines followed by aligning the estimated vertical axis of the model according to the vertical axis of the coordinate system with the obtained rotation matrix. A similar approach is proposed by Furukawa et al. (2009).

Line Filtering In order to reduce the computational overhead and increase the robustness of the method, subsampling of the 3D lines is performed by eliminating cluttered and skewed 3D lines which unlikely belong to window frames following the Manhattan-world assumption. The set of 3D lines $l \in L$ with length $|l|$ and vertical component of the normalized orientation φ_z are subsampled according to

$$\begin{aligned} L' = \{ l \in L : |l| \geq \tau_l \quad \wedge \\ (|\varphi_z(l)| \leq \tau_\varphi \vee |\varphi_z(l)| \geq (1 - \tau_\varphi)) \} , \end{aligned} \tag{C.2}$$

where τ_l and τ_φ are user-defined thresholds defining a minimal length (*e.g.*, 20 cm) and deviation of the vertical and horizontal axes (*e.g.*, 0.05). Table C.1 lists the number of remaining 3D line segments after the filtering step.

Plane Hypotheses From the set of remaining 3D line segments, multiple window plane hypotheses are generated by assuming co-planar window frames. A *RANSAC* estimation is applied to find dominant 3D planes, wherein inliers are identified as 3D lines lying on this plane within a threshold of the thickness of the plane. Each plane is defined by the intercept of close, orthogonal, and co-planar 3D line segments. The normal of the plane is directed towards the camera from which these lines were reconstructed in order to distinguish between indoor and outdoor sides. We assume that window frames generate substantially more inliers compared to painted walls or



(a) Building interior

(b) Building exterior

Figure C.5: Filtered 3D lines and the three most dominant plane hypotheses in the *Office* dataset. Decreasing number of supporting 3D lines. (a) indoor: 717 (green), 348 (red), 177 (blue); (b) outdoor: 1326 (green), 1108 (red), 178 (blue)

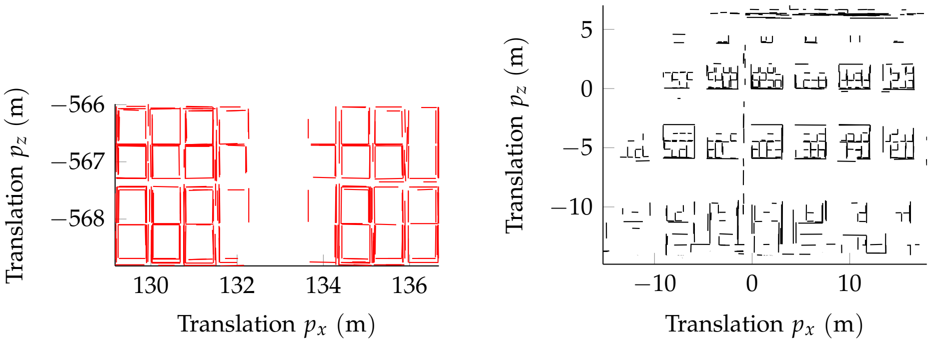
other indoor and outdoor objects. Figure C.4 plots the number of inliers for the first 35 generated 3D planes of the indoor and outdoor model shown in Figure C.1. As expected, the number of inliers decreases rapidly and only a few dominant 3D planes were found. Depending on the complexity of the building, it is mostly sufficient to consider the ten most dominant 3D planes. For the purpose of clarity, only the three most dominant 3D plane hypotheses together with their corresponding inlier lines are illustrated in Figure C.5.

For each pair of computed plane hypotheses, $\hat{\mathbf{T}}$ is now known up to a 2D translation vector within the outdoor plane. The missing parameters can be estimated by first matching every plane hypothesis pair in 2D and then evaluating the matching result to find valid locations.

C.3.4 *Matching Plane Hypotheses*

After computing multiple plane hypotheses, the next step is to determine corresponding plane hypotheses and find valid locations of the indoor model in the outdoor model in order to identify 3D line matches. This is done by performing oriented chamfer matching as described subsequently.

Binary Image Generation For each indoor and outdoor hypothesis, corresponding 3D lines considered as inliers by the plane estimations are projected onto their corresponding planes for generating 2D lines, as illustrated in Figure C.6. It has to be noted that, due to the reconstruction process, the models still contain inaccurate and missing lines, which has to be considered in the matching process. Furthermore, like most buildings, the façade shows highly repetitive structures. In this case, the correct location of the indoor model can not be identified without any further information like adjacent rooms. Instead, all possible valid locations should be returned by the method, whereby the correct one is identified by the user. As chamfer matching requires binary images, the 2D lines are discretized with a user-defined step size (*e.g.*, 5 cm).



(a) Indoor hypothesis 1

(b) Outdoor hypothesis 1

Figure C.6: Projected 3D inliers onto the first plane hypothesis in (a) indoor and (b) outdoor scene

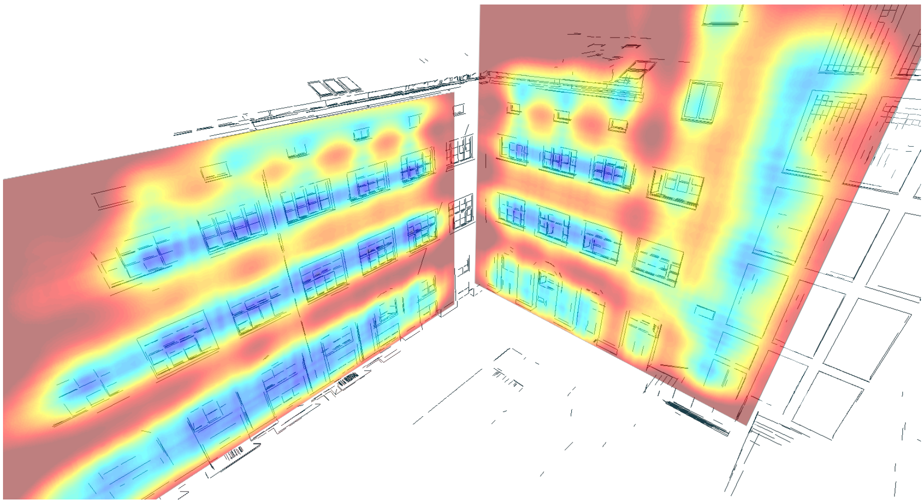
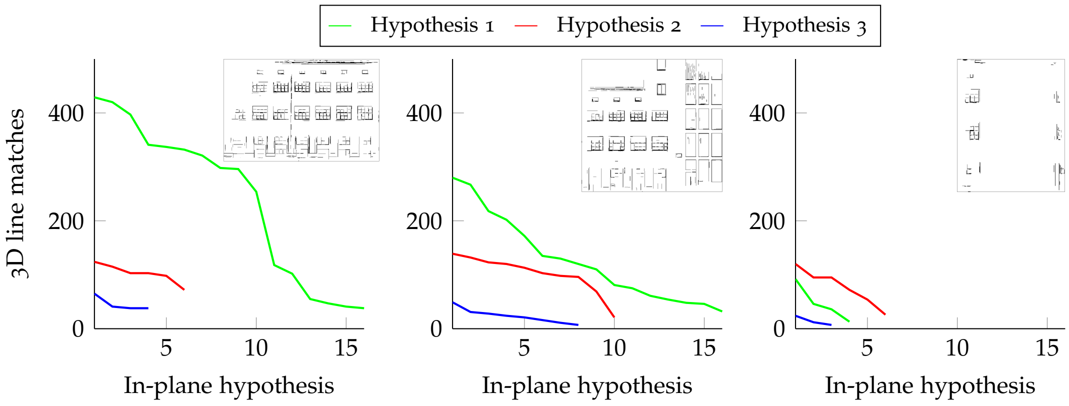


Figure C.7: Chamfer distance maps of indoor hypothesis 1 and outdoor hypotheses 1 and 2 projected on outdoor 3D lines. Both maps are equally scaled, while blue color indicates low distance and therefore likely locations of the indoor model

Oriented Chamfer Matching A popular and efficient technique for shape-based matching is provided by chamfer matching, particulary in presence of incompleteness and clutter. We make use of the oriented chamfer distance (Shotton et al., 2008), which is defined as the mean distance of edge points of a template binary image to their closest edge points in a query binary image, weighted by the orientation differences of closest edge points. This distance can be efficiently computed using *distance transform*, while the orientations of the edge maps can be extracted directly from the 2D line segments.

The resulting chamfer distance map indicates possible locations of the indoor model, the so-called *in-plane hypotheses*. Figure C.7 illustrates the distance maps of matching indoor hypothesis 1 to outdoor hypotheses 1 and 2 projected onto the 3D lines of the outdoor model. Note the low distances for the windows at the first and second floor. However, differences in the scores for different floors are caused by missing edges during the reconstruction process and slightly different window heights for the first and second floor. Multiple in-plane hypotheses are subsequently identified by extracting local minimums in the distance maps.



(a) Outdoor hypothesis 1 (b) Outdoor hypothesis 2 (c) Outdoor hypothesis 3

Figure C.8: Number of 3D line matches for different in-plane hypotheses for the three most dominant indoor and outdoor hypotheses. (a-c) describe different outdoor hypotheses together with their corresponding 2D binary image. Different indoor hypotheses are indicated by different colors, while in-plane hypotheses are sorted by their number of matches

Finding Corresponding 3D Line Segments For each in-plane hypothesis i, a full initial transformation $\hat{\mathbf{T}}_i$ is now available. After transforming all indoor inlier 3D line segments with $\hat{\mathbf{T}}_i$, corresponding 3D line segments can be detected as closest parallel 3D line segments of the outdoor model. Due to the plane estimation, discretization, and multiple window pane layer, the inlier 3D indoor lines are shifted along the normal orientation of the plane until a maximum number of matches is reached. This procedure is repeated for all possible plane combinations and in-plane hypotheses, while the number of detected 3D line matches indicates the quality of the matching. Figure C.8 shows the number of matches for each pair of planes and multiple in-plane hypotheses. Most matches are found by the correct indoor plane hypothesis 1 (green) and the first outdoor plane hypothesis (a), followed by the second façade (b), whereby numerous in-plane locations produce a similar number of matches. Wrong indoor plane hypotheses (red and green) and the wrong outdoor plane hypothesis (c) generate significantly less matches. Figure C.9 illustrates the location of the five most probable in-plane hypotheses. All of them correspond to the first indoor plane hypothesis and first outdoor plane hypothesis.

C.3.5 *3D Refinement*

After obtaining the n most probable in-plane hypotheses and manually choosing the correct one, the parameters of the initial transformation $\hat{\mathbf{T}}$ are still erroneous caused by inaccurate plane estimations, the discretization, or unequal scale of both models, as exemplary shown in Figure C.10a.

A fine alignment is achieved by using the obtained 3D line matches and minimizing Equation C.1. Due to the fact, that corresponding 3D line segments still can vary in their distance - as they could be fragmented during the 3D line generation step - they are extended to infinity. Therefore, the perpendicular distance between matched lines is minimized. Note that this optimization requires both horizontal and vertical line matches in order to eliminate one degree of freedom, but should be satisfied in most cases. Table C.1 summarizes the intermediate results of the alignment and the

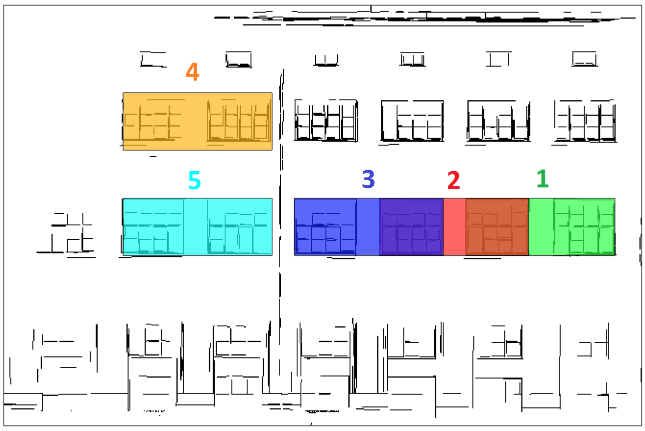
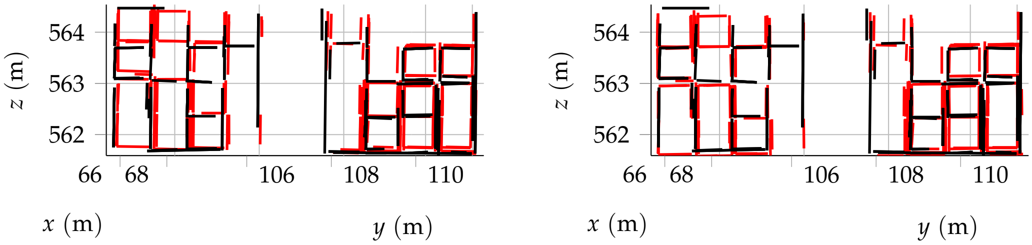


Figure C.9: Most probable locations of the indoor model. The first five hypotheses belong to indoor plane hypothesis 1 and are all located on outdoor plane hypothesis 1



(a) Before 3D optimization

(b) After 3D optimization

Figure C.10: Refining the initial transformation by global 3D optimization: 429 3D line matches for (red) transformed indoor and (black) outdoor lines exemplary shown for the most supported hypothesis

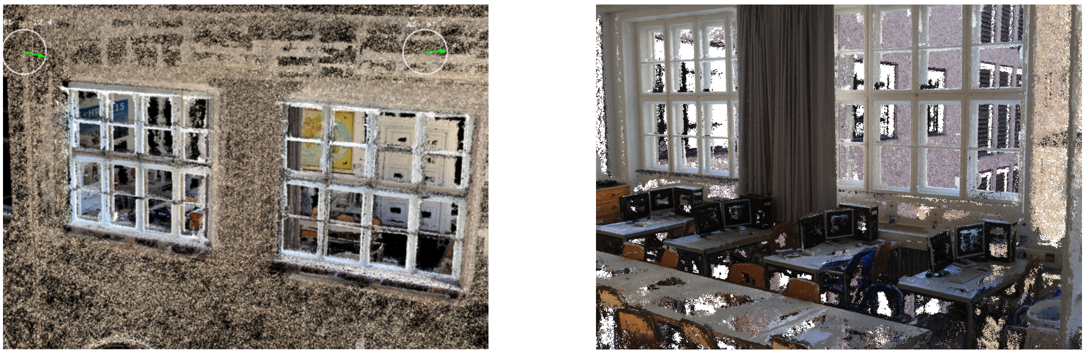
effect of the global optimization. The mean of all perpendicular distances of 3D line matches can be considered as a measure of the alignment accuracy and results in 4.7 cm for the *Office* dataset. A visualization of the aligned 3D line matches before and after the global optimization is illustrated in Figure C.10, while Figure C.11 shows the final alignment of both dense point clouds.

C.4 EXPERIMENTS

Beside the dataset and result in the sections before, another experiment was carried out to illustrate the performance of the method. After giving an overview about the data acquisition and properties of the dataset, intermediate and final results of the alignment are described.

C.4.1 *Dataset Description*

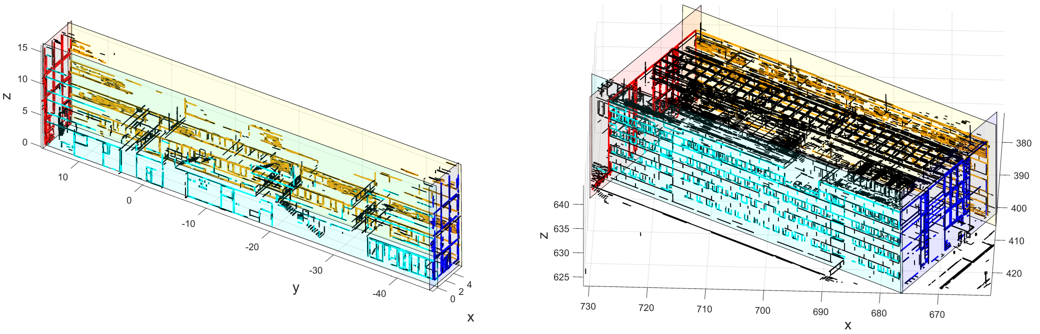
The *Building* dataset contains an outdoor image sequence of a complete building captured from an *UAV* and an indoor hand-held image sequence inside of the building basement. Two large windows at both face sides of the building can be used for stitching the indoor and outdoor model. GPS tags of the aerial images were included in a SfM pipeline to compute a georeferenced, vertically aligned, and



(a) View from Outside

(b) View from Inside

Figure C.11: Aligned point clouds of indoor and outdoor model from different perspectives



(a) Indoor

(b) Outdoor

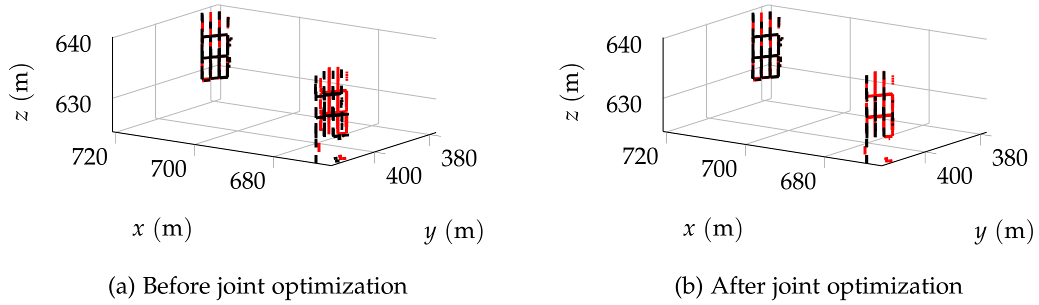
Figure C.12: Filtered 3D lines and the four most dominant plane hypotheses in the *Building* dataset. Decreasing number of supporting 3D lines. (a) indoor: 2827 (orange), 982 (cyan), 412 (red), 272 (blue); (b) outdoor: 2395 (orange), 1617 (cyan), 721 (red), 676 (blue)

correctly scaled reconstruction model. However, one known real-world distance and direction has been included in the indoor reconstruction in order to approximate the orientation and scale of the indoor model.

3D line segments of both models were further generated using the *Line3D* method proposed in Section C.3.2 (*cf.* Figure C.12). A description of the scene and intermediate results for this dataset are given in Table C.1. For further information of this freely available dataset, please refer to Koch et al. (2016).

C.4.2 *Alignment Result*

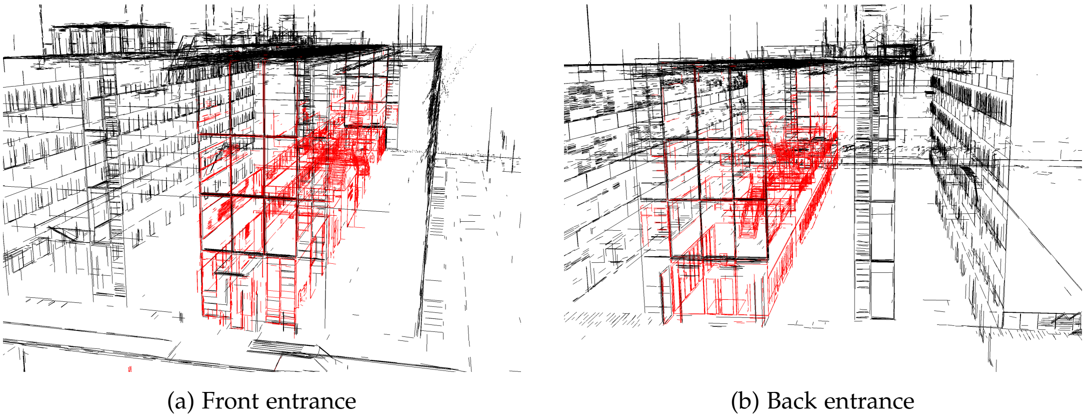
Unlike the dataset used in section 3, the alignment of these models is unique up to a 180° rotation of the indoor model, while the connection can be achieved on both windows sides of the building. The result of the plane hypotheses generation is illustrated in Figure C.12. The four most dominant plane hypotheses represent the four façades of the outdoor model and the two walls and two window sides of the indoor model. In this dataset, the number of matchable plane hypotheses can be reduced by a bounding-box criteria. As the indoor model should not break through the outdoor model, the front and back sides of the indoor model are not matched to the side façades of the outdoor model. Further, as the side walls of the indoor model



(a) Before joint optimization

(b) After joint optimization

Figure C.13: Visualization of (red) transformed indoor and (black) outdoor 3D line matches considering only matches at one face side of the building (a) and matches at both sides (b)



(a) Front entrance

(b) Back entrance

Figure C.14: Final result after joint optimization of front and back entrance from different perspectives

have no connection to the outdoor model and contain different structures, only two main hypotheses remain after the 2D matching step.

172 inlier 3D matches were found when matching one window façade. Due to the building structure, another 157 3D line matches can be added when considering the second hypothesis on the opposite window façade, as shown in Figure C.13b. If only matches at one side of the building are being used, small inaccuracies of the estimated rotation and scale together with the elongated structure of the building (60 m) cause an imprecise fit observed at the opposite side of the building (*cf.* Figure C.13a). Therefore, a joint optimization with matches at both sides is performed which leads to an accurate and robust estimation of **T** with an error of 5.3 cm (*cf.* Figure C.13b). Figure C.14 shows the final alignment of all 3D lines viewed from both sides of the building.

C.5 DISCUSSION AND FUTURE WORK

We have presented an approach for automatically aligning individual indoor and outdoor reconstructions that uses SfM and a 3D line segment reconstruction algorithm. As connecting those kinds of models is mostly restricted to their geometric shapes like windows and doors, 3D lines are well suited for this task. Compared to the extensive generation and analysis of dense 3D points using *Multi-View Stereo*, a comparatively small number of 3D lines offer more interpretable information, at least in detecting and matching geometric shapes.

Table C.1: Properties, intermediate and final results of the experiments *Office* and *Building*. Note the relatively small number of 3D line segments compared to the densified point clouds generated by a standard MVS (Rothermel et al., 2012). Errors are defined as the mean perpendicular distance between 3D line matches before and after global optimization

	Dataset			
	Office		**Building**	
	Indoor	**Outdoor**	**Indoor**	**Outdoor**
Base area (in m²)	75	405	360	1500
Images	247	41	320	228
3D Points (mio)	9	18	13	134
3D lines	4373	3905	10 315	23 801
Filtered 3D lines	1724	2764	6616	21 385
Matches	429		329	
Error pre optim (in cm)	5.7		47.7	
Error post optim (in cm)	4.7		5.3	

The proposed system exploits the planar structures of buildings for generating multiple meaningful matchable hypotheses and is therefore not limited by the complexity of the building. After detecting multiple 3D plane hypotheses, matching can be applied efficiently in 2D by binary image matching methods. However, a more discriminative matching method has to be developed for our task, as standard methods return too many local in-plane minima and hence result in too much computational overhead. This is also the case for reducing the number of meaningful plane hypotheses. A preceding labeling of the 3D line segments using semantic image segmentation could help to include useful priors in the window plane estimation and 2D matching steps.

Beside aligning indoor and outdoor models, this method can also be extended to align individual adjacent room models which are connected by doors. In case of complex building interiors containing multiple rooms, a graph-based approach has to be developed in order to find the correct room constellation.

REFERENCES

Barrow, H. G., Tenenbaum, J. M., Bolles, R. C., and Wolf, H. C. (1977). "Parametric correspondence and Chamfer matching: two new techniques for image matching." In: *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*.

Brust, C.-A., Sickert, S., Simon, M., Rodner, E., and Denzler, J. (2015). "Convolutional patch networks with spatial prior for road detection and urban scene understanding." In: *Proceedings of the IEEE International Conference on Computer Vision Theory and Applications (VISAPP)*.

Cohen, A., Sattler, T., and Pollefeys, M. (2015). "Merging the unmatchable: stitching visually disconnected SfM models." In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2129–2137.

Furukawa, Y., Curless, B., Seitz, S. M., and Szeliski, R. (2009). "Manhattan-World stereo." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1422–1429.

Hofer, M., Maurer, M., and Bischof, H. (2015). “Line3D: efficient 3D scene abstraction for the built environment.” In: *Proceedings of the German Conference on Pattern Recognition (GCPR)*. Springer, pp. 237–246.

Koch, T., d'Angelo, P., Kurz, F., Fraundorfer, F., Reinartz, P., and Körner, M. (2016). “The TUM-DLR multimodal earth observation evaluation benchmark.” In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR-WS)*, pp. 698–705.

Martinovic, A., Knopp, J., Riemenschneider, H., and Van Gool, L. (2015). “3D all the way: semantic segmentation of urban scenes from start to end in 3D.” In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4456–4465.

Pix4Da. *Pix4D: Professional Photogrammetry and Drone Mapping Software*. http://www.pix4d.com/. Accessed: 2019-04-22.

Rothermel, M., Wenzel, K., Fritsch, D., and Haala, N. (2012). “SURE: Photogrammetric surface reconstruction from imagery.” In: *Proceedings of the Low Cost 3D Workshop (LC3D)*. Vol. 8.

Shotton, J., Blake, A., and Cipolla, R. (2008). “Multiscale categorical object recognition using contour fragments.” IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 30(7), pp. 1270–1281.

Snavely, N., Seitz, S. M., and Szeliski, R. (2006). “Photo Tourism: exploring photo collections in 3D.” ACM Transactions on Graphics (TOG) 25(3), pp. 835–846.

Strecha, C., Krull, M., and Betschart, S. (2014). *The Chillon Project: Aerial/ Terrestrial and Indoor Integration*. https://s3.amazonaws.com/mics.pix4d.com/KB/documents/Pix4D-White-Paper-Chillon-Project-.pdf. Accessed: 2019-04-22.

Wu, C. *Visualsfm: a Visual Structure from Motion System*. http://ccwu.me/vsfm/. Accessed: 2019-04-20.

D

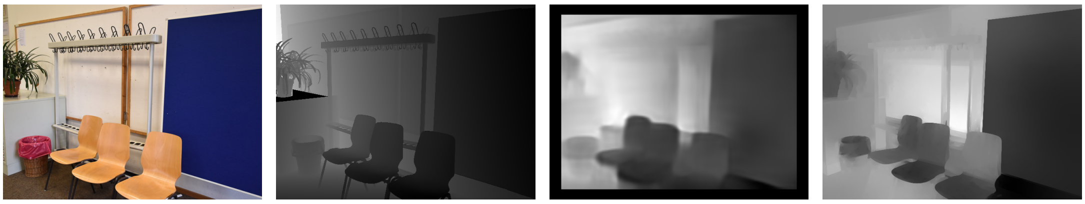
KOCH, T., LIEBEL, L., KÖRNER M., FRAUNDORFER F. (2020) COMPARISON OF MONOCULAR DEPTH ESTIMATION METHODS USING GEOMETRICALLY RELEVANT METRICS ON THE IBIMS-1 DATASET. COMPUTER VISION AND IMAGE UNDERSTANDING. VOLUME 191, 102877

This chapter represents a pre-print version of the published article with identical content. The original article appeared under doi:10.1016/j.cviu.2019.102877.

D.1 INTRODUCTION

Capturing the 3D structure of a scene from a single image is a fundamental question in computer vision and enables manifold scene reconstruction and understanding applications, such as 2D-to-3D conversion (Xie et al., 2016), 3D modeling (Hassner and Basri, 2006), room layout estimation (Izadinia et al., 2017), image refocusing (Shi et al., 2015), foreground-background segmentation (Dhamo et al., 2019), computational cinematography (Devernay and Beardsley, 2010; Phan and Androutsos, 2013), robot navigation (Mancini et al., 2018), autonomous driving (ref), or augmented reality systems (Liu et al., 2018). The process of predicting a depth map of a scene using one or more images is commonly known as *depth estimation* and is usually derived from correspondences across stereo images or motion sequences which provide relatively rich information for understanding 3D structures. In contrast, a broad range of research has dealt with the task of predicting pixel-wise depth maps from monocular images, which is generally referred to as monocular depth estimation (MDE). Among the multitude of different approaches, single-image depth estimation (SIDE) addresses depth prediction from a single view without prior knowledge and, thus, constitutes the most challenging scenario of this discipline. However, recent years have witnessed the fast development of *deep learning* methods and their massive impact on the computer vision domain, which has also affected the progress of SIDE by implicitly learning relevant scene priors to cope with this task. Current state-of-the-art methods replace traditional handcrafted methods and employ SIDE architectures to address the problem of SIDE as a pixel-level regression task. The remarkable results of such methods, exemplary shown in Figure D.1, demonstrate the power of such deep networks by inferring geometrical information solely from monocular RGB or grayscale images.

While these methods produce nicely intuitive results, proper evaluating the estimated depth maps is crucial for subsequent analysis and improvement of the methods, as well as their usability for further 3D understanding scenarios. Consistent and reliable relative depth estimates are, for instance, a key requirement for path planning approaches in robotics (Mancini et al., 2018), augmented reality applications (Liu et al., 2018), or computational cinematography (Devernay and Beardsley, 2010), while preserving the planarity of predicted walls and floors of a room plays a decisive role in room layout estimation applications (Zhuo et al., 2015).



(a) RGB input image (b) Ground truth depth (c) Prediction using method of Eigen and Fergus (2015) (d) Prediction using method of Liu et al. (2015)

Figure D.1: Depth maps produced by different methods, scoring similar errors using standard metrics

Nevertheless, the evaluation schemes and error metrics commonly used so far mainly consider the overall accuracy by reporting global statistics of depth residuals which do not give insight into the depth estimation quality at salient and important regions, like planar surfaces or geometric discontinuities. Hence, fairly reasonable reconstruction results, as shown in Figures D.1c and D.1d, are evaluated with similar errors, although they apparently show different characteristics in terms of ordinal relations, smoothness of planar regions, and defects at object boundaries.

For this reason, we provide a set of new geometrically interpretable error metrics targeting the aforementioned issues allowing for a precise analysis of the performance of depth estimation methods under different perspectives. At the same time, we present a new evaluation dataset[1] acquired from diverse indoor scenarios containing high-resolution RGB images aside highly accurate depth maps from laser scans to overcome the shortage of available datasets providing ground truth data of sufficient quality and quantity.

This work extends our previous work on the evaluation of SIDE (Koch et al., 2018) by providing a more detailed description of our dataset and error metrics, further information on our acquisition procedure and dataset content, and a comprehensive comparison towards other datasets. In addition, we present additional qualitatively and quantitatively results and further experiments that analyze the performance of current state-of-the-art methods for specific situations, such as the presence of textured regions and variations in the scene illumination.

The remainder of the paper is structured as follows: Section D.2 starts with a comprehensive presentation of the current state-of-the-art in deriving depth maps from single and stereo images and reviews existing RGB-D datasets that are used for training and benchmarking purposes. A thorough description of the proposed geometric quality metrics is provided in Section D.3. Section D.4 is devoted to a detailed description of the new IBims-1 RGB-D indoor dataset and a quantitative and qualitative comparison towards the related NYU-v2 (Silberman et al., 2012) dataset. In Section D.5, an assessment of the performance of several current SIDE methods regarding both established and proposed error metrics on IBims-1 is outlined, which reveals novel insights into the performance and differences among the methodologies. Besides the benchmarking protocol, Section D.6 presents additional experiments that aimed to highlight specific properties of the methods, such as robustness towards image augmentations and the influence of texture and illumination cues on the depth estimation. Section D.7 concludes the paper with a concise summary and shares the newly gained insights for further improvements in the field of SIDE.

[1] The dataset is freely available at www.lmf.bgu.tum.de/ibims1

D.2 RELATED WORK

The task of image-based depth perception is a long-standing and active research field, which has been already addressed by a variety of different techniques. The following section provides an overview of both established and novel algorithms with a focus on most recent learning-based methods. Since those data demanding methods rely on a multitude of aligned RGB and depth image pairs for training, the availability of RGB-D datasets has recently increased significantly. Therefore, we introduce and discuss existing datasets used in the field of SIDE in the second part of this section.

D.2.1 *Methodologies*

Recovering depth information from images can be addressed by *single-view* or *multi-view* approaches. The following sections provides an overview of different groups for deriving image-based depth maps. A summary including relevant literatures is listed in Table D.1.

MULTI-VIEW Traditionally, depth information is derived by geometric constraints from multiple observations of a scene using stereo camera setups or leveraging camera motion. The former rely on a prior calibration of the stereo setup and dense point correspondences across the stereo images to estimate depth via geometric triangulation. The task of optimal pixel-wise disparity estimation is usually addressed by local, semi-global, or global optimization methods (Szeliski, 2010). While local methods (Yoon and Kweon, 2006) evaluate pixel correspondences in a point-wise approach, yielding fast, but often inaccurate correspondences due to their sensitivity towards appearance changes and occlusions, global (Felzenszwalb and Huttenlocher, 2006; Kolmogorov and Zabih, 2001) and semi-global methods (Hirschmuller, 2005), on the other hand, make explicit smoothness assumptions and solve for a global optimization problem formulated as energy minimization frameworks, resulting into accurate and less noisy depth maps, but requiring significantly increased computation times. A prominent representative for semi-global methods constitutes the well-known *semi-global-matching (SGM)* algorithm (Hirschmuller, 2005). Methods that leverage monocular camera motion are utilizing *Structure-from-Motion (SfM)* or *Simultaneous Localization and Mapping (SLAM)* methods to transform multiple single-view images to a stereo problem, which can be addressed by multi-view stereo (MVS) methods subsequently (Szeliski, 2010). Extensive studies in the field of two or more frame stereo correspondence algorithms can be found in (Hartley and Zisserman, 2003; Scharstein and Szeliski, 2002; Seitz et al., 2006). A further line of approaches was developed with the emergence of light field cameras using an array of micro-lenses placed in front of the image sensor (Doorn et al., 2011; Heber and Pock, 2016).

SINGLE-VIEW ACTIVE METHODS Another research direction endeavors to ease the multi-view requirement by addressing the task of depth estimation by a sequence of images from the same perspective. Depth information is obtained either by variations of the camera parameters (shape from focus/defocus (Favaro and Soatto, 2005; Suwajanakorn et al., 2015)), by different lighting conditions of the scene (photometric

Table D.1: Overview of different approaches for image-based depth estimation

Group	Approach	Method	Literature
Multi-view	Calibrated stereo setup	Local, semi-global, global	Yoon and Kweon, 2006, Hirschmuller, 2005, Felzenszwalb and Huttenlocher, 2006; Kolmogorov and Zabih, 2001
	Unordered image stacks	SfM + MVS	Hartley and Zisserman, 2003; Seitz et al., 2006; Szeliski, 2010
	Light-field cameras		Doorn et al., 2011; Heber and Pock, 2016
Single-view	Active	Shape from focus/defocus	Favaro and Soatto, 2005; Suwajanakorn et al., 2015
		Lightning conditions	Ackermann and Goesele, 2015
		Polarization cues	Kadambi et al., 2015; Ngo et al., 2015
	Passive	Shape from shading	Horn, 1970; Zhang et al., 1999
		Atmospheric optics	Nayar and Narasimhan, 1999
	Learning-based	Parametric	Baig and Torresani, 2016; Furukawa et al., 2017; Hane et al., 2015; Hoiem et al., 2007; Ladicky et al., 2014; Li et al., 2014; Liu et al., 2010; Ranftl et al., 2016; Saxena et al., 2006; 2008; 2009; Shi et al., 2015; You et al., 2014
		Non-parametric	Choi et al., 2015; Karsch et al., 2014; Kong and Black, 2015; Konrad et al., 2012; Konrad et al., 2013; Liu et al., 2014
	Deep learning-based	Supervised	Chakrabarti et al., 2016; Eigen and Fergus, 2015; Eigen et al., 2014; Fu et al., 2018; Hao et al., 2018; Heo et al., 2018; Hu et al., 2019; Kim et al., 2016; Laina et al., 2016; Lee et al., 2018; Li et al., 2015; Li et al., 2017; Liu et al., 2018; Liu et al., 2015; 2016; Ramamonjisoa and Lepetit, 2019; Roy and Todorovic, 2016; Wang et al., 2015; 2016; Xu et al., 2018; Yang and Zhou, 2018; Zhuo et al., 2015; Zoran et al., 2015
		Unsupervised	Garg et al., 2016; Godard et al., 2017; Kuznietsov et al., 2017; Ummenhofer et al., 2017; Yin and Shi, 2018; Zhan et al., 2018; Zhuo et al., 2015

stereo (Ackermann and Goesele, 2015)) or by utilizing polarization cues (Kadambi et al., 2015; Ngo et al., 2015).

Single-view passive methods Most prominently, *shape from shading (SfS)* methods (Horn, 1970) exploit intensity or color gradients of a single image under the assumption of homogeneous lighting and Lambertian surface properties. Although these methods work on single-shots, they only perform well for largely known environments or synthetic data but rather poor on real images in unconstrained environments (Zhang et al., 1999). Another early approach aimed at exploiting light sources and illumination conditions, such as haze and fog in an image to recover the relative scene depth by relying on atmospheric optical models (Nayar and Narasimhan, 1999).

Single-view learning-based methods As one of the first learning-based approaches, Torralba and Oliva (2002) focused on absolute depth estimation for a query image by incorporating the size of known objects depicted in the image. Instead of decomposing the image into its constituent elements, the absolute scene depth of the image is derived from the global image structure represented as a set of features from Fourier and wavelet transforms. The features of the query image were finally compared towards a model trained with 4000 images and corresponding scene depths in a cluster-weighted modeling approach. With the release of first RGB-D datasets (Geiger et al., 2012; Saxena et al., 2009; Silberman et al., 2012), data-driven approaches became feasible and rapidly began to outperform established model-based methods. A pioneer work of a supervised learning-based approach was firstly proposed by Saxena et al. (2006) by training a discriminatively-trained *Markov random field (MRF)* incorporating multi-scale local and global-image features to infer depth. An extension of this work to 3D scene reconstruction was proposed later (Saxena et al., 2009). Since then, a variety of approaches have been proposed to exploit the monocular cues using hand-crafted features together with graphical models (Baig and Torresani, 2016; Furukawa et al., 2017; Hane et al., 2015; Hoiem et al., 2007; Li et al., 2014; Ranftl et al., 2016; Saxena et al., 2008; Shi et al., 2015; You et al., 2014). Better depth estimates have been achieved by incorporating semantic labels (Ladicky et al., 2014; Liu et al., 2010).

Single-view non-parametric learning-based methods Another cluster of work estimate depth using non-parametric learning-based methods (Choi et al., 2015; Karsch et al., 2014; Kong and Black, 2015; Konrad et al., 2012; Konrad et al., 2013; Liu et al., 2014). These methods assume similarities between RGB values and depth cues across a large set of images. First, similar images of the input image are retrieved from a RGB-D database by feature-based matching. The depth complements of the nearest neighbors are combined and cross-bilateral filtered for smoothing the final depth map (Konrad et al., 2013), warped towards the input image using SIFT flow (Karsch et al., 2014; Liu et al., 2011), or optimized via a conditional random field (CRF) (Liu et al., 2014).

Single-view deep learning-based methods In conjunction with the undeniable influence of deep learning within the field of computer vision, the research was driven towards the use of convolutional neural networks (CNNs) for depth estimation. Since 2014, some works have significantly improved SIDE performance with the use of deep models, demonstrating the superiority of deep features over hand-crafted features (Chakrabarti et al., 2016; Eigen and Fergus, 2015; Eigen et al., 2014; Fu et al., 2018; Kim et al., 2016; Laina et al., 2016; Lee et al., 2018; Li et al., 2015; Li et al., 2017; Liu et al., 2018; Liu et al., 2015; 2016; Roy and Todorovic, 2016; Wang et al., 2015; 2016; Xu et al., 2018; Zhuo et al., 2015; Zoran et al., 2015). These methods pursue the problem of SIDE as a regression problem by building upon successful architectures and learning a deep CNN to estimate the continuous depth map. The first work using deep models was proposed by Eigen et al. (2014) in a two-scale architecture. A coarse global prediction is performed with one network in a first stage, while another network locally refines the prediction in a successive second stage. An extension to this approach uses deeper models and additionally predicts normals and semantic labels (Eigen and Fergus, 2015).

Some works have harnessed the power of pre-trained CNNs in the form of fully convolutional networks (Chakrabarti et al., 2016; Eigen and Fergus, 2015; Laina et al., 2016; Li et al., 2017). The convolutional layers from networks such as AlexNet (Krizhevsky et al., 2012), VGG (Simonyan and Zisserman, 2014) and ResNet (He et al., 2016) are fine-tuned, while the fully connected layers are re-learned from scratch to encode a spatial feature mapping of the scene. One main limitation using CNNs for depth prediction is decrease of resolution of the output map due to repeated pooling operations in the deep feature extractors. In order to preserve the local structures of output depth maps, several authors have attempted to cope with this problem by up-sampling (Chakrabarti et al., 2016; Eigen and Fergus, 2015; Li et al., 2017), up-convolution blocks (Laina et al., 2016), skip connections between the up-sampling blocks (Li et al., 2017) and space-increasing discretization (Fu et al., 2018).

Improving the quality of predicted depth maps was also addressed by combining CNNs and graphical models, such as *conditional random fields* (CRFs) (Kim et al., 2016; Li et al., 2015; Liu et al., 2015; 2016; Wang et al., 2015; Xu et al., 2017; 2018). A *deep convolutional neural field (DCNF)* combining CNNs and CRFs in a unified framework for estimating depth on each superpixel while enforcing smoothness within a CRF was proposed by Liu et al. (2015); Liu et al. (2016). Li et al. (2015) and Wang et al. (2015) use hierarchical CRFs to refine their patch-wise CNN predictions from superpixel down to pixel level. CRFs can be exploited to fuse the multi-scale information derived from inner layers of a CNN (Xu et al., 2017; 2018). A combination of CNNs and regression forests with very shallow architectures at each tree node reduces the need for big data (Roy and Todorovic, 2016). Exploiting the Fourier frequency domain in a deep learning algorithm was proposed by Lee et al. (2018).

After the first success of applying deep architectures for SIDE, authors began to focus on tackling major challenges, such as distorted depth discontinuities (Hao et al., 2018; Hu et al., 2019; Ramamonjisoa and Lepetit, 2019) or planar regions (Heo et al., 2018; Liu et al., 2018; Wang et al., 2016; Yang and Zhou, 2018).

UNSUPERVISED DEEP LEARNING-BASED Recently, unsupervised or semi-supervised learning is introduced to learn depth estimation (Garg et al., 2016; Godard et al., 2017; Kuznietsov et al., 2017; Ummenhofer et al., 2017; Yin and Shi, 2018; Zhan et al., 2018; Zhou et al., 2017). This is accomplished by an intermediate task of a view synthesis, and allows training by only using stereo pairs as input with known baselines. These methods design reconstruction losses to estimate the disparity map by recovering a right view with a left view.

Use of synthetic data With the emergence of synthetic datasets, first work was done to exhibit the possibility to render noise-free and dense depth maps in a very large scale. However, the large domain gaps between synthetic data and real data is still a very challenging task. First works in this field are trying to handle this gap (Guo et al., 2018; Zheng et al., 2018).

Ordinal depth prediction Some applications only require relative or ordinal depth, such as 2D-to-3D conversion (Karsch et al., 2014), image refocusing (Anwar et al., 2017), or foreground-background segmentation (Camplani and Salgado, 2014). Methods in this field predict dense relative depths from pairwise relationships (closer-than and further-than relationships) estimates for rare points in the input image (Chen et al., 2016; Zoran et al., 2015).

D.2.2 *Existing RGB-D Datasets*

In order to train supervised SIDE methods as well as to evaluate and compare them with other approaches, any dataset containing corresponding RGB and depth image pairs can be considered, which also comprises, *e.g.*, benchmarks originally designed for the evaluation of MVS approaches.

This variety of freely available datasets can be categorized according to different criteria (*cf.* Table D.2). Some of them exhibit an adequate number of samples for training deep models, others concentrate on few, but highly accurate, RGB-D image pairs allowing for exhaustive analysis and comparison of different methodologies. The amount and quality of depth maps also depends on the choice of the sensor used for the acquisition campaign. In general, RGB-D image pairs are commonly generated either by active sensors, suchs as RGB-D cameras or laser scanners, or passively by the use of stereo images. While active RGB-D sensors, such as the Microsoft Kinect version 1 and 2, the Occipital Structure Sensor, and the Intel RealSense are pre-calibrated setups, ready to produce aligned depth maps in a large quantity without manual effort, LiDAR sensors are slow and usually need an additional camera and registration technique. However, the quality of generated depth maps from LiDAR are superior to RGB-D sensors in terms of resolution, completeness, range, and accuracy. More recently, researches started to make use of the big amount of image data from freely available image databases, such as Flickr, to generate RGB-D image pairs utilizing stereo vision algorithms. With the generation of synthetic data, data-depending deep learning methods can be fed with innumerable training data.

The following datasets can currently be considered for the task of SIDE. Among the datasets that rely on precise laser scan data, Strecha et al. (2008) propose a MVS benchmark providing overlapping images with camera poses for six different outdoor scenes and a ground truth point cloud obtained by a laser scanner. More recently, two MVS benchmarks, the `ETH3D` (Schöps et al., 2017) and the `Tanks & Temples` (Knapitsch et al., 2017) datasets, have been released, which stand out due to their high resolution indoor and outdoor images and accurate ground-truth point clouds acquired from a laser scanner. Although these MVS benchmarks contain high-resolution images and accurate ground truth data obtained from a laser scanner, the setup is not designed for SIDE methods. Usually, a scene is scanned from multiple aligned laser scans and images are acquired in a sequential matter. The scans can be

Table D.2: Comparison of existing datasets related to SIDE evaluation with respect to different dataset characteristics. Interval distinguishes between still image acquisition (still) and continuous image acquisition (cont). Density specifies the completeness of provided depth maps. Higher resolutions are specified in brackets, if available in the datasets

Benchmark	**Setting**	**Sensor**	**Scenes**	**Images**	**Inverval**	**Range** (in m)	**Density**	**Resolution** (in Mpx)
`MegaDepth` (Li and Snavely, 2018)	Outdoor	RGB	196	130K	Still	relative	dense	1.9
`DIW` (Chen et al., 2016)	Various	RGB	—	470K	Still	relative	2 points	0.15
`ReDWeb` (Xian et al., 2018)	Various	RGB	—	3.6K	Still	relative	dense	0.19
`SceneNet RGB-D` (McCormac et al., 2017)	Indoor	Synthetic	57	5M	Cont.	1–5 m	dense	0.08
`SUNCG` (Song et al., 2017)	Indoor	Synthetic	45k	130k	Cont.	1–8 m	dense	0.31
`360-D` (Zioulis et al., 2018)	Indoor	Various	—	22k	Still	1–10 m	dense	0.13
`NYU-v2` (Silberman et al., 2012)	Indoor	RGB-D	464	654	Still	1–10 m	gaps	0.31
`Matterport3D` (Chang et al., 2017)	Indoor	RGB-D	90	200k	Cont.	1–10 m	gaps	0.8
`ScanNet` (Dai et al., 2017)	Indoor	RGB-D	707	2.5M	Cont.	0.4–3.5 m	gaps	0.31
`2D-3D-S` (Armeni et al., 2017)	Indoor	RGB-D	6	25k	Cont	1–10 m	gaps	1.3
`ETH3D` (Schöps et al., 2017)	Various	LiDAR	25	898	Cont.	1–20 m	gaps	0.4 (24)
`Tanks & Temples` (Knapitsch et al., 2017)	Various	LiDAR	14	150k	Cont.	1–20 m	dense	2
`Kitti` (Geiger et al., 2012)	Street	LiDAR	—	697	Cont.	1–80 m	sparse	0.5
`Strecha` (Strecha et al., 2008)	Outdoor	LiDAR	6	30	Still	1–10 m	dense	6
`Make3D` (Saxena et al., 2009)	Outdoor	LiDAR	—	534	Still	1–80 m	sparse	0.017
`LIVE Color+3D Database` (Su et al., 2017)	Outdoor	LiDAR	—	98	Still	2–100 m	dense	2.07
IBims-1 (Koch et al., 2018)	Indoor	LiDAR	70	100	Still	0.3–25 m	dense	0.31 (1.5)

used to generate depth maps aligned with the captured RGB images, but, however, it cannot be guaranteed that corresponding depth maps are dense. Occlusions in the images result in gaps in the depth maps especially at object boundaries which are, however, a key aspect of our metrics. Despite the possibility of acquiring a large number of image pairs, they mostly comprise only a limited scene variety and are highly redundant due high visual overlap. Currently, SIDE methods are tested on mainly three different datasets. `Make3D` (Saxena et al., 2009), as one example, contains 534 outdoor images and aligned depth maps acquired from a custom-built 3D scanner, but suffers from a very low resolution of the depth maps and a rather limited scene variety. The `Kitti` dataset (Geiger et al., 2012) contains street scenes captured out of a moving car. The dataset contains RGB images together with depth maps from a Velodyne laser scanner. However, depth maps are only provided in a very low resolution which furthermore suffer from irregularly and sparsely spaced points.

The most frequently used dataset for training and evaluating SIDE in indoor scenarios is the `NYU depth v2` (Silberman et al., 2012) dataset containing 464 indoor scenes with aligned RGB and depth images from video sequences obtained from a Microsoft Kinect v1 sensor. A subset of this dataset is mostly used for training deep networks, while another 654 image and depth pairs serve for evaluation. This large number of image pairs and the various indoor scenarios facilitated the fast progress of SIDE methods. However, active RGB-D sensors, like the Kinect, suffer from a short operational range, occlusions, gaps, and erroneous specular surfaces. The recently released `Matterport3D` (Chang et al., 2017), `ScanNet` (Dai et al., 2017), and `2D-3D-S` (Armeni et al., 2017) datasets provide even larger amounts of indoor scenes collected from RGB-D cameras, such as the Matterport Camera or the Structure sensor (Occipital, 2016). These datasets are valuable additions to the `NYU-v2` dataset but also suffer from the same weaknesses, as the used sensors have a similar design to the Kinect v1 sensor.

Recently, RGB-D datasets have been published using solely RGB images, such as `DIW` (Chen et al., 2016), `MegaDepth` (Li and Snavely, 2018), and `ReDWeb` (Xian et al., 2018). These datasets provide depths maps generated from stereo images utilizing freely available large-scale data platforms (*e.g.*, Flickr). They offer a huge variety of different scenes containing both indoor and outdoor scenes and can be easily computed using established MVS methods. However, the scale is unknown and the provided depth maps are therefore only relatively scaled, which only allows for ordinal depth estimation. Nevertheless, first investigations on training deep networks on these images reveal better generalization capabilities, but, however, they are ineligible when a metric scale is needed.

The LiDAR-based `LIVE Color+3D Database` (Su et al., 2017) offers highly-accurate registered RGB-D image pairs for 98 outdoor scenes similar to `Make3D`, but with an increased resolution and dense depth maps. The large range of scene depths and the high quality of the depth maps allow for detailed investigations of SIDE methods in outdoor scenarios, however, the scene variety is rather limited.

With the appearance of synthetic datasets, such as `SceneNet RGB-D` (McCormac et al., 2017), `SUNCG` (Song et al., 2017), and `360-D` (Zioulis et al., 2018), first attempts were made to train deep models with rendered RGB-D image pairs of this multitude of synthetically generated indoor scenes. However, the rendered RGB images are still far from realistic shots and are therefore not suited for testing the applicability of SIDE methods in real world environments.

D.3 NOVEL EVALUATION METRICS FOR DEPTH ESTIMATION

This section describes established metrics and our new proposed ones allowing for a more detailed analysis.

D.3.1 *Commonly Used Error Metrics*

Established error metrics consider global statistics between a predicted depth map $\boldsymbol{Y}$ and its ground truth depth image $\boldsymbol{Y}^*$ with T depth pixels. Beside visual inspections of depth maps or projected 3D point clouds, the following error metrics are exclusively used in all relevant recent publications (Eigen and Fergus, 2015; Eigen et al., 2014; Laina et al., 2016; Li et al., 2017; Xu et al., 2017):

Absolute relative difference: $\text{rel}\,(\boldsymbol{Y},\boldsymbol{Y}^*) = \frac{1}{T}\sum_{i,j}\left|y_{i,j} - y^*_{i,j}\right| / y^*_{i,j}$

Squared relative difference: $\text{srel}\,(\boldsymbol{Y},\boldsymbol{Y}^*) = \frac{1}{T}\sum_{i,j}\left|y_{i,j} - y^*_{i,j}\right|^2 / y^*_{i,j}$

RMS (linear): $\text{RMS}\,(\boldsymbol{Y},\boldsymbol{Y}^*) = \sqrt{\frac{1}{T}\sum_{i,j}\left|y_{i,j} - y^*_{i,j}\right|^2}$

RMS (log): $\log\,(\boldsymbol{Y},\boldsymbol{Y}^*) = \sqrt{\frac{1}{T}\sum_{i,j}\left|\log y_{i,j} - \log y^*_{i,j}\right|^2}$

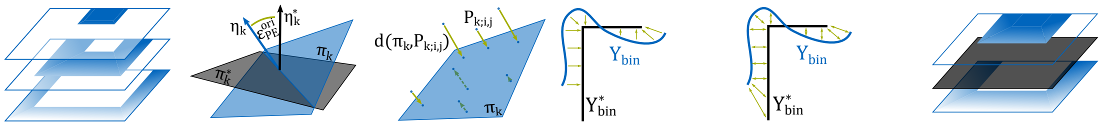
Threshold: percentage of $\boldsymbol{Y}$ such that $\max(\frac{y_i}{y_i^*}, \frac{y_i^*}{y_i}) = \sigma < thr$

The absolute relative difference error measures the relative per-pixel error linear to the absolute distance. In other words, an error of 0.1 m at a depth of 1 m is penalized equally to an error of 1 m at a depth of 10 m. An alternative with a squared influence of the relative per-pixel error is given by the squared relative difference. In contrast, the RMS error equally penalizes an error of 0.1 m at both depths. The threshold error on the other hand considers per-pixel proportions rather than per-pixel differences and measures the ratio of pixels, for which the relative difference between prediction and ground truth depths is below a threshold (thresholds are usually set to 1.25, 1.25^2, and 1.25^3).

Even though these statistics are good indicators for the general quality of predicted depth maps, they could be delusive. Particularly, the standard metrics are not able to directly assess the planarity of planar surfaces or the correctness of estimated plane orientations. Furthermore, it is of high relevance that depth discontinuities are precisely located, which is not reflected by the standard metrics. A general weakness of most current state-of-the-art SIDE methods is that the outputs tend to have spatially distorted or blurry object edges. While these local structures only affect a rather small part of the entire image, missing or blurry depth discontinuities have only a minor effect on the global error metrics, impeding a fair comparison of different methods.

D.3.2 *Proposed Error Metrics*

In order to allow for a more meaningful analysis of predicted depth maps and a more complete comparison of different algorithms, we present a set of new quality measures that specify on different characteristics of depth maps which are crucial for



(a) ε_{DRA} (b) ε_{PE}^{plan} (c) ε_{PE}^{orie} (d) ε_{DBE}^{acc} (e) ε_{DBE}^{comp} (f) $\varepsilon_{DDE}^{0,-,+}$

Figure D.2: Visualizations of our proposed error metrics. The *distance-related assessment* (a) applies standard metrics for different depth range intervals. The flatness and orientations of predicted planar regions can be evaluated with our *planarity errors* (b and c). The location accuracy and completeness of depth discontinuities is rated by the *depth boundary errors* (d and e), while the consistency of depth predictions with respect to a virtual depth plane can be assessed with our *directed depth errors* (f)

many applications. These are meant to be used in addition to the traditional error metrics introduced in Section D.3.1. Visual illustrations of our metrics explained below are depicted in Figure D.2. When talking about depth maps, the following questions arise that should be addressed by our new metrics:

- How is the quality of predicted depth maps for different absolute scene depths?
- Can planar surfaces be reconstructed correctly?
- Can all depth discontinuities be represented? How accurately are they localized?
- Are depth estimates consistent over the entire image area?

D.3.2.1 *Distance-Related Assessment*

Established global statistics are calculated over the full range of depth comprised by the image and therefore do not consider different accuracies for specific absolute scene ranges. Hence, applying the standard metrics for specific range intervals by discretizing existing depth ranges into discrete bins (*e.g.*, one-meter depth slices) allows investigating the performance of predicted depths for close and far ranged objects independently.

D.3.2.2 *Planarity Error (PE)*

Man-made objects, in particular, can often be characterized by planar structures like walls, floors, ceilings, openings, and diverse types of furniture. However, global statistics do not directly give information about the shape correctness of objects within the scene. Predicting depths for planar objects is challenging for many reasons. Primarily, these objects tend to lack texture and only differ by smooth color gradients in the image, from which it is hard to estimate the correct orientation of a 3D plane with three-degrees-of-freedom. In the presence of textured planar surfaces, it is even more challenging for a SIDE approach to distinguish between a real depth discontinuity and a textured planar surface, *e.g.*, a painting on a wall. As most methods are trained on large indoor scenes, like NYU-v2, a correct representation of planar structures is an important task for SIDE, but can hardly be evaluated using established standard metrics. For this reason, we propose to use a set of annotated images defining various planar surfaces (walls, table tops and floors) and evaluate the flatness and orientation of predicted 3D planes $\boldsymbol{\pi}_k = (\boldsymbol{\eta}_k, o_k)$ compared to ground truth 3D planes $\boldsymbol{\pi}_k^* = \left(\boldsymbol{\eta}_k^*, o_k^*\right)$. Each plane is specified by a normal vector $\boldsymbol{\eta}$ and an offset to the origin o. In detail, a masked depth map $\boldsymbol{Y}_k$ of a particular planar surface and an intrinsic matrix is used together in order to project the masked depth map to 3D points $\boldsymbol{P}_{k;i,j}$, where 3D planes $\boldsymbol{\pi}_k$ are robustly fitted to both the ground truth and predicted 3D point clouds $\mathcal{P}_k^* = \left\{\boldsymbol{P}_{k;i,j}^*\right\}_{i,j}$ and $\mathcal{P}_k = \left\{\boldsymbol{P}_{k;i,j}\right\}_{i,j}$, respectively. The planarity error

$$\varepsilon_{\mathrm{PE}}^{\mathrm{plan}}\left(\boldsymbol{Y}_k\right) = \mathbb{V}\left[\sum_{\boldsymbol{P}_{k;i,j}\in\mathcal{P}_k} d\left(\boldsymbol{\pi}_k, \boldsymbol{P}_{k;i,j}\right)\right] \tag{D.1}$$

is then quantified by the standard deviation of the averaged distances d between the predicted 3D point cloud and its corresponding 3D plane estimate. The orientation error

$$\varepsilon_{\mathrm{PE}}^{\mathrm{orie}}\left(\boldsymbol{Y}_k, \boldsymbol{\pi}_k^*\right) = \operatorname{acos}\left(\boldsymbol{\eta}_k^\top \cdot \boldsymbol{\eta}_k^*\right) \tag{D.2}$$

is defined as the 3D angle difference between the normal vectors of predicted and ground truth 3D planes. Figures D.2b and D.2c illustrate the proposed planarity errors. Note that for each individual planar mask the predicted depth maps are median scaled w.r.t. the ground truth depth map. This eliminates scaling differences of compared methods, which would influence the planarity error by favoring underestimated depth predictions.

D.3.2.3 *Location Accuracy of Depth Boundaries (DBE)*

Beside planar surfaces, captured scenes, especially indoor scenes, cover a large variety of scene depths caused by any object in the scene. Depth discontinuities between two objects are represented as strong gradient changes in the depth maps. In this context, it is important to examine whether predicted depths maps are able to represent all relevant depth discontinuities in an accurate way or if they even create fictitious depth discontinuities confused by texture. An analysis of depth discontinuities can be best expressed by detecting and comparing edges in predicted and ground truth depth maps. In order to evaluate predicted depth maps, edges $\boldsymbol{Y}_{\mathrm{bin}}$ are extracted and compared to a set of ground truth edges $\boldsymbol{Y}_{\mathrm{bin}}^*$ via *truncated chamfer distance* of the binary edge images. Specifically, a *Euclidean distance transform* is applied to the ground truth edge image $\boldsymbol{E}^* = DT\left(\boldsymbol{Y}_{\mathrm{bin}}^*\right)$, while distances exceeding a given threshold θ are truncated to a maximum distance θ. We define the depth boundary errors (DBEs), comprised of an accuracy measure

$$\varepsilon_{\mathrm{DBE}}^{\mathrm{acc}}(\boldsymbol{Y}_{\mathrm{bin}}, \boldsymbol{Y}_{\mathrm{bin}}^*) = \frac{1}{\sum_i \sum_j y_{\mathrm{bin};i,j}} \sum_i \sum_j e_{i,j}^* \cdot y_{\mathrm{bin};i,j} \tag{D.3}$$

by multiplying the predicted binary edge map with the distance map and a subsequent accumulation of the pixel distances towards the ground truth edge. Since this measure does not consider any missing or dispensable edges in the predicted depth image, we also define a completeness error

$$\varepsilon_{\mathrm{DBE}}^{\mathrm{comp}}(\boldsymbol{Y}_{\mathrm{bin}}, \boldsymbol{Y}_{\mathrm{bin}}^*) = \frac{1}{\sum_i \sum_j y_{\mathrm{bin};i,j}^* + y_{\mathrm{bin};i,j}} \sum_i \sum_j e_{i,j}^* \cdot y_{\mathrm{bin};i,j} + e_{i,j} \cdot y_{\mathrm{bin};i,j}^* \tag{D.4}$$

by accumulating both ground truth and predicted edges multiplied with their corresponding distance maps of ground truth and predicted edges $\boldsymbol{E}^*$ and $\boldsymbol{E} =$

$DT\left(\boldsymbol{Y}_{\text{bin}}\right)$. Therefore, the completeness error penalizes both missing and extra edges in the predictions in an equal manner. A visual explanation of the DBEs are illustrated in Figures D.2d and D.2e.

D.3.2.4 *Directed Depth Error (DDE)*

For many applications, it is of high interest that depth images are consistent over the whole image area. Although the absolute depth error, the squared depth error and the RMS errors give information about the correctness between predicted and ground truth depths, they do not provide information if the predicted depth is estimated too short or too far. For this purpose, we define the directed depth errors (DDEs)

$$\varepsilon_{\text{DDE}}^{0}\left(\boldsymbol{Y},\boldsymbol{Y}^{*},\boldsymbol{\pi}^{*}\right)=\frac{\left|\left\{y_{i,j}|d_{\text{sgn}}(\boldsymbol{\pi}^{*},\boldsymbol{P}_{i,j})=0\wedge d_{\text{sgn}}(\boldsymbol{\pi}^{*},\boldsymbol{P}_{i,j}^{*})=0\right\}\right|}{T} \tag{D.5}$$

$$\varepsilon_{\text{DDE}}^{+}\left(\boldsymbol{Y},\boldsymbol{Y}^{*},\boldsymbol{\pi}^{*}\right)=\frac{\left|\left\{y_{i,j}|d_{\text{sgn}}(\boldsymbol{\pi}^{*},\boldsymbol{P}_{i,j})>0\wedge d_{\text{sgn}}(\boldsymbol{\pi}^{*},\boldsymbol{P}_{i,j}^{*})<0\right\}\right|}{T} \tag{D.6}$$

$$\varepsilon_{\text{DDE}}^{-}\left(\boldsymbol{Y},\boldsymbol{Y}^{*},\boldsymbol{\pi}^{*}\right)=\frac{\left|\left\{y_{i,j}|d_{\text{sgn}}(\boldsymbol{\pi}^{*},\boldsymbol{P}_{i,j})<0\wedge d_{\text{sgn}}(\boldsymbol{\pi}^{*},\boldsymbol{P}_{i,j}^{*})>0\right\}\right|}{T} \tag{D.7}$$

as the proportions of correct, too far and too close predicted depth pixels $\varepsilon_{\text{DDE}}^{0}$, $\varepsilon_{\text{DDE}}^{+}$ and $\varepsilon_{\text{DDE}}^{-}$. In practice, a reference depth plane $\boldsymbol{\pi}^{*}$ is defined at a certain distance (*e.g.*, at 3 m) orthogonal to the camera view and all predicted depths pixels which lie in front and behind this plane are masked and assessed according to their correctness using the reference depth maps.

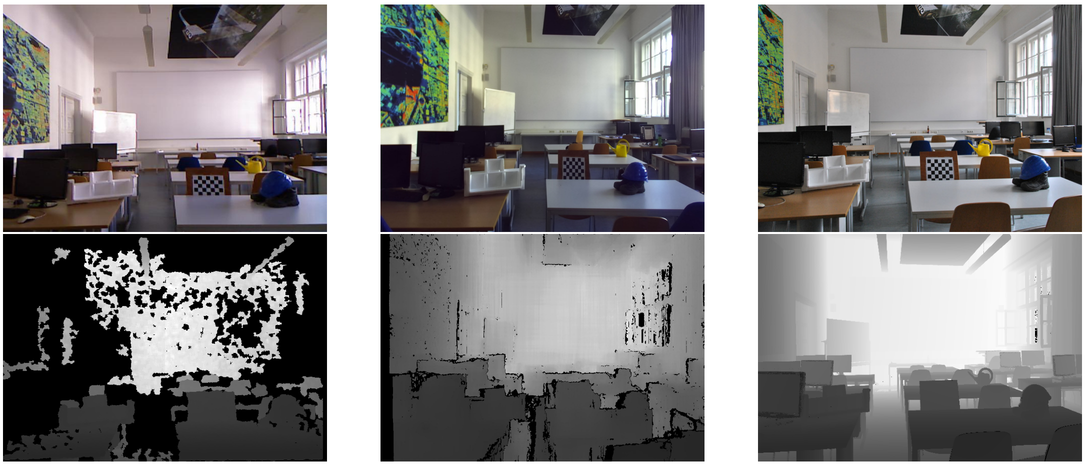
D.4 THE ibims-1 DATASET

As described in the previous sections, our proposed metrics require extended ground truth which is not yet available in standard datasets. Hence, we compiled a new dataset according to these specifications.

D.4.1 *Sensor Comparison*

For creating such a reference dataset, high-quality optical RGB images and depth maps had to be acquired. Practical considerations included the choice of suitable instruments for the acquisition of both parts. Furthermore, a protocol to calibrate both instruments, such that image and depth map align with each other, had to be developed.

For the creation of depth maps, we considered various sensors and instruments. Common mass market RGB-D products, such as MICROSOFT Kinect, not only allow for fast and convenient capturing of scenes, but also provide registered images and depth maps at the same time. However, the overall quality – especially in terms of resolution, accuracy and depth range – of the resulting depth maps and images turn out to be insufficient for the intended usage as reference data. Stereo rigs, such as the STEREOLABS ZED camera, outperform RGB-D products in several crucial areas, such as outdoor scenes. They are equally easy to use but also show deficits in certain areas. As the stereo reconstruction only produces results for textured



(a) Kinect v1 (b) ZED Cam (c) Laser scan + DSLR

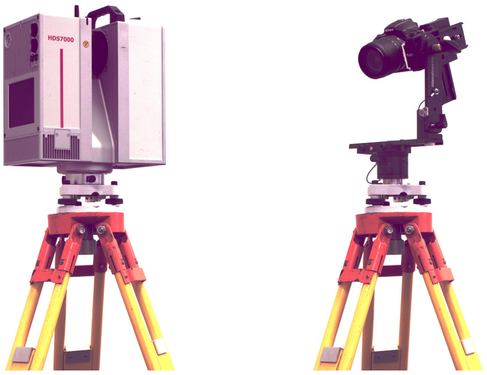
Figure D.3: Comparison of the depth map quality of different sensors

surfaces, the produced depth maps are often incomplete and suffer from noise. Precise geodetic instruments, such as tacheometers, laser trackers, or laser scanners, can provide highly accurate distance measurements. Among them, laser scanners excel in recording highly accurate dense point clouds in 360°. Figure D.3 shows a comparison of depth maps acquired from different sensors capturing the same scene. Beside differences in the image quality and intrinsics of the RGB images, the depth map generated with the Kinect v1 lack from numerous areal gaps as well as distorted object boundaries. The depth map provided by the ZED Cam on the other hand features almost dense depth estimates due to an internal interpolation of texture-less regions but, however, show the same deficits around object boundaries caused by the parallax effect. In addition, the overall noise level – especially for high distances – is relatively large compared to the Kinect v1. The high density and extremely high accuracy of the laser scanner allows for generating accurate, dense and detailed depth maps of superior quality compared to the other sensors.

As we want to generate highly accurate depth maps for high-resolution images, we finally chose a laser scanner as our sensor of choice. They do, however, fall short of expectations regarding provided imagery. As only a few instruments can capture RGB images at all, this is, in practice, most commonly done using an auxiliary camera. For this reason, we decided to design our own acquisition setup, as it is explained in the following section.

D.4.2 *Acquisition Process*

In order to record the ground truth for our dataset, we used a highly accurate Leica HDS7000 laser scanner, which stands out for high point cloud density and very low noise level. Dependent on the scene depth of the individual images in our dataset we varied the point spacing of the acquired scans to ensure at least one depth value for each pixel in a down-sampled version of the RGB image of 640 × 480 px. However, for most scenes we exceeded the required point density by a multiple in order to provide nearly dense depth maps in a higher resolution as well (1500 × 1000 px). As our laser scanner does not provide RGB images along with the point clouds, an additional camera was used in order to capture optical imagery. The usage of a



(a) Laserscanner (b) Camera

Figure D.4: Our hardware setup used for the acquisition of `IBims-1` with a laser scanner (a) and a DSLR camera (b) mounted on a survey tripod. A custom panoramic tripod is used in order to achieve a coincidence of the optical center of the camera and the origin of the laser scanner coordinate system to avoid occlusions in the resulting depth maps

reasonably high-quality camera sensor and lens allows for capturing images in high resolution with only slight distortions and a high stability regarding the intrinsic parameters. For our data acquisition, we chose two calibrated DSLR cameras: one Nikon D5500 digital single-lens reflex (DSLR) camera equipped with a Nikon AF-S Nikkor 18–105 mm lens, mechanically fixed to a focal length of 18 mm and a Nikon D3000 DSLR camera equipped with the same lens, mechanically fixed to focal lengths of 18 mm and 21 mm.

Using our sensor setup, synchronous acquisition of point clouds and RGB imagery is not possible. In order to acquire depth maps without parallax effects, the camera was mounted on a custom panoramic tripod head which allows to freely position the camera along all six degrees of freedom. An illustration of our setup is depicted in Figure D.4. This setup can be interchanged with the laser scanner, ensuring coincidence of the optical center of the camera and the origin of the laser scanner coordinate system after a prior calibration of the system. It is worth noting that every single RGB-D image pair of our dataset was obtained by an individual scan and image capture with the aforementioned strategy in order to achieve dense depth maps without gaps due to occlusions.

D.4.3 *Registration and Processing*

The acquired images were undistorted using the intrinsic camera parameters obtained from the calibration process. In order to register the camera towards the local coordinate system of the laser scanner, we manually selected a sufficient number of corresponding 2D and 3D points and estimated the camera pose using EPnP (Moreno-Noguer et al., 2007). This registration of the camera relative to the point cloud yielded only a minor translation, thanks to the pre-calibrated platform. Using this procedure, we determined the 6D pose of a virtual depth sensor which we use to derive a matching depth map from the 3D point cloud. In order to obtain a depth value for each pixel in the image, the images were sampled down to two different resolutions. We provide a high-quality version with a resolution of 1500 × 1000 px and a cropped `NYU-v2`-like version with a resolution of 640 × 480 px. After the pose estimation of the camera, 3D points were projected to the virtual sensor with the

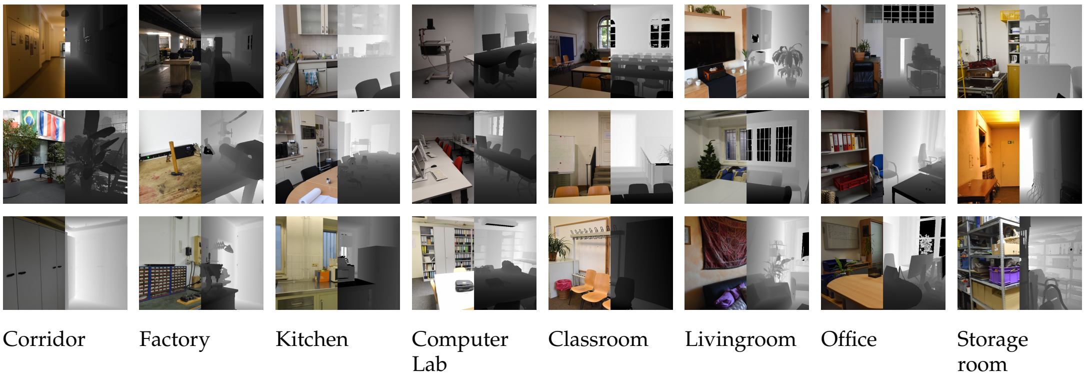


Figure D.5: Sample RGD-D image pairs of our `IBims-1` dataset covering different scenes. Illustrations are composed of the RGB image (left) and the corresponding depth map (right)

respective resolution. For each pixel, a depth value was calculated, representing the depth value of the 3D point with the shortest distance to the virtual sensor. It is worth highlighting that depth maps were derived from the 3D point cloud for both versions of the images separately. Hence, no down-sampling artifacts are introduced for the lower-resolution version of the depth maps.

D.4.4 *Registration Accuracy*

In order to present a high-quality RGB-D reference dataset, it is crucial that RGB images and depth images are aligned properly. The reprojection errors of the 2D-3D correspondences used for the camera pose estimations provide a first evidence of the registration accuracy of our dataset. For each of the 100 RGB-D image pairs of our dataset we manually selected 8-10 point correspondences. The mean reprojection error for all 2D-3D correspondences is 0.81 px with respect to the NYU-like resolution.

Since the reprojection error is only calculated on the basis of a single points, it is difficult to make a general statement about the overall registration accuracy. For this reason we also investigate the alignment on the basis of edges with the assumption that most depth discontinuities in a edge map correspond to intensity changes in the RGB image. We therefore compute dominant edges in depth maps and RGB images respectively using a Sobel operator and compare them using a *directed chamfer distance*. Note, that we only consider edges in the RGB image which are located in the local neighborhood of extracted depth edges (*e.g.*, within 10 px) for excluding gradients caused by texture or illumination changes. In average, around 450 edge pixels were extracted and compared for each RGB-D image pair. The averaged chamfer distance considering all images is 1.20 px. Since some depth edges do not correspond to intensity changes in the RGB image and vice versa, this metric serves only as a vague proof of the registration accuracy, but, however, yields an overall quality measure showing how accurate our RGB and depth maps are aligned.

D.4.5 *Contents*

Following the described procedure in Sections D.4.2 and D.4.3, we compiled a dataset, which we henceforth refer to as the *independent benchmark images and matched scans v1 (`IBims-1`)* dataset. The dataset is mainly composed of reference data for the

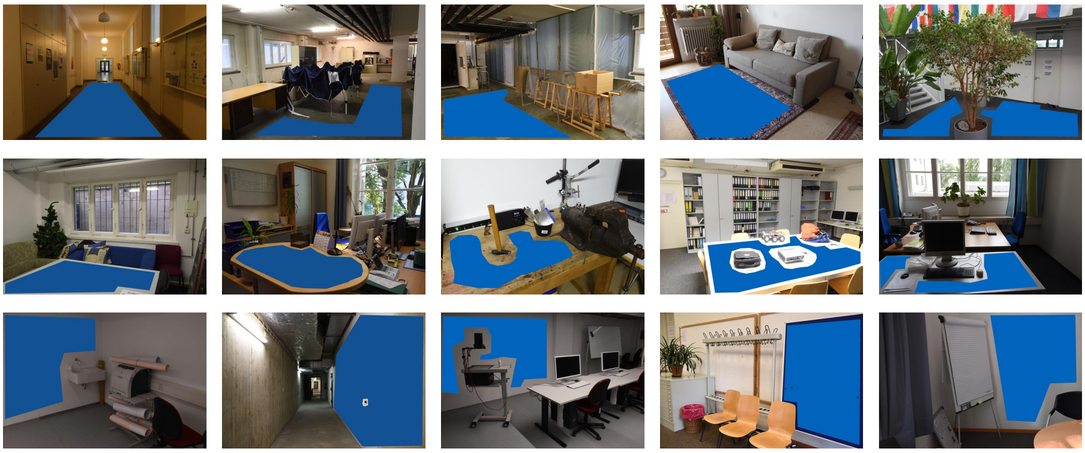


Figure D.6: Annotation samples showing provided plane masks (🟦) for *floors* (top), *table tops* (mid) and *walls* (bottom)

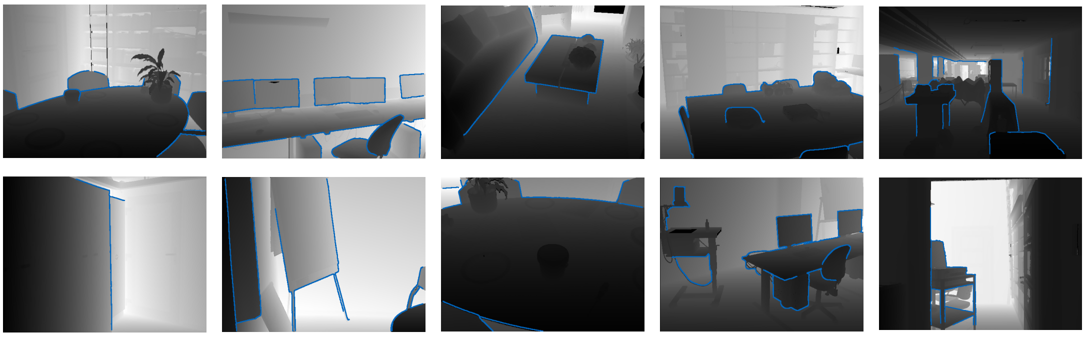
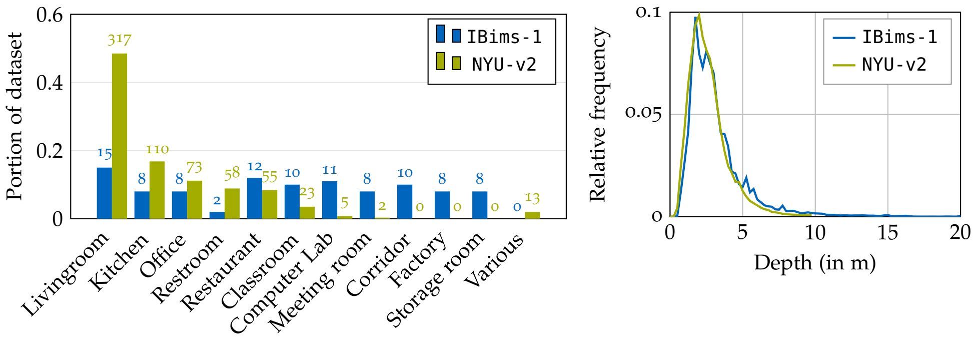


Figure D.7: Annotation samples showing provided edge masks (—) for distinct depth discontinuities

direct evaluation of depth maps, as produced by SIDE methods. This main part of the dataset contains 100 RGB-D image pairs in total. As described in the previous sections, pairs of images and depth maps were acquired and are provided in two different versions, namely a high-quality version and a `NYU-v2`-like version. Example pairs of images and matching depth maps from `IBims-1` are shown in Figure D.5.

Additionally, several manually created masks are provided. Unreliable or invalid pixels in the depth map are labeled by two different sets of binary masks. One of which flags transparent objects, mainly windows, which could be assigned with an ambiguous depth. While the laser scanner captured points behind those objects, it may be intended to obtain the distance of the transparent object for certain applications. The other mask for invalid pixels indicates faulty values in the 3D point cloud. Those mainly originate from scanner-related errors, such as reflecting surfaces, as well as regions out of range. Three further sets of masks label planar surfaces of three different types, *i.e.*, tables, floors, and walls. Each instance is contained in a separate mask. Examples for planar masks are shown in Figure D.6, while statistics of the plane annotations are listed in Table D.3. It is worth mentioning that the plane masks do not coincide with the object boundaries, but rather keeping a buffer area of several pixel towards the object boundaries. The reason for this is that these masks are used for investigating the capability of predicting planar regions. Object boundaries often cause distortions in the predicted depth map which is target of our DBE but should not influence the PE.



(a) Distribution of samples for each scene type. Absolute numbers are given above

(b) Distribution of depth values

Figure D.8: IBims-1 dataset statistics compared to the NYU-v2 dataset. Scene variety (a) and distribution of depth values (b)

Table D.3: Statistics of plane annotations in NYU-v2 and IBims-1. Number of instances (Inst.) of a specific plane type (Type) occurred in the dataset (Images), the average size of each object mask (Avg. Size), and accuracy of fitted 3D reference planes. The larger deviations in planes fitted to the images of NYU-v2 can be attributed to the inaccurate and noisy measurements of the utilized RGB-D sensor. A reliable assessment of planarity errors based on NYU-v2 is therefore only possible to a limited extent

Dataset	**Type**	**Images**	**Inst.**	**Avg. Size**	**Mean Dev.**	**Std. Dev.**
				(in px)	(in mm)	(in mm)
NYU-v2	Floor	132	132	29389	17.42	14.25
NYU-v2	Table	44	44	27989	17.80	17.19
NYU-v2	Wall	168	168	34975	28.17	22.66
IBims-1	Floor	47	51	22813	1.57	1.85
IBims-1	Table	46	54	15704	1.18	1.50
IBims-1	Wall	82	140	46744	1.79	2.38

In order to allow for evaluation following the proposed DBE metric, we provide distinct edges for all images. Location accuracy and sharp edges are of high importance for generating a set of ground truth depth transitions which cannot be guaranteed by existing datasets acquired from RGB-D sensors. Ground truth edges are extracted from our dataset by applying a Canny edge detector on the depth maps. Since the scenes in our dataset exhibit various depth ranges, the selection of dominant edges vary with the depth range of the individual RGB-D image pairs. For this reason, we only consider distinct depth edges that exceed a depth change of at least 15% of the overall depth range in the individual image. Figure D.7 shows examples of the ground truth edges for different scenes from IBims-1.

Additionally, we provide an *auxiliary dataset* which consists of four parts: (1) Four outdoor RGB-D image pairs, containing vegetation, building, cars and larger ranges than indoor scenes. (2) Special cases which are expected to mislead SIDE methods. These show 85 RGB images of printed samples from the NYU-v2 and the Pattern dataset (Asuni and Giachetti, 2014) hung on a wall. Those could potentially give valuable insights, as they reveal what kind of image features SIDE methods exploit. No depth maps are provided for those images, as the region of interest is supposed to be approximately planar and depth estimates are, thus, easy to assess qualitatively. (3) 56 geometrical and radiometrical augmentations for each image of our core dataset to test the robustness of SIDE methods. (4) Up to three additional handheld images for many RGB-D image pairs of our core dataset with viewpoint changes towards the reference images which allows to validate MVS algorithms with high-quality ground truth depth maps.

D.4.6 *Comparison of `IBims-1` and `NYU-v2`*

So far, the `NYU-v2` dataset is still the most comprehensive and accurate indoor dataset for training data-demanding deep learning methods. Since this dataset has most commonly been used for training the considered SIDE methods, `IBims-1` is designed to contain similar scenarios. Our acquired scenarios include various indoor settings, such as offices, lecture, and living rooms, computer labs, as well as more challenging ones, such as long corridors, potted plants and factory rooms. A comparison regarding the scene variety between `NYU-v2` and `IBims-1` can be seen in Figure D.8a. Furthermore, `IBims-1` features statistics comparable to `NYU-v2`, such as the distribution of depth values, shown in Figure D.8b, and a comparable field of view.

However, comparing the depth map quality of both datasets, raw depth maps of `NYU-v2` show a large amount of missing and erroneous depth values due to parallax effect, limited range (up to 10 m), and relatively high noise level, as this is already investigated in Zennaro et al. (2015). In total, 36% of all depth values in the raw depth maps in `NYU-v2` are missing. Missing values were interpolated using the colorization method of Levin et al. (2004), which results in erroneous measurements and artefacts, such as flying pixels. Moreover, transparent and specular surfaces are not masked in `NYU-v2` resulting in distorted depth values in the dataset. Due to the high point density and accuracy of the scans in `IBims-1`, no interpolation is needed for NYU-like resolution in `IBims-1`, resulting in dense and valid depth values. Figure D.9 visualizes the quality of `NYU-v2` and compares it towards `IBims-1`. In contrast to the imprecise and incomplete depth maps in `NYU-v2`, the seamless depth maps in `IBims-1` facilitate the extraction of accurate and complete depth discontinuities. The high accuracy of these depth maps also guarantees the extraction of accurate 3D planes in the range of a few millimeters, while deviations of more than 2 cm were noted when making use of the `NYU-v2` dataset[2], as shown in Table D.3. Although in principle `NYU-v2` allows to generally assess the planarity of depth predictions, the results would not satisfy our accuracy requirements for providing reliable conclusions about the performance of the methods.

D.5 EVALUATION OF SIDE METHODS

In this section, we evaluate the quality of existing SIDE methods using both established and proposed metrics for our reference test dataset, as well as for the commonly used `NYU-v2` dataset. As outlined in our review of the state-of-the-art in Section D.2.1, a multitude of different deep learning-based SIDE approaches have been developed over the past few years. Naturally, not all of them can be subjected to detailed investigation. We chose an exemplary subset of the available approaches

[2] Plane annotations for `NYU-v2` were also made available on our webpage

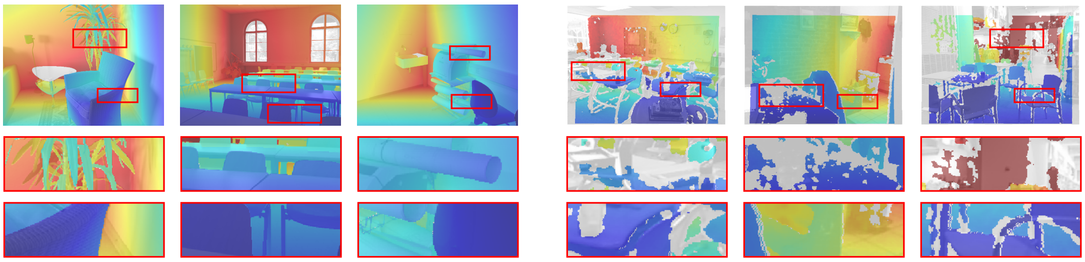


Figure D.9: Visualization of registration accuracy and depth completeness of IBims-1 (left) and NYU-v2 (right). Overlay of greyscale RGB images and colored depth maps for various samples (invalid or missing depth values are depicted in grey). Top: full image. Middle and bottom row: detailed views

for the evaluation experiments, that either represent a milestone in the development of SIDE, or constitute current approaches that address specific aspects of particular interest, which have been identified in the development of our geometrically interpretable error metrics. In order to allow a fair comparison, only methods that were trained on indoor scenes, namely the NYU-v2 dataset, were examined. This preliminary selection was further narrowed down to accessible methods for which we received either source code or predictions for our dataset, which ultimately led to a comparison of eight methods, namely those proposed by Eigen et al. (2014), Eigen and Fergus (2015), Liu et al. (2015), Laina et al. (2016), Li et al. (2017), PlaneNet (Liu et al., 2018) and Sharpnet (Ramamonjisoa and Lepetit, 2019). Since all of these methods were solely trained on the NYU-v2 dataset, differences in the results are expected to arise from the developed methodology rather than the training data. For the evaluation using our dataset, only valid depth areas were considered by applying the provided corresponding masks to the raw depth maps. The quantitative results on both datasets with all error metrics are listed in Table D.4. A detailed analysis of the individual metrics is given in the following sections. Although a runtime evaluation would be of great interest for many application fields, the realization of a revealing comparison was infeasible, since runtime is highly dependent on implementation details and utilized frameworks, which varied between the examined methods. Furthermore, the lack of available source code for some methods prevents a comparison on the same hardware setup.

D.5.1 *Established Global Error Metrics*

The results of evaluation using commonly used global metrics on IBims-1 and NYU-v2 listed in Table D.4 by computing the statistical error metrics on the complete images. This is the standard evaluation procedure in all recent publications. The revealed lower overall scores for our dataset are expected since the dataset is previously unseen by these methods. As the methods are trained to predict depths in the range of the NYU-v2 dataset (*i.e.*, 1–10 m), they are not able to estimate depths beyond this range which are also encompassed in our dataset. This highly affects the RMS error, which turned out to be almost three times as large as in NYU-v2. Moreover, our dataset uncovers different generalization capabilities of the methods, as the order of the rankings has changed between NYU-v2 and IBims-1. However, the ranking according to different standard metrics did not change substantially among the methods, as most metrics are highly correlated to each other. This proves our claim

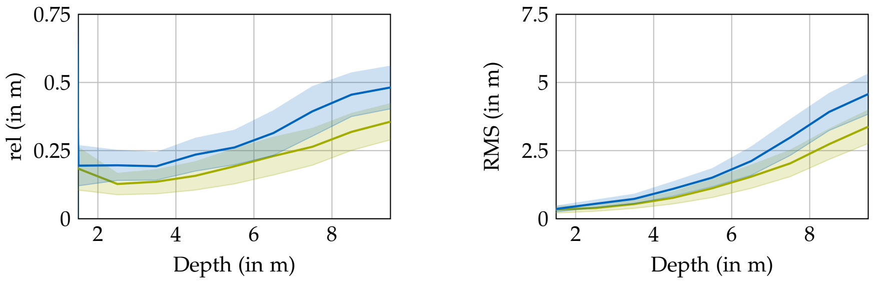


Figure D.10: Distance-related global errors (left: relative error and right: RMS) for the shared depth range of NYU-v2 (mean: —,±0.5 std: 🟩) and IBims-1 (mean: —, ±0.5 std: 🟦) using the method of Li et al. (2017)

for further sophisticated evaluation criteria, which are analyzed in the following sections.

D.5.2 *Distance-related Assessment*

In order to get a better understanding of these results, we evaluated the considered methods on specific range intervals, which we set to 1 m in our experiments. Figure D.10 shows the error band of the relative and RMS errors of the method proposed by Li et al. (2017) applied to both datasets. The result clearly shows a comparable trend on both datasets for the shared depth range. This proves our first assumption, that the overall lower scores originate from the huge differences at depth values beyond the 10 m depth range. On the other hand, the results reveal the generalization capabilities of the networks, which achieve similar results on images from camera with slightly different intrinsics and image quality, as well as for unseen scenarios. A comparison of the performance on a larger depth range for different methods and error metrics, as shown in Figure D.11, clearly shows a trend of decreasing accuracy over an increasing distance. Best results are achieved in a very close range up to 4 m, which corresponds to the maximum of the depth distribution of the NYU-v2 dataset on which the methods were trained (*cf.* Figure D.8b). Training on this highly imbalanced dataset with current state-of-the-art methods results in predicting depths below a RMS error of 1 m for distances up to 5 m, but linearly increases together with the scene depth for distances greater than 5 m. While most methods do not differ significantly in predicting depth values at various ranges and correspond to the ranking in Table D.4, the method of Ramamonjisoa and Lepetit (2019) performs notably better at larger distances. However, since the results exhibit a deficiency in close ranges up to 2–3 m, which corresponds to the peak of the depth distribution in IBims-1, errors in this range decisively contribute to the global errors listed in Table D.4. Such enhanced distinction and assessment of the performance would not have been feasible by solely relying on established global error metrics.

D.5.3 *Planarity*

To investigate the quality of reconstructed planar structures, we evaluated the different methods with the planarity and orientation errors $\varepsilon_{\mathrm{PE}}^{\mathrm{plan}}$ and $\varepsilon_{\mathrm{PE}}^{\mathrm{orie}}$, respectively, as defined in Section D.3.2.2, for different planar objects. In particular, we distinguished

Table D.4: Quantitative results for standard metrics on NYU-v2 and standard metrics, proposed PE, DBE, and DDE metrics on IBims-1 applying different SIDE methods (**best**, **second best**). Higher the better for ↑ and lower the better for ↓

Method	**Dataset**	**Standard Metrics** ($\sigma_i = 1.25^i$)						**PE (cm/°)**		**DBE (px)**		**DDE (%) for** d **=3 m**		
		rel ↓	$\log_{10}$ ↓	RMS ↓	σ_1 ↑	σ_2 ↑	σ_3 ↑	$\varepsilon_{\text{PE}}^{\text{plan}}$ ↓	$\varepsilon_{\text{PE}}^{\text{orie}}$ ↓	$\varepsilon_{\text{DBE}}^{\text{acc}}$ ↓	$\varepsilon_{\text{DBE}}^{\text{comp}}$ ↓	$\varepsilon_{\text{DDE}}^{0}$ ↑	$\varepsilon_{\text{DDE}}^{-}$ ↓	$\varepsilon_{\text{DDE}}^{+}$ ↓
Eigen et al., 2014	NYU-v2	0.22	0.09	0.76	0.61	0.89	0.97	—	—	—	—	—	—	—
Eigen and Fergus, 2015 (AlexNet)	NYU-v2	0.19	0.08	0.67	0.69	0.91	**0.98**	—	—	—	—	—	—	—
Eigen and Fergus, 2015 (VGG)	NYU-v2	0.16	**0.07**	0.58	0.75	0.95	**0.99**	—	—	—	—	—	—	—
Laina et al., 2016	NYU-v2	**0.14**	**0.06**	**0.51**	**0.82**	0.95	**0.99**	—	—	—	—	—	—	—
Liu et al., 2015	NYU-v2	0.21	0.09	0.68	0.66	0.91	**0.98**	—	—	—	—	—	—	—
Li et al., 2017	NYU-v2	**0.15**	**0.06**	0.53	0.79	**0.96**	**0.99**	—	—	—	—	—	—	—
Liu et al., 2018	NYU-v2	**0.14**	**0.06**	**0.51**	0.81	**0.96**	**0.99**	—	—	—	—	—	—	—
Ramamonjisoa and Lepetit, 2019	NYU-v2	**0.14**	**0.06**	**0.46**	**0.84**	**0.97**	**0.99**	—	—	—	—	—	—	—
Eigen et al., 2014	IBims-1	0.32	0.17	1.55	0.36	0.65	0.84	7.70	24.91	9.97	9.99	70.37	27.42	2.22
Eigen and Fergus, 2015 (AlexNet)	IBims-1	0.30	0.15	1.38	0.40	0.73	0.88	7.52	21.50	4.66	8.68	77.48	18.93	3.59
Eigen and Fergus, 2015 (VGG)	IBims-1	**0.25**	**0.13**	1.26	0.47	0.78	**0.93**	**5.97**	**17.65**	4.05	8.01	79.88	18.72	**1.41**
Laina et al., 2016	IBims-1	0.26	**0.13**	1.20	0.50	0.78	0.91	**6.46**	19.13	6.19	9.17	81.02	17.01	1.97
Liu et al., 2015	IBims-1	0.30	**0.13**	1.26	0.48	0.78	0.91	8.45	28.69	**2.42**	**7.11**	79.70	14.16	6.14
Li et al., 2017	IBims-1	**0.22**	**0.11**	**1.09**	**0.58**	**0.85**	**0.94**	7.82	22.20	3.90	8.17	**83.71**	**13.20**	3.09
Liu et al., 2018	IBims-1	0.29	0.17	1.45	0.41	0.70	0.86	7.26	**17.24**	4.84	8.86	71.24	28.36	**0.40**
Ramamonjisoa and Lepetit, 2019	IBims-1	0.26	**0.11**	**1.07**	**0.59**	**0.84**	**0.94**	9.95	25.67	**3.52**	**7.61**	**84.03**	**9.48**	6.49

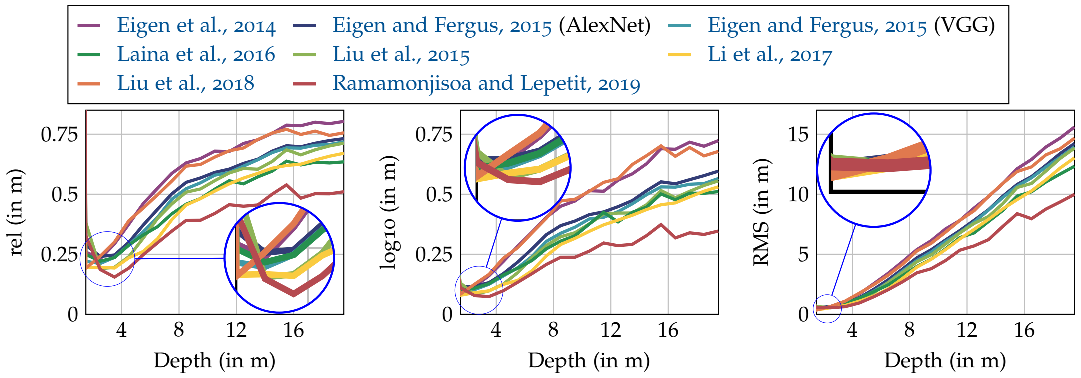


Figure D.11: Comparing distance-related global errors up to 20 m on `IBims-1` for the examined methods. From left to right: relative error, log10 error and RMS error

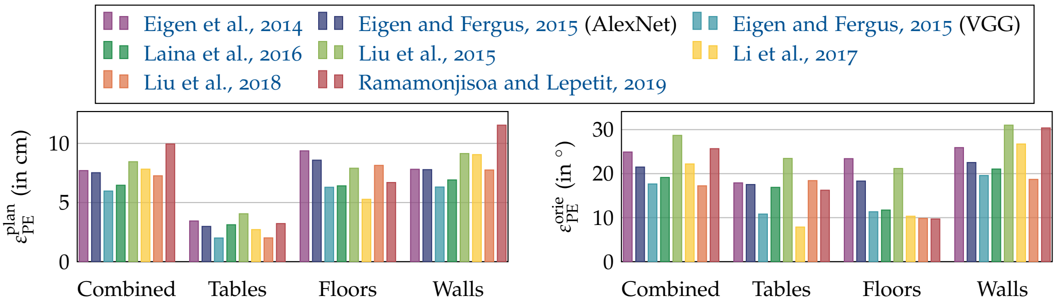
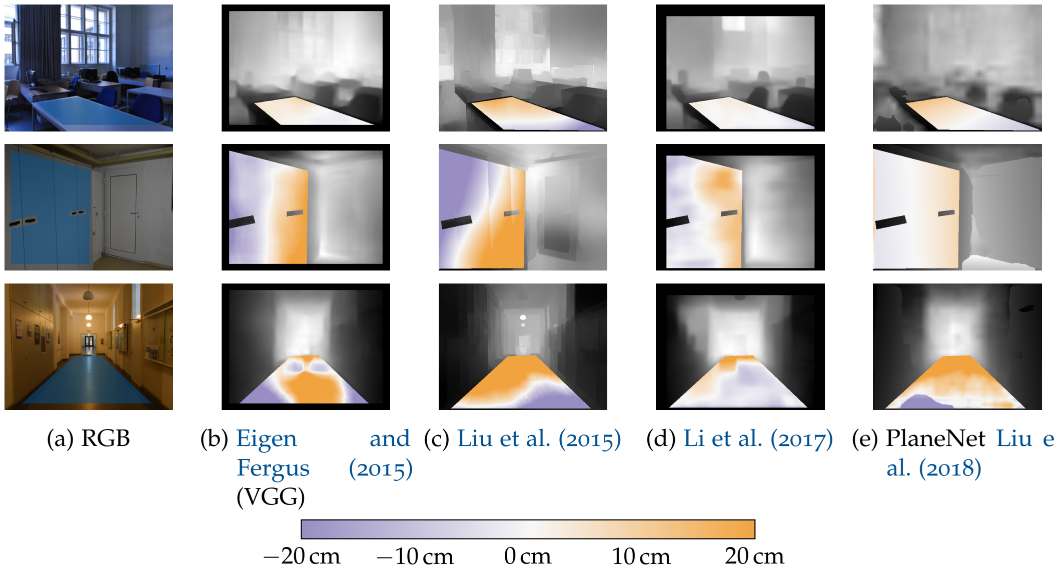


Figure D.12: Results for the planarity metrics $\varepsilon_{\mathrm{PE}}^{\mathrm{plan}}$ (left) and $\varepsilon_{\mathrm{PE}}^{\mathrm{orie}}$ (right) on `IBims-1` for individual plane types and a combination of all (Combined)

between horizontal and vertical planes and used masks from our dataset. Beside a combined error, including all planar labels, we separately computed the errors for the individual objects as well. Results for averaged errors among all types of planar regions are listed in Table D.4, while results for individual plane types are shown in Figure D.12. The results reveal different performances for individual classes, especially orientations of floors and table tops were predicted in a significantly higher accuracy, while the absolute orientation error for walls is surprisingly high. Considering the flatness of the predictions, tables can be reconstructed more reliable than floors or walls. Apart from the general performance of all methods, substantial differences between the considered methods can be determined. It is notable that the method of Li et al. (2017) achieved much better results in predicting orientations of horizontal planes but also performed rather bad on vertical surfaces. In contrast, orientation results for Liu et al. (2015) exhibit large errors for all types of planes. Reason for this could lie in problems of smooth depth transitions for adjacent superpixels representing flat, but textured or differently illuminated areas. This oversegmentation results in strong depth changes in planar regions. The method of Ramamonjisoa and Lepetit (2019) revealed large differences in the accuracy of the reconstruction of planar objects, notably for floors and walls. In striving at preserving accurate and sharp depth transitions, this network tends to be more sensitive to texture changes and high frequencies, yielding fragmented and falsely determined planes. The performance of PlaneNet (Liu et al., 2018), which focuses on the preservation of planar regions, strongly depends on a prior semantic segmentation of the input



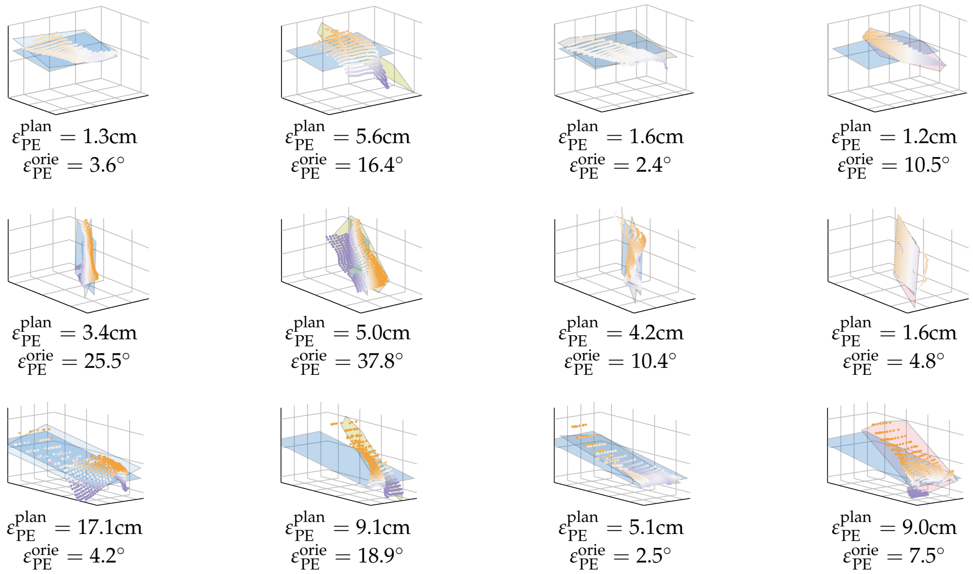
(a) RGB (b) Eigen and Fergus (2015) (VGG) (c) Liu et al. (2015) (d) Li et al. (2017) (e) PlaneNet Liu et al. (2018)

Figure D.13: Visual results after applying *planarity errors* (PEs) on different planar regions (top: table, middle: wall, bottom: floor). RGB with corresponding plane masks (🟦) (a). Predictions using different methodologies (b-e). Colors in the predictions correspond to orthogonal differences of projected depths towards the reference plane

image. For each detected planar region in the segmentation step, the method estimates reasonable 3D plane parameters, but, however, pixel-accurate segmentations of planar regions often fails, which results in imprecise and fragmented 3D planes. Visual results showing residuals of projected depth maps and ground truth 3D planes are depicted in Figure D.13, which reveals different depth map characteristics based on the used methodology. 3D illustrations, displaying projected 3D points, fitted 3D plane and ground truth 3D plane for the scenes in Figure D.13 are shown in Figure D.14. Despite the considerably lower accuracy of fitted ground truth 3D planes in `NYU-v2`, planarity errors can principally be determined in the same manner, although, as already outlined in Section D.4.6 inaccurate ground truth 3D planes limit the reliability of the derived results. The evaluations have shown that, similar to the global metrics, better overall results can be achieved, which is partly attributed to the slight domain shift between both dataset. However, similar to the results on `IBims-1`, a difference in the performance regarding the reconstruction of planar regions could be observed which results in a similar ranking of the investigated methods.

D.5.4 *Location Accuracy of Depth Boundaries*

The high quality of our reference dataset facilitates an accurate assessment of predicted depth discontinuities. As ground truth edges, we used the provided edge maps from our dataset and computed the accuracy and completeness errors $\varepsilon_{\text{DBE}}^{\text{acc}}$ and $\varepsilon_{\text{DBE}}^{\text{comp}}$, respectively, introduced in Section D.3.2.3. We set the distance threshold of the *truncated chamfer distance* to $\theta = 10\,\text{px}$, which also defines the upper bound of the accuracy and completeness errors. Quantitative results for all methods are listed in Table D.4. Comparing the accuracy error of all methods, Liu et al. (2015) and Ramamonjisoa and Lepetit (2019) achieved best results in preserving actual depth



$\varepsilon_{\text{PE}}^{\text{plan}} = 1.3\text{cm}$ $\varepsilon_{\text{PE}}^{\text{orie}} = 3.6°$

$\varepsilon_{\text{PE}}^{\text{plan}} = 5.6\text{cm}$ $\varepsilon_{\text{PE}}^{\text{orie}} = 16.4°$

$\varepsilon_{\text{PE}}^{\text{plan}} = 1.6\text{cm}$ $\varepsilon_{\text{PE}}^{\text{orie}} = 2.4°$

$\varepsilon_{\text{PE}}^{\text{plan}} = 1.2\text{cm}$ $\varepsilon_{\text{PE}}^{\text{orie}} = 10.5°$

$\varepsilon_{\text{PE}}^{\text{plan}} = 3.4\text{cm}$ $\varepsilon_{\text{PE}}^{\text{orie}} = 25.5°$

$\varepsilon_{\text{PE}}^{\text{plan}} = 5.0\text{cm}$ $\varepsilon_{\text{PE}}^{\text{orie}} = 37.8°$

$\varepsilon_{\text{PE}}^{\text{plan}} = 4.2\text{cm}$ $\varepsilon_{\text{PE}}^{\text{orie}} = 10.4°$

$\varepsilon_{\text{PE}}^{\text{plan}} = 1.6\text{cm}$ $\varepsilon_{\text{PE}}^{\text{orie}} = 4.8°$

$\varepsilon_{\text{PE}}^{\text{plan}} = 17.1\text{cm}$ $\varepsilon_{\text{PE}}^{\text{orie}} = 4.2°$

$\varepsilon_{\text{PE}}^{\text{plan}} = 9.1\text{cm}$ $\varepsilon_{\text{PE}}^{\text{orie}} = 18.9°$

$\varepsilon_{\text{PE}}^{\text{plan}} = 5.1\text{cm}$ $\varepsilon_{\text{PE}}^{\text{orie}} = 2.5°$

$\varepsilon_{\text{PE}}^{\text{plan}} = 9.0\text{cm}$ $\varepsilon_{\text{PE}}^{\text{orie}} = 7.5°$

(a) Eigen and Fergus (2015) (VGG)

(b) Liu et al. (2015)

(c) Li et al. (2017)

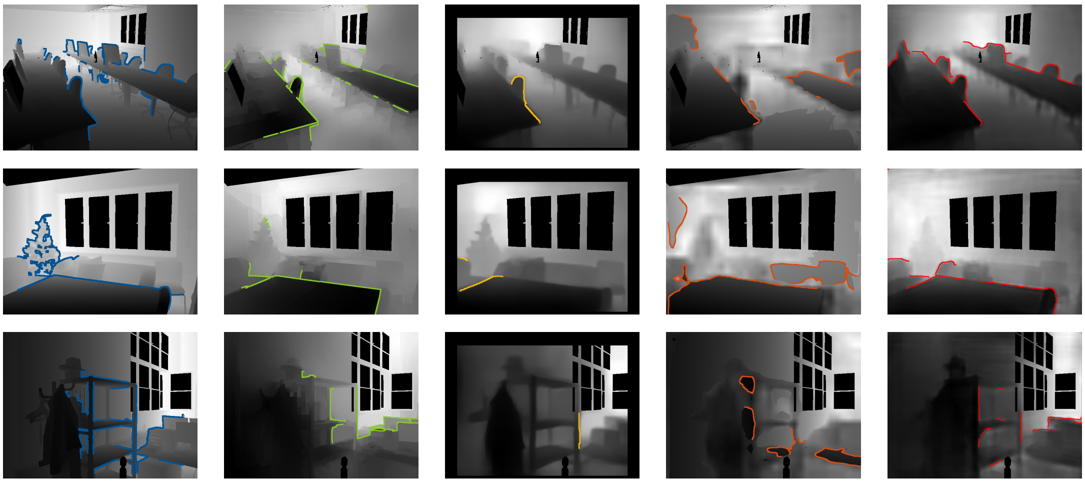
(d) PlaneNet (Liu et al., 2018)

Figure D.14: 3D visualizations of predicted 3D planes from Figure D.13. Ground truth 3D planes (🟦), projected 3D points from predictions and fitted 3D planes. Color coding of the 3D points is similar to definitions in Figure D.13

boundaries, while other methods tended to produce smooth edges, and thus failed to reconstruct precise and complete depth transitions. This smoothing property and the small output resolution of some methods also affected the completeness error, resulting in missing edges expressed by larger values for $\varepsilon_{\text{DBE}}^{\text{comp}}$. A comparison of depth boundaries from different methods can be seen in Figure D.15. Preserving sharp depth discontinuities is a main challenge using CNN-based methods, due to the intensive number of strided convolutions and spatial poolings, which reduce the output resolution, and, thus, local details of the image. However, methods that explicitly address this aspect have proven to enhance the reconstruction of object contours, which is also evident in the proposed DBE metrics.

D.5.5 *Directed Depth Error*

The DDE aims to identify predicted depth values which lie on the correct side of a predefined reference plane but also distinguishes between overestimated and underestimated predicted depths. This measure could be useful for applications, such as image refocusing and 3D cinematography. For the quantitative results listed in Table D.4 we defined a reference plane at 3 m distance and computed the proportions of correct $\varepsilon_{\text{DDE}}^{0}$, overestimated $\varepsilon_{\text{DDE}}^{+}$, and underestimated $\varepsilon_{\text{DDE}}^{-}$ depth values towards this plane according to the error definitions in Section D.3.2.4. A visual illustration of correctly and falsely predicted depths is depicted in Figure D.16 and a comparsion of different thresholds of d is shown in Figure D.17. The results show that, apart from the approach of Ramamonjisoa and Lepetit (2019), the methods tended to underestimate depth, although the amount of correctly estimated depth values almost reaches 85% for the methods of Li et al. (2015) and Ramamonjisoa and Lepetit (2019). For shorter distances up to 3 m, the methods of Eigen et al. (2014)



(a) Ground truth (b) Liu et al., 2015 (c) Li et al., 2017 (d) PlaneNet (Liu et al., 2018) (e) (Ramamonjisoa and Lepetit, 2019)

Figure D.15: Visual results after applying *depth boundary errors* (DBEs) on `IBims-1`. Overlay of ground truth depth map with ground truth edge (—) (a) and depth map predictions with extracted edges (colored) using different methods (b-e)

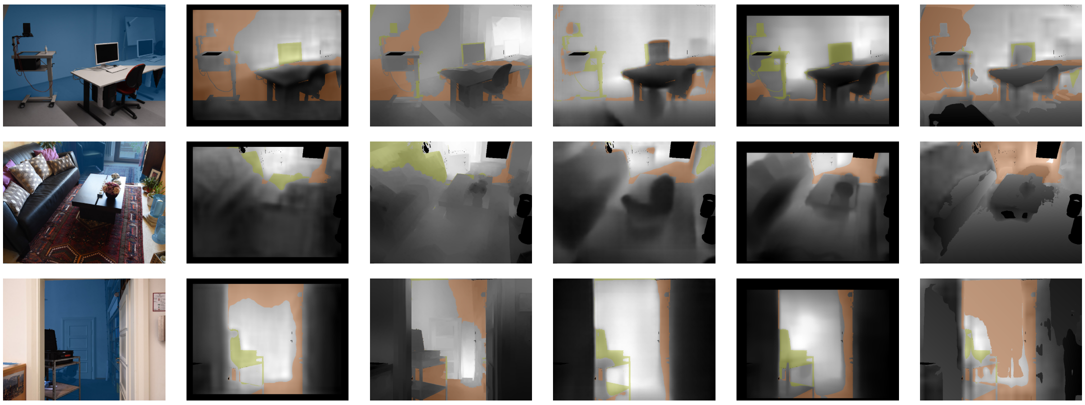
and PlaneNet (Liu et al., 2018) tend to underestimate to a larger extend compared to other methods, while the method of Liu et al. (2015) rather overestimated short distances. It is worth noting that the method of Ramamonjisoa and Lepetit (2019) exhibited a largely well-balanced distribution of over- and underestimated depths.

D.6 INFLUENCE ON THE PERFORMANCE OF SIDE METHODS

Furthermore, additional experiments were conducted to investigate the general behavior of SIDE methods, *i.e.*, the robustness of predicted depth maps to geometrical and color transformations, the planarity of predicted textured vertical surfaces, and the influence of different illumination in the scene.

D.6.1 *Augmentation*

In order to assess the robustness of SIDE methods w.r.t. simple geometrical and color transformation and noise, we derived a set of augmented images from our dataset. For geometrical transformations we flipped the input images horizontally—which is expected to not change the results significantly—and vertically, which is expected to expose slight overfitting effects. As images in the `NYU-v2` dataset usually show a considerable amount of pixels on the floor in the lower part of the picture, this is expected to notably influence the estimated depth maps. For color transformations, we consider swapping of image channels, shifting the hue by some offset h and scaling the saturation by a factor s. We change the gamma values to simulate over- and under-exposure and optimize the contrast by histogram stretching. Blurred versions of the images are simulated by applying Gaussian blur with increasing standard deviation σ. Furthermore, we consider noisy versions of the images by applying Gaussian additive noise and salt and pepper noise with increasing variance



(a) RGB (b) Eigen et al., 2014 (c) Liu et al., 2015 (d) Laina et al., 2016 (e) Li et al., 2017 (f) PlaneNet (Liu et al., 2018)

Figure D.16: Visual results after applying *directed depth errors* (DDEs) on `IBims-1`. Ground truth depth plane at $d = 3\,\text{m}$ separating foreground from background (🟦) (a). Differences between ground truth and predictions (b-f). Color coded are depth values that are either estimated too short (🟧) or too far (🟩)

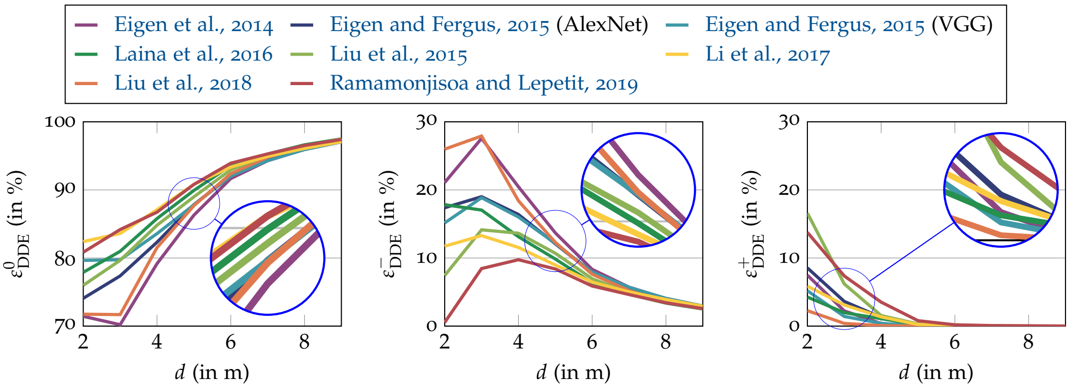
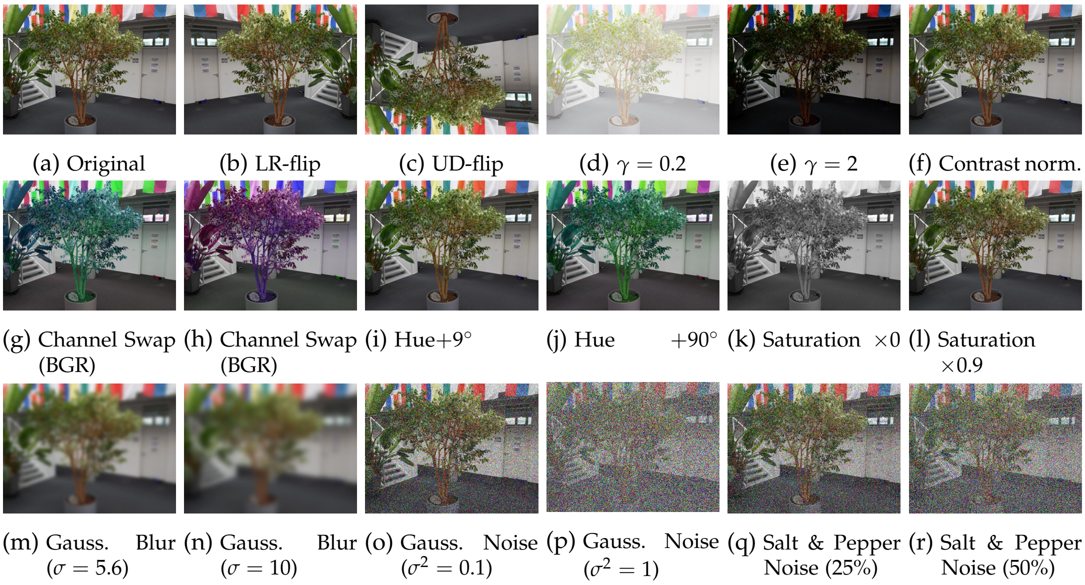


Figure D.17: *Directed depth errors* (DDEs) for different distances d of the virtual plane seperating foreground and background. From left to right: Proportions of correct ($\varepsilon^0_{\text{DDE}}$), too-close ($\varepsilon^-_{\text{DDE}}$) and too-far ($\varepsilon^+_{\text{DDE}}$) predicted pixels for different methods

and amount of affected pixels, respectively. Examples from this auxiliary dataset are shown in Figure D.18.

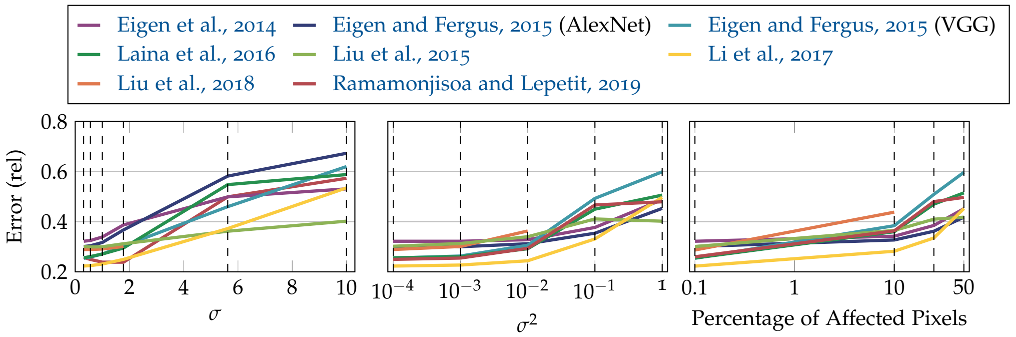
Table D.5 shows results for these augmented images using the global relative error metric for selected methods. As expected, the geometrical transformations yielded contrasting results. While the horizontal flipping did not influence the results by a large margin, flipping the images vertically increased the error by up to 60%. Slight overexposure influenced the result notably, underexposure seems to have been less problematic. Histogram stretching had no influence on the results, suggesting that this is already a fixed or learned part of the methods. The methods also seem to be robust to color changes, which is best seen in the results for $s = 0$, *i.e.*, grayscale input images which yielded an equal error to the reference. The results for blurring the input images with a Gaussian kernel of various standard deviations, as well as adding a different amount of Gaussian and salt and pepper noise to the input images are depicted in Figure D.19. Minor blurring did not change the results, as


(a) Original (b) LR-flip (c) UD-flip (d) $\gamma = 0.2$ (e) $\gamma = 2$ (f) Contrast norm.

(g) Channel Swap (BGR) (h) Channel Swap (BGR) (i) Hue+9° (j) Hue +90° (k) Saturation ×0 (l) Saturation ×0.9

(m) Gauss. Blur ($\sigma = 5.6$) (n) Gauss. Blur ($\sigma = 10$) (o) Gauss. Noise ($\sigma^2 = 0.1$) (p) Gauss. Noise ($\sigma^2 = 1$) (q) Salt & Pepper Noise (25%) (r) Salt & Pepper Noise (50%)

Figure D.18: Different geometric and radiometric augmentation samples applied to `IBims-1`



(a) Gaussian Blur (b) Gaussian Noise (c) Salt & Pepper Noise

Figure D.19: Quality of SIDE results for different methods after applying different augmentations with increasing intensity on `IBims-1`. Vertical lines (- -) correspond to discrete augmentation intensities

the examined methods considerably down-sample the input images and are thus robust to blurring up to a certain standard deviation. However, the performance of all methods starts to linearly decrease for blurring the image with $\sigma > 2$, whereby the methods of Eigen et al. (2014) and Liu et al. (2015) are more robust for larger blurring than the other methods. PlaneNet (Liu et al., 2018) could not handle blurring the image for standard deviations of the Gaussian distribution $\sigma > 2$ due to a failed vanishing point estimation.

The results for adding noise to the images, shown in Figures D.19b and D.19c, give certain thresholds for the maximum tolerable amount of noise. All of the considered methods were able to cope with up to 10 % of Salt and Pepper noise and Gaussian noise with variance of 0.01 until the quality of results decreased notably. The AlexNet version of Eigen and Fergus (2015) seems to be more robust to noise as opposed to the VGG version, which is, however, less sensitive to blurred input images. Again, the method of Liu et al. (2015) performed best on large noise levels, while PlaneNet (Liu et al., 2018) could not cope with a large amount of noise.

Table D.5: Quantitative results on the augmented IBims-1 dataset exemplary listed for the global relative distance error. Errors showing relative differences for various image augmentations towards the predicted original input image (Reference)

Method	Reference	Geometric		Contrast			Ch. Swap		Hue		Saturation	
		LR	UD	$\gamma = 0.2$	$\gamma = 2$	Norm.	BGR	BRG	+9°	+90°	×0	×0.9
Eigen et al., 2014	0.322	−0.003	0.087	0.056	0.015	0.000	0.017	0.018	0.001	0.021	0.003	−0.001
Eigen and Fergus, 2015 (AlexNet)	0.301	0.006	0.147	0.105	0.023	−0.002	0.017	0.008	0.002	0.017	0.007	−0.001
Eigen and Fergus, 2015 (VGG)	0.254	0.003	0.150	0.109	0.008	0.000	0.010	0.013	0.000	0.012	0.009	−0.001
Laina et al., 2016	0.255	−0.004	0.161	0.078	0.022	−0.001	0.007	0.009	0.000	0.007	0.003	−0.001
Liu et al., 2015	0.301	−0.004	0.079	0.021	0.011	−0.001	0.006	0.004	0.000	0.009	0.004	0.001
Li et al., 2017	0.222	0.001	0.152	0.024	0.004	0.001	0.016	0.014	0.003	0.019	0.015	0.001
Liu et al., 2018	0.287	0.003	0.204	0.069	0.025	−0.001	0.009	0.027	0.000	0.010	0.027	0.002
Ramamonjisoa and Lepetit, 2019	0.257	0.008	0.156	0.003	−0.003	0.000	0.012	0.010	−0.002	0.012	0.004	0.001

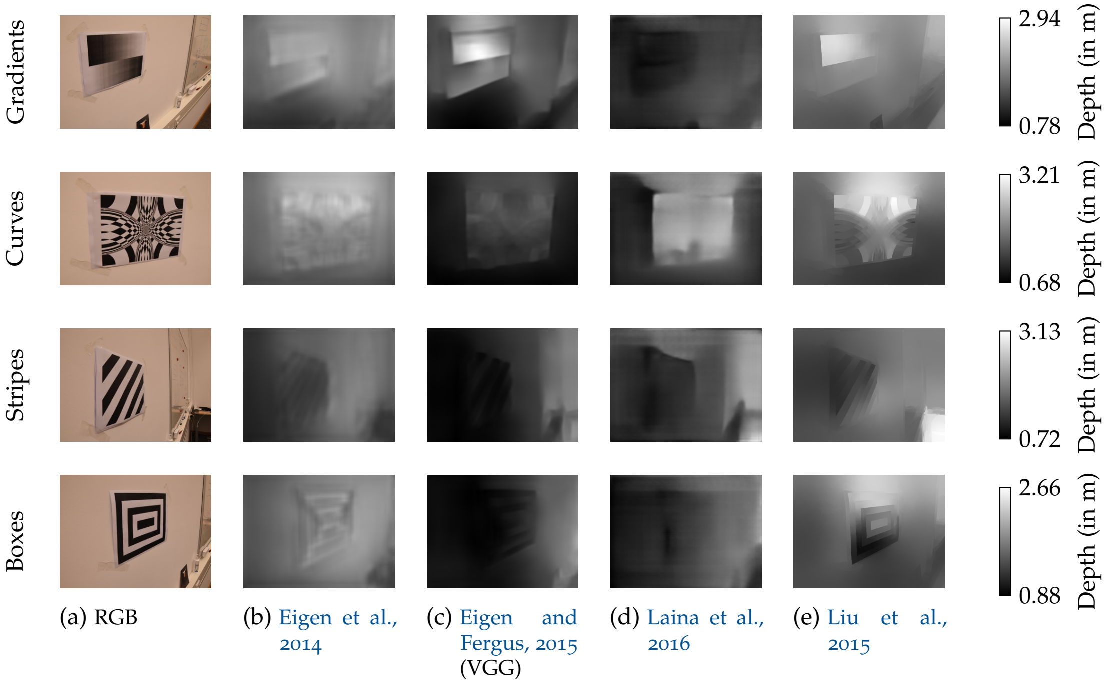


Figure D.20: Predictions for different printed samples from the Pattern dataset (Asuni and Giachetti, 2014) on a planar surface (rows). Predictions using different methods (b-f) of the input images (a). Predicted depth maps are color-coded according to the colormaps shown in the last column

D.6.2 *Textured Planar Surfaces*

Experiments with printed patterns and `NYU-v2` samples on a planar surface exploit which features influence the predictions of SIDE methods. As to be seen in Figure D.20, gradients seem to serve as a strong hint to the network. All of the tested methods estimated incorrectly depth in the depicted scene, none of them, however, identified the actual planarity of the picture. All of the examined networks respond to these patterns. However, this effect is less severe for Laina et al. (2016), which respond with only a constant offset to the alternating gradients in the pattern. Edges in the input also seem to influence the result as to be seen in Stripes and Boxes. Again, Laina et al. (2016) gave a constant offset, while the result of Liu et al. (2015) clearly contained artifacts of the superpixel approach, which is even more evident in Curves. Although `NYU-v2`, which served as training data for all methods, also contains such textured surfaces in terms of paintings and drawings on walls, the networks are unable to distinguish between intensity changes due to real depth discontinuities and solely texture. Further research in this field is needed in order to improve the applicability of SIDE in the fields of 3D room modeling or robot navigation.

D.6.3 *Illumination*

Illumination plays a significant role in recovering the 3D structure of a scene, especially for indoor scenarios where different types of natural and artificial illumination come together. This can be considered as a combination of under- and overexposure and intensity-based gradients on planar regions. As both effects were already discussed in sections Sections D.6.1 and D.6.2 separately, this experiment represents a real world scenario revealing these effects for current state-of-the-art methods. For this experiment we captured a static scene containing a table covering small objects in the foreground, as well as a white wall in the background separated by a floor lamp. We generated one ground truth depth map using a Kinect v1 and changed the scene illumination by various artificial lights, such as diffuse lighting from a floor lamp, and directional lighting from a spot appended on the floor lamp and a flashlight illuminates the scene from different viewpoints outside of the image. Depending on the illumination type, shadows cause strong gradients especially on the background wall. RGB images, predictions and quantitative results of the examined methods are visualized in Figure D.21. The results clearly show the impact of directional lighting of the spot creating depth changes according to the strong gradients on the right side of the wall, while diffuse lighting did not influence the results notably. While comparable performances of the different methods – especially for diffuse lighting – can be observed when using the global error metric, more distinguishable results can be noted applying the *planarity errors*. As in the evaluation in Section D.5, Liu et al. (2015) experiences difficulties in estimating the correct plane, while PlaneNet (Liu et al., 2018) successfully segmented the wall in each image and produces accurate 3D planes, although problems in the predictions of objects on the tables can be noticed, resulting in larger errors for the global metric.

D.7 CONCLUSIONS

We presented a novel set of quality criteria and a new high-quality RGB-D dataset for the evaluation of SIDE methods. We pointed out, that established error metrics which are used to assess the quality of predicted depth maps do not consider meaningful geometric properties, such as the preservation of depth boundaries and planar regions, the depth consistency across the image, and the depth range in the image. In order to gradually establish SIDE methods in industrial applications, different properties of the derived depth maps are decisive which highly depend on the application field. For instance, 3D indoor room modeling emphasizes accurate and correct plane estimations rather the reconstruction of detailed and small-scaled furnishings. Developing realistic occlusion-aware augmented reality applications, on the other hand, requires the reconstruction of precise and sharp depth discontinuities in occluded contours (Ramamonjisoa and Lepetit, 2019). With the growing popularity of 3D movies, SIDE techniques are partially used to substitute the costly and time-consuming stereoscopic video recording process or the manual 2D-to-3D conversion of single RGB images to arrive stereo pairs (Xie et al., 2016). Since most 3D animations consist mainly of a few discrete depth layers, the consistency of depth estimates for certain depth ranges becomes an important issue. In the field of autonomous driving, the accuracy assessment of distance estimates is often considered as non-linear, since higher accuracies are required for objects close to the camera than for faraway objects (Liebel and Körner, 2019). Therefore, a distance-related assessment of the depth maps would provide valuable insights into the performance of the methods for different depth ranges.

As all of these application samples focus on different geometric properties of SIDE, measurable evaluation metrics are needed to compare and understand the

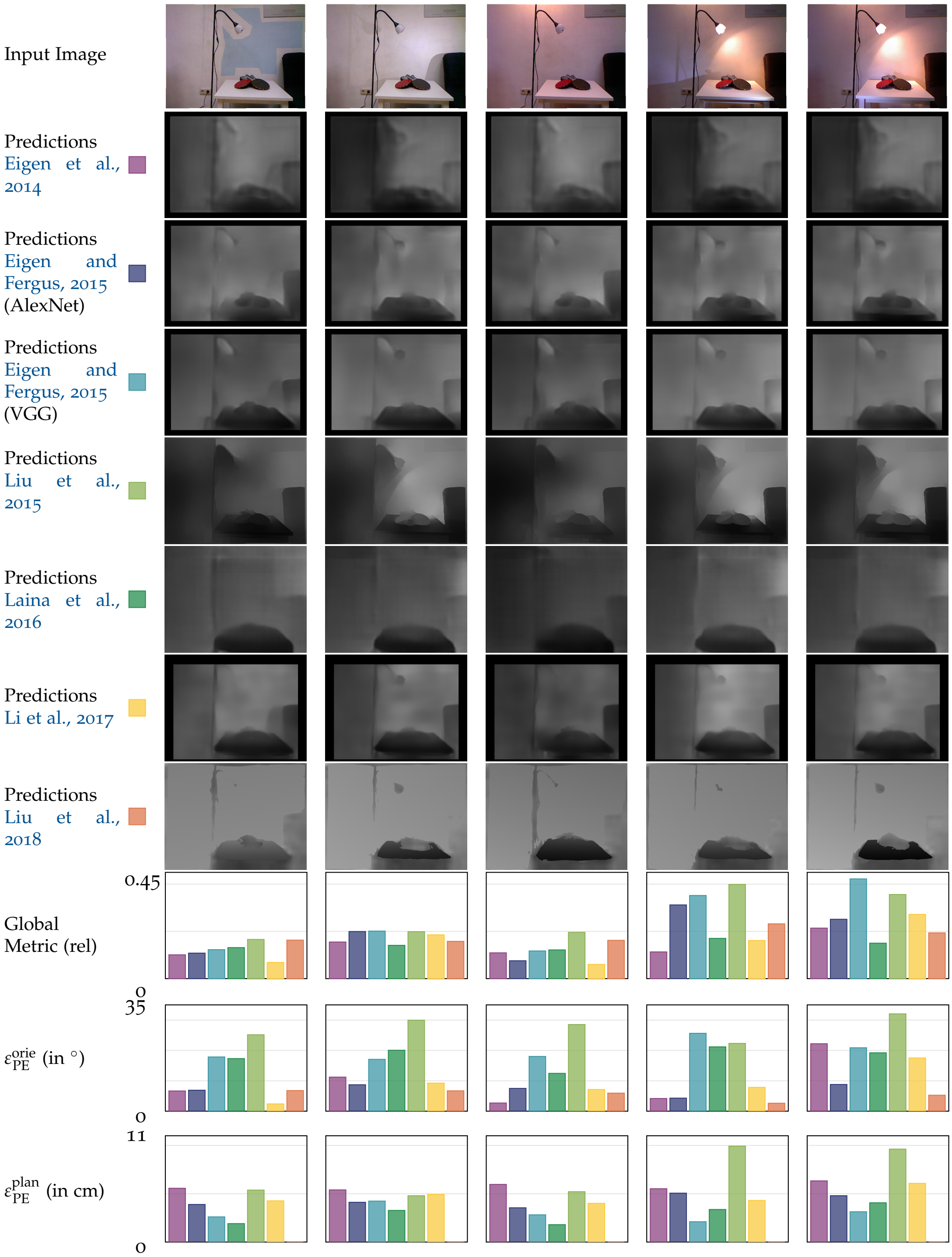


Figure D.21: Influence of different illumination on SIDE methods for a static scene. From top to bottom: Input RGB images, predictions using different SIDE methods, and errobars for global relative distance error and *planarity error* for annotated wall in the top-left image (🟦)

performance of both existing and novel methodologies in this field. We elaborated simple, but geometrically interpretable error metrics for the mentioned properties above. Particulary, these are *distance-related error metrics*, *planarity errors*, *depth boundary errors*, and *directed depth errros*. Since these metrics require precise, dense, and noise-free RGB-D image pairs, existing RGB-D datasets can not fully satisfy these high demands. For this reason, we introduced a new high-quality indoor RGB-D dataset, recorded with a custom acquisition setup combining a *laser scanner* and a *DSLR* camera to capture accurately aligned RGB-D image pairs. In our experiments, we were able to assess the quality of current state-of-the-art SIDE approaches w.r.t. to above mentioned properties, and unlike commonly used global metrics, our proposed set of quality criteria enabled us to unveil even subtle differences between the considered methods. In particular, our experiments have shown that the prediction of planar surfaces, which is crucial for many reconstruction applications, is lacking accuracy and CNN-based methods tend to produce smooth predictions resulting in blurry or vanishing depth boundaries. Although new methods that tackle specific aspects of the analyzed properties have been proposed recently, they still struggle to find a good trade-off between these aspects. Intuitively, a method that is designed and trained to predict sharp edges at depth discontinuities based on a single image, such as Sharpnet (Ramamonjisoa and Lepetit, 2019), tends to be sensitive to texture changes. Hence, a drop in the planarity metrics could be observed. Detecting planar regions in images and accurately predicting continuous depth values for such areas, as proposed in the PlaneNet approach of Liu et al. (2018), on the other hand, comes at the cost of disregarding finer details in favor of dominant planes. Our experiments showed that, again, the increased performance with respect to the targeted property is opposed by notable shortcomings in other aspects, most prominently the detection of edges. Additional experiments were conducted to test the robustness of the methods in terms of geometrical and radiometrical distortions, in the presence of textured planar surfaces and under varying lighting conditions. The results have proven a high robustness to minor blurring or noising of the input image, as well as to radiometrical changes. On the other hand, gradients and sharp intensity changes of planar objects, either caused by texture or illumination, can easily jar the methods in producing large depth changes. We believe that our dataset is suitable for future developments in this regard, as our images are provided in a very high resolution and contain new sceneries with extended scene depths. Together with our new proposed error metrics, it serves as an independent evaluation protocol for indoor depth prediction and helps to improve future developments in this field.

ACKNOWLEDGMENTS

This research was funded by the German Research Foundation (DFG) for Tobias Koch and the Federal Ministry of Transport and Digital Infrastructure (BMVI) for Lukas Liebel. We thank our colleagues from the Chair of Geodesy for providing all the necessary equipment and our student assistant Leonidas Stöckle for his help during the data acquisition campaign.

REFERENCES

Ackermann, J. and Goesele, M. (2015). "A survey of photometric stereo techniques." Foundations and Trends in Computer Graphics and Vision 9(**3-4**), pp. 149–254.

Anwar, S., Hayder, Z., and Porikli, F. (2017). "Depth estimation and blur removal from a single out-of-focus image." In: *Proceedings of the British Machine Vision Conference (BMVC)*.

Armeni, I., Sax, S., Zamir, A. R., and Savarese, S. (2017). "Joint 2D-3D-semantic data for indoor scene understanding." arXiv preprint arXiv:1702.01105.

Asuni, N. and Giachetti, A. (2014). "Testimages: a large-scale archive for testing visual devices and basic image processing algorithms." In: *Smart Tools and Apps for Graphics - Eurographics Italian Chapter Conference*. The Eurographics Association, pp. 63–70.

Baig, M. H. and Torresani, L. (2016). "Coupled depth learning." In: *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–10.

Camplani, M. and Salgado, L. (2014). "Background foreground segmentation with RGB-D Kinect data: an efficient combination of classifiers." Journal of Visual Communication and Image Representation 25(**1**), pp. 122–136.

Chakrabarti, A., Shao, J., and Shakhnarovich, G. (2016). "Depth from a single image by harmonizing overcomplete local network predictions." In: *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, pp. 2658–2666.

Chang, A., Dai, A., Funkhouser, T., Halber, M., Nießner, M., Savva, M., Song, S., Zeng, A., and Zhang, Y. (2017). "Matterport3D: learning from RGB-D data in indoor environments." In: *Proceedings of the IEEE International Conference on 3D Vision (3DV)*, pp. 667–676.

Chen, W., Fu, Z., Yang, D., and Deng, J. (2016). "Single-image depth perception in the wild." In: *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, pp. 730–738.

Choi, S., Min, D., Ham, B., Kim, Y., Oh, C., and Sohn, K. (2015). "Depth analogy: data-driven approach for single image depth estimation using gradient samples." IEEE Transactions on Image Processing (TIP) 24(**12**), pp. 5953–5966.

Dai, A., Chang, A. X., Savva, M., Halber, M., Funkhouser, T., and Nießner, M. (2017). "ScanNet: richly-annotated 3D reconstructions of indoor scenes." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2432–2443.

Devernay, F. and Beardsley, P. (2010). "Stereoscopic cinema." In: *Image and Geometry Processing for 3-D Cinematography*. Springer, pp. 11–51.

Dhamo, H., Tateno, K., Laina, I., Navab, N., and Tombari, F. (2019). "Peeking behind objects: layered depth prediction from a single image." Pattern Recognition Letters 125, pp. 333–340.

Doorn, A. J. van, Koenderink, J. J., and Wagemans, J. (2011). "Light Fields and Shape from Shading." Journal of Vision 11(**3**), pp. 21.1–21.21.

Eigen, D. and Fergus, R. (2015). "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture." In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2650–2658.

Eigen, D., Puhrsch, C., and Fergus, R. (2014). "Depth map prediction from a single image using a multi-scale deep network." In: *Proceedings of Advances in Neural Information Processing Systems (NIPS)*. Vol. 2, pp. 2366–2374.

Favaro, P. and Soatto, S. (2005). "A geometric approach to shape from defocus." IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 27(**3**), pp. 406–417.

Felzenszwalb, P. F. and Huttenlocher, D. P. (2006). "Efficient belief propagation for early vision." International Journal of Computer Vision (IJCV) 70(**1**), pp. 41–54.

Fu, H., Gong, M., Wang, C., Batmanghelich, K., and Tao, D. (2018). "Deep ordinal regression network for monocular depth estimation." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2002–2011.

Furukawa, R., Sagawa, R., and Kawasaki, H. (2017). "Depth estimation using structured light flow–analysis of projected pattern flow on an object's surface." In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 4640–4648.

Garg, R., Carneiro, G., and Reid, I. (2016). "Unsupervised CNN for single view depth estimation: geometry to the rescue." In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 740–756.

Geiger, A., Lenz, P., and Urtasun, R. (2012). "Are we ready for autonomous driving? the kitti vision benchmark suite." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3354–3361.

Godard, C., Mac Aodha, O., and Brostow, G. J. (2017). "Unsupervised monocular depth estimation with left-right consistency." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6602–6611.

Guo, X., Li, H., Yi, S., Ren, J., and Wang, X. (2018). "Learning monocular depth by distilling cross-domain stereo networks." In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 484–500.

Hane, C., Ladicky, L., and Pollefeys, M. (2015). "Direction matters: Depth estimation with a surface normal classifier." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 381–389.

Hao, Z., Li, Y., You, S., and Lu, F. (2018). "Detail preserving depth estimation from a single image using attention guided networks." In: *Proceedings of the IEEE International Conference on 3D Vision (3DV)*, pp. 304–313.

Hartley, R. and Zisserman, A. (2003). *Multiple view geometry in computer vision*. Cambridge University Press.

Hassner, T. and Basri, R. (2006). "Example based 3D reconstruction from single 2D images." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR-WS)*, pp. 8–15.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778.

Heber, S. and Pock, T. (2016). "Convolutional networks for shape from light field." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3746–3754.

Heo, M., Lee, J., Kim, K.-R., Kim, H.-U., and Kim, C.-S. (2018). "Monocular depth estimation using whole strip masking and reliability-based refinement." In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 36–51.

Hirschmuller, H. (2005). "Accurate and efficient stereo processing by semi-global matching and mutual information." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 2, pp. 807–814.

Hoiem, D., Efros, A. A., and Hebert, M. (2007). "Recovering surface layout from an image." International Journal of Computer Vision (IJCV) 75(**1**), pp. 151–172.

Horn, B. K. P. (1970). *Shape from shading: a method for obtaining the shape of a smooth opaque object from one view*. Tech. rep. Cambridge, MA, USA: MIT - AI.

Hu, J., Ozay, M., Zhang, Y., and Okatani, T. (2019). "Revisiting single image depth estimation: toward higher resolution maps with accurate object boundaries." In: *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1043–1051.

Izadinia, H., Shan, Q., and Seitz, S. M. (2017). "Im2cad." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5134–5143.

Kadambi, A., Taamazyan, V., Shi, B., and Raskar, R. (2015). "Polarized 3D: high-quality depth sensing with polarization cues." In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 3370–3378.

Karsch, K., Liu, C., and Kang, S. B. (2014). "Depth transfer: Depth extraction from video using non-parametric sampling." IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 36(**11**), pp. 2144–2158.

Kim, S., Park, K., Sohn, K., and Lin, S. (2016). "Unified depth prediction and intrinsic image decomposition from a single image via joint convolutional neural fields." In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 143–159.

Knapitsch, A., Park, J., Zhou, Q.-Y., and Koltun, V. (2017). "Tanks and Temples: benchmarking large-scale scene reconstruction." ACM Transactions on Graphics (TOG) 36(**4**), p. 78.

Koch, T., Liebel, L., Fraundorfer, F., and Körner, M. (2018). "Evaluation of CNN-based single-image depth estimation methods." In: *Proceedings of the European Conference on Computer Vision Workshops (ECCV-WS)*. Springer, pp. 331–348.

Kolmogorov, V. and Zabih, R. (2001). "Computing visual correspondence with occlusions using graph cuts." In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 508–515.

Kong, N. and Black, M. J. (2015). "Intrinsic depth: improving depth transfer with intrinsic images." In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 3514–3522.

Konrad, J., Brown, G., Wang, M., Ishwar, P., Wu, C., and Mukherjee, D. (2012). "Automatic 2D-to-3D image conversion using 3D examples from the internet." In: *Proceedings of the Stereoscopic Displays and Applications*. International Society for Optics and Photonics, 82880F.

Konrad, J., Wang, M., Ishwar, P., Wu, C., and Mukherjee, D. (2013). "Learning-based, automatic 2D-to-3D image and video conversion." IEEE Transactions on Image Processing (TIP) 22(**9**), pp. 3485–3496.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "Imagenet classification with deep convolutional neural networks." In: *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, pp. 1097–1105.

Kuznietsov, Y., Stückler, J., and Leibe, B. (2017). "Semi-supervised deep learning for monocular depth map prediction." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6647–6655.

Ladicky, L., Shi, J., and Pollefeys, M. (2014). "Pulling things out of perspective." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 89–96.

Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., and Navab, N. (2016). "Deeper depth prediction with fully convolutional residual networks." In: *Proceedings of the IEEE International Conference on 3D Vision (3DV)*, pp. 239–248.

Lee, J.-H., Heo, M., Kim, K.-R., and Kim, C.-S. (2018). "Single-image depth estimation based on Fourier domain analysis." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 330–339.

Levin, A., Lischinski, D., and Weiss, Y. (2004). "Colorization using optimization." ACM Transactions on Graphics (TOG) 23(**3**), pp. 689–694.

Li, B., Shen, C., Dai, Y., Hengel, A. van den, and He, M. (2015). "Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1119–1127.

Li, J., Klein, R., and Yao, A. (2017). "A two-streamed network for estimating fine-scaled depth maps from single RGB images." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3372–3380.

Li, X., Qin, H., Wang, Y., Zhang, Y., and Dai, Q. (2014). "DEPT: depth estimation by parameter transfer for single still images." In: *Proceedings of the Asian Conference on Computer Vision (ACCV)*. Springer, pp. 45–58.

Li, Z. and Snavely, N. (2018). "MegaDepth: learning single-view depth prediction from internet photos." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2041–2050.

Liebel, L. and Körner, M. (2019). "MultiDepth: single-image depth estimation via multi-task regression and classification." arXiv preprint arXiv:1907.11111.

Liu, B., Gould, S., and Koller, D. (2010). "Single image depth estimation from predicted semantic labels." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1253–1260.

Liu, C., Yuen, J., and Torralba, A. (2011). "Sift flow: dense correspondence across scenes and its applications." IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 33(**5**), pp. 978–994.

Liu, C., Yang, J., Ceylan, D., Yumer, E., and Furukawa, Y. (2018). "PlaneNet: piece-wise planar reconstruction from a single RGB image." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2579–2588.

Liu, F., Shen, C., and Lin, G. (2015). "Deep convolutional neural fields for depth estimation from a single image." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5162–5170.

Liu, F., Shen, C., Lin, G., and Reid, I. (2016). "Learning depth from single monocular images using deep convolutional neural fields." IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 38(**10**), pp. 2024–2039.

Liu, M., Salzmann, M., and He, X. (2014). "Discrete-continuous depth estimation from a single image." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 716–723.

Mancini, M., Costante, G., Valigi, P., and Ciarfuglia, T. A. (2018). "J-MOD 2: joint monocular obstacle detection and depth estimation." IEEE Robotics and Automation Letters 3(**3**), pp. 1490–1497.

McCormac, J., Handa, A., Leutenegger, S., and Davison, A. J. (2017). "Scenenet RGB-D: can 5m synthetic images beat generic imagenet pre-training on indoor segmentation." In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Vol. 4, pp. 2697–2706.

Moreno-Noguer, F., Lepetit, V., and Fua, P. (2007). "Accurate non-iterative o (n) solution to the pnp problem." In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1–8.

Nayar, S. K. and Narasimhan, S. G. (1999). "Vision in bad weather." In: *Proceedings of the IEEE International Conference on Computer Vision (CVPR)*. Vol. 2, pp. 820–827.

Ngo, T. T., Nagahara, H., and Taniguchi, R.-i. (2015). "Shape and light directions from shading and polarization." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2310–2318.

Occipital, I (2016). *Structure sensor-3d scanning, augmented reality, and more for mobile devices.*

Phan, R. and Androutsos, D. (2013). "Robust semi-automatic depth map generation in unconstrained images and video sequences for 2D to stereoscopic 3D conversion." IEEE Transactions on Multimedia 16(**1**), pp. 122–136.

Ramamonjisoa, M. and Lepetit, V. (2019). "SharpNet: fast and accurate recovery of occluding contours in monocular depth estimation." In: *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCV-WS)*, tbd.

Ranftl, R., Vineet, V., Chen, Q., and Koltun, V. (2016). "Dense monocular depth estimation in complex dynamic scenes." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4058–4066.

Roy, A. and Todorovic, S. (2016). "Monocular depth estimation using neural regression forest." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5506–5514.

Saxena, A., Chung, S. H., and Ng, A. Y. (2006). "Learning depth from single monocular images." In: *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, pp. 1161–1168.

– (2008). "3-d depth reconstruction from a single still image." International Journal of Computer Vision (IJCV) 76(**1**), pp. 53–69.

Saxena, A., Sun, M., and Ng, A. Y. (2009). "Make3d: learning 3D scene structure from a single still image." IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 31(**5**), pp. 824–840.

Scharstein, D. and Szeliski, R. (2002). "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms." International Journal of Computer Vision (IJCV) 47(**1-3**), pp. 7–42.

Schöps, T., Schönberger, J. L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., and Geiger, A. (2017). "A multi-view stereo benchmark with high-resolution images and multi-camera videos." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2538–2547.

Seitz, S. M., Curless, B., Diebel, J., Scharstein, D., and Szeliski, R. (2006). "A comparison and evaluation of multi-view stereo reconstruction algorithms." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 519–528.

Shi, J., Tao, X., Xu, L., and Jia, J. (2015). "Break ames room illusion: depth from general single images." ACM Transactions on Graphics (TOG) 34(**6**), p. 225.

Silberman, N., Hoiem, D., Kohli, P., and Fergus, R. (2012). "Indoor segmentation and support inference from rgbd images." In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 746–760.

Simonyan, K. and Zisserman, A. (2014). "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556.

Song, S., Yu, F., Zeng, A., Chang, A. X., Savva, M., and Funkhouser, T. (2017). "Semantic scene completion from a single depth image." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 190–198.

Strecha, C., Von Hansen, W., Van Gool, L., Fua, P., and Thoennessen, U. (2008). "On benchmarking camera calibration and multi-view stereo for high resolution imagery." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8.

Su, C.-C., Cormack, L. K., and Bovik, A. C. (2017). "Bayesian depth estimation from monocular natural images." Journal of Vision 17(**5**), pp. 22–22.

Suwajanakorn, S., Hernandez, C., and Seitz, S. M. (2015). "Depth from focus with your mobile phone." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3497–3506.

Szeliski, R. (2010). *Computer vision: algorithms and applications*. Springer Science & Business Media.

Torralba, A. and Oliva, A. (2002). "Depth estimation from image structure." IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 24(**9**), pp. 1226–1238.

Ummenhofer, B., Zhou, H., Uhrig, J., Mayer, N., Ilg, E., Dosovitskiy, A., and Brox, T. (2017). "DeMoN: depth and motion network for learning monocular stereo." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 5, pp. 5038–5047.

Wang, P., Shen, X., Lin, Z., Cohen, S., Price, B., and Yuille, A. L. (2015). "Towards unified depth and semantic prediction from a single image." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2800–2809.

Wang, P., Shen, X., Russell, B., Cohen, S., Price, B., and Yuille, A. L. (2016). "Surge: surface regularized geometry estimation from a single image." In: *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, pp. 172–180.

Xian, K., Shen, C., Cao, Z., Lu, H., Xiao, Y., Li, R., and Luo, Z. (2018). "Monocular relative depth perception with web stereo data supervision." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 311–320.

Xie, J., Girshick, R., and Farhadi, A. (2016). "Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks." In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 842–857.

Xu, D., Ricci, E., Ouyang, W., Wang, X., and Sebe, N. (2017). "Multi-scale continuous CRFs as sequential deep networks for monocular depth estimation." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 161–169.

Xu, D., Wang, W., Tang, H., Liu, H., Sebe, N., and Ricci, E. (2018). "Structured attention guided convolutional neural fields for monocular depth estimation." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3917–3925.

Yang, F. and Zhou, Z. (2018). "Recovering 3D planes from a single image via convolutional neural networks." In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 85–100.

Yin, Z. and Shi, J. (2018). "GeoNet: unsupervised learning of dense depth, optical flow and camera pose." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1983–1992.

Yoon, K.-J. and Kweon, I. S. (2006). "Adaptive support-weight approach for correspondence search." IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 28(**4**), pp. 650–656.

You, X., Li, Q., Tao, D., Ou, W., and Gong, M. (2014). "Local metric learning for exemplar-based object detection." IEEE Transactions on Circuits and Systems for Video Technology 24(**8**), pp. 1265–1276.

Zennaro, S, Munaro, M., Milani, S., Zanuttigh, P., Bernardi, A, Ghidoni, S., and Menegatti, E. (2015). "Performance evaluation of the 1st and 2nd generation Kinect for multimedia applications." In: *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, pp. 1–6.

Zhan, H., Garg, R., Weerasekera, C. S., Li, K., Agarwal, H., and Reid, I. (2018). "Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 340–349.

Zhang, R., Tsai, P.-S., Cryer, J. E., and Shah, M. (1999). "Shape from shading: a survey." IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 21(**8**), pp. 690–706.

Zheng, C., Cham, T.-J., and Cai, J. (2018). "T2Net: synthetic-to-realistic translation for solving single-image depth estimation tasks." In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 798–814.

Zhou, T., Brown, M., Snavely, N., and Lowe, D. G. (2017). "Unsupervised learning of depth and ego-motion from video." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6612–6619.

Zhuo, W., Salzmann, M., He, X., and Liu, M. (2015). "Indoor scene structure analysis for single image depth estimation." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 614–622.

Zioulis, N., Karakottas, A., Zarpalas, D., and Daras, P. (2018). "OmniDepth: dense depth estimation for indoors spherical panoramas." In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, pp. 453–471.

Zoran, D., Isola, P., Krishnan, D., and Freeman, W. T. (2015). "Learning ordinal relationships for mid-level vision." In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 388–396.