Technische Universität München
TUM School of Computation, Information and Technology

# Robust Learned 3D-Perception

Patrick Winfried Ruhkamp

Vollständiger Abdruck der von der TUM School of Computation, Information and Tech-nology der Technischen Universität München zur Erlangung eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

*Vorsitz:* Prof. Dr. Stefan Leutenegger

*Prüfer der Dissertation:* 1. Prof. Dr. Nassir Navab

2. Prof. Dr. Aleš Leonardis

Die Dissertation wurde am 14.12.2023 bei der Technischen Universität München eingereicht und durch die TUM School of Computation, Information and Technology am 26.04.2024 angenommen.

**Patrick Winfried Ruhkamp**
*Robust Learned 3D-Perception*
Dissertation, Version 1.0

**Technische Universität München**
TUM School of Computation, Information and Technology
Lehrstuhl für Informatikanwendungen in der Medizin
Boltzmannstraße 3
85748 Garching bei München

# Abstract

Robustly perceiving the 3D information of a scene is a fundamental task in computer vision and essential for autonomous systems and many applications ranging from AR/VR, 3D reconstruction, or robotics. This dissertation addresses the lack of reliable 3D ground truth data, particularly in outdoor settings or for scenes with photometrically complex objects, through data-driven learning methods that bypass the need for annotated data for robust learned 3D perception. It focuses on dense pixel-wise depth estimation and 6D object pose estimation.

A spatial-temporal attention mechanism is introduced to enhance self-supervised depth estimation from monocular image sequences regarding spatial coherence and temporal consistency. Both are vital for dynamic outdoor environments like autonomous driving, tackling issues like frame-to-frame drift and scale drift across sequences. By introducing a novel Temporal Consistency Metric (TCM), depth consistency across consecutive frames can be objectively quantified. However, the observed artifacts and noisy measurements of the LiDAR sensor used as ground-truth, questions the reliability and accuracy of active sensors for training and evaluation.

Novel multi-modal dataset acquisition techniques with different depth sensors and polarimetric imaging are introduced to systematically analyze sensor characteristics, integrating robotic forward-kinematics and advanced calibration methods. This leads to highly accurate datasets, focusing on photometrically challenging objects, and unveils sensor-specific artifacts affecting depth measurements and subsequent 3D vision tasks.

This data acquisition identifies the limitations of current RGB-D-based methods for 6D object pose estimation, especially under conditions with unreliable depth data for photometrically challenging objects. Polarimetric imaging, which encodes shape information, is proposed for robust 6D object pose estimation. The resulting Polarimetric Pose Prediction model, short PPP-Net, a supervised multi-modal hybrid approach utilizing polarimetric data, significantly surpasses existing RGB-only and RGB-D methods, especially for photometrically challenging objects. Additionally, novel network and loss components are proposed, along with neural rendering and an invertible polarimetric physical model, for self-supervised 6D object pose estimation using RGB+polarization data. This substantially improves pose estimation robustness and accuracy for challenging objects without relying on annotated real data, thereby addressing the shortcomings in current self-supervised RGB-D approaches.

# Zusammenfassung

Die robuste Wahrnehmung der 3D-Informationen einer Szene ist eine grundlegende Aufgabe des maschinellen Sehens und essentiell für autonome Systeme und viele Anwendungen, die von AR/VR, 3D-Rekonstruktion oder Robotik reichen. Diese Dissertation befasst sich mit dem Mangel an verlässlichen 3D-Ground-Truth-Daten, insbesondere im Außenbereich oder bei photometrisch komplexen Objekten, durch datengesteuerte Lernmethoden, die die Notwendigkeit für annotierte Daten für robuste 3D-Wahrnehmung umgehen. Der Schwerpunkt liegt dabei auf der pixelweisen Tiefenschätzung und der 6D-Objektposenschätzung.

Es wird ein spatial-temporal-attention-Mechanismus eingeführt, um die selbstüberwachte Tiefenschätzung aus monokularen Bildsequenzen hinsichtlich räumlicher Kohärenz und zeitlicher Konsistenz zu verbessern. Beides ist für dynamische Umgebungen im Freien, wie z.B. autonomes Fahren, von entscheidender Bedeutung, um Probleme wie Frame-to-Frame-Drift und Skalendrift über Sequenzen hinweg zu bewältigen. Durch die Einführung einer neuartigen temporalen Konsistenzmetrik (TCM) kann die Tiefenkonsistenz zwischen aufeinanderfolgenden Bildern objektiv quantifiziert werden. Die beobachteten Artefakte und verrauschten Messungen des LiDAR-Sensors, der als Ground-Truth verwendet wird, stellen jedoch die Zuverlässigkeit und Genauigkeit aktiver Sensoren für Training und Auswertung in Frage.

Neuartige multimodale Datenerfassungstechniken mit verschiedenen Tiefensensoren und polarimetrischer Bildgebung werden eingeführt, um die Sensoreigenschaften systematisch zu analysieren, wobei robotergestützte Vorwärtskinematik und fortschrittliche Kalibrierungsmethoden integriert werden. Dies führt zu hochpräzisen Datensätzen, die sich auf photometrisch schwierige Objekte konzentrieren, und deckt sensorspezifische Artefakte auf, die Tiefenmessungen und nachfolgende 3D-Vision-Aufgaben beeinflussen.

Diese Datenerfassung zeigt die Grenzen aktueller RGB-D-basierter Methoden zur Schätzung der 6D-Objektpose auf, insbesondere unter Bedingungen mit unzuverlässigen Tiefendaten für photometrisch schwierige Objekte. Für eine robuste 6D-Objektpositionsschätzung wird die Integration polarimetrischer Bilder vorgeschlagen, welche Forminformationen kodieren. Das daraus resultierende Polarimetric Pose Prediction Modell, kurz PPP-Net, ein überwachter multimodaler hybrider Ansatz, der polarimetrische Daten nutzt, übertrifft die bestehenden reinen RGB- und RGB-D-Methoden deutlich, insbesondere bei photometrisch schwierigen Objekten. Zusätzlich werden neuartige Netzwerkkomponenten und Verlustfunktionen zusammen mit neuronalem Rendering und einem invertierbaren polarimetrischen physikalischen Modell für die selbstüberwachte 6D-Objektposenschätzung unter Verwendung von RGB+Polarisation Daten vorgeschlagen. Dies verbessert die Robustheit und Genauigkeit der Posenschätzung für schwierige Objekte erheblich, ohne auf annotierte reale Daten angewiesen zu sein, und behebt so die Unzulänglichkeiten aktueller selbstüberwachter RGB-D-Ansätze.

# Acknowledgments

First and foremost, I would like to thank Nassir for allowing me to pursue my Ph.D. at CAMP and for your continuous support, also outside of academia.

Next, I want to thank Ben, who has been more than just a mentor for me during my Ph.D.

To all my friends and colleagues at TUM, CAMP, and to everyone at all other stations during my Ph.D.: Thanks for everything!

I also want to say thanks to all my students. It was fantastic working with you and learning from you.

And most importantly:

I owe a debt of gratitude to my family, friends, and loved ones. I appreciate your support and belief in me. Your love, patience, and encouragement have supported me throughout this journey.

# Contents

# Part I

## Introduction

# 1

# Introduction

## 1.1 Motivation & Objectives

**Robust learned 3D-perception enables machines to see, think, and act.**

In a world where autonomous systems seamlessly interact with their surroundings, accurate and robust 3D perception is essential. Picture a robot navigating a cluttered room, autonomous vehicles navigating the bustling streets, or a robotic arm delicately grasping a glass from a table. In all these scenarios, a profound 3-dimensional understanding of the environment is the key to unlocking the potential of these machines.

Imagine for a moment how humans perceive their surroundings. We use the power of binocular vision, harnessing stereo cues to transform the images from our two eyes into a rich 3D representation. Once we estimate the distance, we seamlessly interact with the objects around us, whether picking up a glass or navigating the world. This fusion of knowing how far something is away and understanding its precise position and orientation within the 3D space allows us to interact effortlessly. We state that 3D perception in the context of this dissertation is the combination of depth estimation and where and how objects are positioned and orientated within the scene. Here, we neglect any other understanding of the scene, such as semantics, scene graphs, or others.

[51] This dissertation investigates methodologies enabling machines to learn robust 3D perception, essential in challenging scenarios, particularly unbound scenes like outdoor scenarios in autonomous driving for self-supervised monocular depth prediction [51] and 6D pose estimation [42, 165, 174] for photometrically complex objects with translucency or reflectivity. Many approaches in machine-based 3D perception involve binocular passive camera systems or active sensor technologies like LiDAR (Light Detection and Ranging). The binocular system, mimicking human stereoscopic vision, computes depth through the disparity between two camera perspectives [139]. However, this method faces constraints such as the requirement for a specific camera baseline, which might not be feasible in compact systems, and the computational intensity associated with image processing and synchronization. Conversely, active sensing technologies, which include for instance LiDAR sensors and Time-of-Flight (ToF) cameras, offer distance measurements by analyzing the reflection time of emitted light pulses. While effective in certain aspects, these technologies are hindered by limitations in form factor, energy efficiency, operational range, data sparsity, artifacts such as multi-path interference (MPI), and susceptibility to environmental factors like reflective surfaces or translucent materials. Given these limitations, this dissertation uses data-driven deep learning techniques, without the reliance on annotated data or active sensors, for robust learned 3D perception in challenging scenarios, like unbound outdoor scenes or with photometrically complex objects.

In scenarios such as autonomous driving, accurately perceiving and understanding the surrounding 3D environment is essential. In self-supervised monocular depth prediction, consecutive images captured by a moving monocular camera are used to enforce photometric consistency between adjacent frames after projectively transforming them by the regressed dense depth and relative camera poses. Despite significant advancements in this field [51], achieving consistency in temporal and geometric depth across frames without losing depth accuracy remains a significant issue [9].

> How can we fully leverage the spatial-temporal relation between consecutive frames to predict spatially coherent and temporally consistent depth estimates while maintaining high accuracy? Moreover, how can we objectively quantify depth consistency?

The reliability and accuracy of active sensors, e.g., as observed for LiDAR [103], are questioned due to artifacts and noise in measurements. This issue is especially pertinent in outdoor scenes where obtaining detailed and accurate depth ground truth for sensor comparison is challenging. Conversely, it is feasible in indoor settings to reconstruct individual objects with high precision using advanced 3D scanners before they are positioned in the scene, offering a more reliable basis for comparison.

> How can active depth sensors' characteristics and associated artifacts be systematically analyzed, also considering challenging scenarios?

Polarization of light, a passively observable physical property, offers valuable insights where depth measurements can be noisy. This polarimetric information is particularly advantageous in environments where photometric challenges limit the effectiveness of active sensors, as they encode shape cues of objects. Leveraging these shape cues can be instrumental in estimating the 6D pose of an object, which includes accurately determining its exact position and orientation rather than just measuring its distance from the camera.

> How can we integrate the physical properties of polarized light into a learning pipeline for robust 3D perception tasks like 6D object pose estimation?

Even though polarization might alleviate specific problems inherent in active depth sensors for the 6D object pose estimation task of photometrically challenging objects, there is still a substantial requirement for a large volume of annotated data for supervised training. This reliance on extensive annotation is impractical, particularly as the process of annotating such data can be laborious, time-intensive, and prone to introducing errors or inconsistencies in the annotations.

> Can we avoid the need for annotated real data - potentially by leveraging polarization for self-supervision?

## 1.2 Contributions

We will answer the questions above with the contributions described here.

The first contribution improves the temporal consistency of self-supervised monocular depth estimation.

**1. Spatial-Temporal Constraints for Consistent Monocular Depth Estimation**

> **Patrick Ruhkamp***, Daoyi Gao*, Hanzhi Chen*, Nassir Navab, and Benjamin Busam. "Attention meets Geometry: Geometry Guided Spatial-Temporal Attention for Consistent Self-Supervised Monocular Depth Estimation". IEEE **3DV 2021**. [132] ( *authors contributed equally)

> **Patrick Ruhkamp***, Daoyi Gao*, Hanzhi Chen*, Nassir Navab, and Benjamin Busam. "Spatial-Temporal Attention through Self-Supervised Geometric Guidance". **ICCV 2021** Workshop SSLAD. [133] ( *authors contributed equally)

This dissertation introduces a spatial-temporal attention mechanism to enhance self-supervised depth estimation from monocular image sequences regarding spatial coherence and temporal consistency. Both are vital for dynamic outdoor environments like autonomous driving, tackling issues like frame-to-frame drift and scale drift across sequences. Introducing a novel Temporal Consistency Metric (TCM) can quantify depth consistency across consecutive frames objectively.

Given the observed artifacts and noisy measurements of the LiDAR sensor used as ground-truth from before, the second contribution systematically analyzes depth sensor characteristics.

**2. Sensor Characteristics and Dense 3D Perception**

> HyunJun Jung*, **Patrick Ruhkamp***, Gunagyao Zhai, Nikolas Brasch, Yitong Li, Yannick Verdie, Jifei Song, Yiren Zhou, Anil Armagan, Slobodan Ilic, Ales Leonardis, Nassir Navab, and Benjamin Busam. "On the Importance of Accurate Geometry Data for Dense 3D Vision Tasks". IEEE **CVPR 2023**. [79] ( *authors contributed equally)

> HyunJun Jung*, **Patrick Ruhkamp***, Nassir Navab, and Benjamin Busam. "Multi-Modal Dataset Acquisition for Photometrically Challenging Object". **ICCV 2023** Workshop on Transparent & Reflective Objects in the Wild Challenges. [78] ( *authors contributed equally)

> HyunJun Jung*, Guangyao Zhai*, Shun-Cheng Wu*, **Patrick Ruhkamp***, Hannah Schieber*, Giulia Rizzoli, Pengyuan Wang, Hongcheng Zhao, Lorenzo Garattoni, Sven Meier, Daniel Roth, Nassir Navab, and Benjamin Busam. "HouseCat6D – A Large-Scale Multi-Modal Category Level 6D Object Perception Dataset with Household Objects in Realistic Scenarios". IEEE **CVPR 2024**. [80] ( *authors contributed equally)

Novel multi-modal dataset acquisition techniques are introduced to systematically analyze sensor characteristics, including different depth sensors and polarimetric imaging, by integrating robotic forward-kinematics and advanced calibration methods. This leads to highly accurate data acquisitions, focusing on photometrically challenging objects. Due to the sensor artifacts, depth data show noisy and incorrect measurements as expected, translating to dense 3D vision tasks when used for supervision.

Given the observations of corrupt depth data, especially for photometrically challenging objects, from above, multi-modal RGB+polatization is proposed for robust 6D object pose estimation as the third contribution.

### 3. Physical Constraints for 6D Object Pose Estimation

> Daoyi Gao[*], Yitong Li[*], **Patrick Ruhkamp**[*], Iuliia Skobleva[*], H[*], HyunJun Jung, Pengyuan Wang, Arturo Guridi and Benjamin Busam. "Polarimetric Pose Prediction". **ECCV 2022**. [42] ( [*]authors contributed equally)
>
> **Patrick Ruhkamp**, Daoyi Gao, HyunJun Jung, Nassir Navab, and Benjamin Busam. "Polarimetric Information for Multi-Modal 6D Pose Estimation of Photometrically Challenging Objects with Limited Data". **ICCV 2023** Workshop on Transparent & Reflective Objects in the Wild Challenges. [134] ( [*]authors contributed equally)
>
> **Patrick Ruhkamp**[*], Daoyi Gao[*], Nassir Navab and Benjamin Busam. "S2P3: Self-Supervised Polarimetric Pose Prediction". **IJCV 2024**. [135] ( [*]authors contributed equally)

The previous data acquisition identifies the limitations of current RGB-D-based methods for 6D object pose estimation in terms of accuracy and robustness, especially under conditions with unreliable depth data for photometrically challenging objects. Polarimetric imaging, which encodes shape information, is proposed for robust 6D object pose estimation. The resulting Polarimetric Pose Prediction model, short PPP-Net, a supervised multi-modal hybrid approach utilizing polarimetric data, significantly surpasses existing RGB-only and RGB-D methods, especially for photometrically challenging objects.

Additionally, novel network and loss components are proposed, along with neural rendering and an invertible polarimetric physical model, for self-supervised 6D object pose estimation using RGB+polarization data, resulting in the Self-Supervised Polarimetric Pose Prediction pipeline, short $S^2P^3$. This substantially improves pose estimation robustness and accuracy for challenging objects without relying on annotated real data, thereby addressing the shortcomings in current self-supervised RGB-D approaches and avoiding the need for accurate ground-truth data.

# Part II

## Spatial-Temporal Constraints for Consistent Monocular Depth Estimation

**How can we fully leverage the spatial-temporal relation between consecutive frames to predict spatially coherent and temporally consistent depth estimates while maintaining high accuracy? And how can we objectively quantify depth consistency?**

# 2 3D Perception from Monocular Camera Ego-Motion

*“Stand out of my light.”*

**– Diogenes to Alexander the Great**

## 2.1 Introduction

Depth estimation from a single image is essential in computer vision, forming the basis for robust 3D perception and numerous downstream applications. The onset of deep learning saw researchers applying supervised learning to predict dense depth from monocular RGB images, utilizing datasets annotated with ground truth from active depth sensors [34, 91]. However, creating these large-scale, accurately annotated datasets is often unfeasible, expensive, and time-consuming.

In self-supervised learning, a binocular camera setup can mimic human vision, assuming photometric consistency, where a point seen from one camera should match its appearance in the other camera’s image. Neural networks regress depth maps by minimizing the photometric discrepancy after projectively transforming pixels between images using the predicted depth and camera parameters [50]. When a stereo setup is unavailable, a moving camera can act as multiple stereo systems with varying poses between consecutive frames. Depth and pose estimation are coupled problems, as evidenced by structure from motion and SLAM/VO techniques. An additional network can estimate the relative 6D camera pose between frames, with transformation parameters becoming part of the photometric consistency optimization [51].

Recent advances have improved depth accuracy, but challenges remain in ensuring robust predictions across consecutive frames regarding temporal consistency and spatial coherence [9]. Issues like drift and scale drift between frames persist. Additional supervision signals like velocity data or constraints from other sensors could be used. Also, extensive optimization approaches akin to bundle adjustment in video sequences can enforce consistency, but these approaches can be computationally expensive.

The question arises: How can we algorithmically enhance consistency and coherence in depth predictions to achieve robust 3D perception? Drawing parallels with Natural Language Processing (NLP), where words in a sentence are akin to multiple consecutive images, we explore the potential of transformer-like principles in depth estimation. However, integrating transformers into monocular depth prediction is not straightforward, as the scene content changes between frames, and the attention mechanism may correlate scene content across images that are spatially not reasonable. Our approach first integrates spatial attention per frame, which serves as a 3D-positional or spatial encoding for subsequent temporal attention across frames. We propose epipolar constraints and novel self-supervised loss functions to further guide the attention.

### 2.1.1 Motivation

The enhancement of self-supervised monocular depth prediction accuracy has been a key focus recently, as highlighted by significant research efforts [19, 51]. Despite these advances, the challenge of predicting temporally and geometrically consistent depth across multiple frames still needs to be explored. Such consistency is crucial for a wide range of 3D vision applications, including 3D reconstruction [115], SLAM [179], pose estimation [13], medical applications [16], AR/MR [105], computational photography [14], and autonomous driving [48].

Inconsistent depth predictions can significantly impair downstream tasks, such as 6D object pose estimation [35, 168, 188], also in autonomous driving [64, 65] or RGB-D reconstruction [115]. Classical Structure from Motion (SfM) and visual odometry approaches have traditionally tackled geometric consistency by employing computationally intensive techniques like local and global bundle adjustment [113, 114]. Some depth prediction methods have recently attempted to enforce consistency, either by incorporating additional ground truth signals like velocity [53] or by considering whole sequences with recurrent units [122].

The standard evaluation of depth accuracy in monocular self-supervised methods often uses median scaling against depth ground truth [51], a process applied independently per image, neglecting the consistency of predictions across frames. This approach must be revised for real-world applications, as it fails to capture pixel-wise variations and overall depth scale consistency.

Traditional methods enforcing geometric consistency compromise depth accuracy, leading to blurred edges and smooth depth discontinuities [9]. In response, our proposed spatial-temporal attention model uniquely correlates geometrically meaningful and spatially coherent features. This approach maintains temporal aggregations across frames without negatively af-

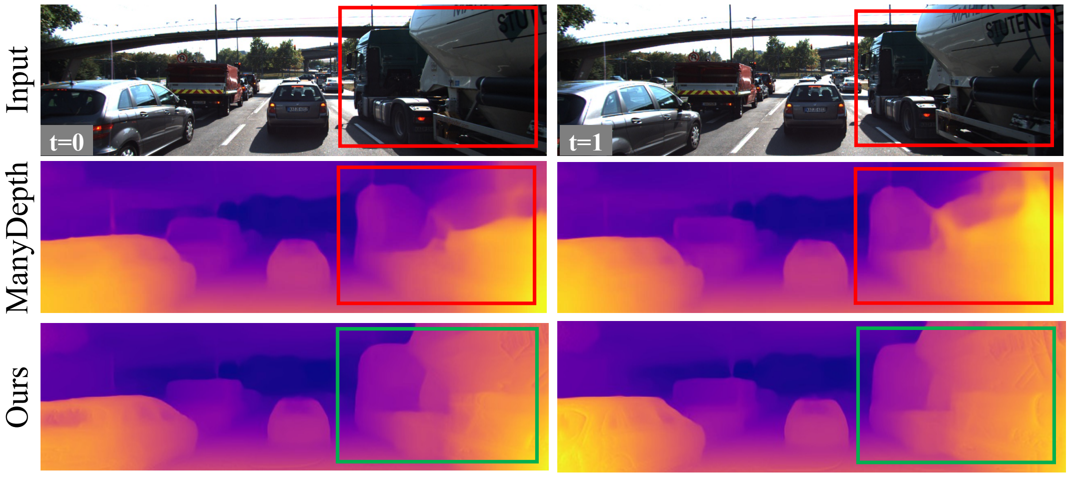


Fig. 2.1 **Qualitative Depth Results:** Strong baselines in self-supervised depth prediction, such as ManyDepth [171], often manifest noticeable flickering effects between consecutive images. In contrast, our method excels in estimating temporally consistent depth across successive frames. Remarkably, it showcases proficiency even when confronted with large dynamically moving objects in the scene.

fecting depth accuracy, providing focused and accurate attention between frames, a significant advancement over previous methods [9].

We advocate for a greater emphasis on temporal consistency in depth predictions by proposing a new metric to quantify this aspect. The qualitative 3D reconstruction from consecutive depth predictions, as shown in Fig. 2.1, underscores this need. Although the recent ManyDepth approach [171] achieves impressive accuracy, its predictions suffer from inconsistencies, leading to less reliable scene reconstructions. In contrast, our model achieves elevated accuracy metrics and ensures highly consistent depth predictions, even for dynamically moving objects (Fig. 2.1), resulting in improved scene reconstructions.

### 2.1.2 Contributions

Our primary objective is to robustly predict depth from monocular image sequences that are accurate, coherent, and temporally consistent despite challenging factors like occlusions, disocclusions, moving objects, and complex camera movements. We integrate scene information from temporally adjacent frames into our method, achieving this without additional supervision signals or reliance on complex, non-real-time networks.

In self-supervised monocular depth estimation, the primary focus has historically been improving accuracy on a per-frame basis. While there have been efforts to enhance temporal consistency, these often involve additional geometric loss terms applied between consecutive frames during training. However, implementing these geometric consistency constraints can lead to diminished depth accuracy, manifesting as blurred edges and smoothed depth discontinuities [9]. Additionally, no established metric can effectively quantify the temporal consistency of consecutive depth predictions.

To tackle these challenges, we introduce **TC-Depth**, our **t**emporally **c**onsistent **depth** estimation pipeline. **TC-Depth** is designed to balance geometric consistency and depth accuracy. It leverages an innovative spatial-temporal attention module to explicitly learn temporally consistent features for depth prediction. Combined with our geometric regularization techniques, our method achieves high accuracy while ensuring unparalleled consistency across frames.

Further validation of our approach is provided through a comprehensive ablation study to shed light on our novel components' contributions to consistency and accuracy. Notably, we demonstrate how the introduction of photometric cycle consistency significantly enhances the performance of our attention mechanism. Moreover, we have developed the Temporal Consistency Metric (TCM), specifically designed to quantify the performance of coherent depth prediction across temporal frames.

The primary contributions in this chapter are in summary:

## Contributions

1. A novel **spatial attention** mechanism that aggregates and utilizes local geometric information, enhancing the depth prediction.
2. A **temporal attention** module designed to operate across tuples of monocular frames, promoting global consistency and coherence.
3. A novel **cycle consistency regularization** strategy which offers **geometric guidance**. This guidance aids the fusion of feature embeddings in the **spatial-temporal attention mechanism**.
4. The introduction of a **Temporal Consistency Metric** (TCM) that provides a quantifiable measure for evaluating the consistency of depth predictions across frames.

## 2.2 Related Work

### 2.2.1 Self-Supervised Depth Estimation

In computer vision, dense depth prediction is essential in numerous applications and is a fundamental building block. The accuracy and consistency of depth predictions can significantly impact the success of subsequent tasks. For instance, in autonomous vehicles, the reliable estimation of objects in 3D holds paramount importance, especially in safety-critical situations [64, 65]. Similarly, within the domain of RGB-D reconstruction [115], precision and consistency of depth information are indispensable for generating accurate and visually coherent models of the surrounding environment. Addressing the challenge of maintaining geometric consistency in visual data is not a novel pursuit. Classical techniques like Structure from Motion (SfM) and visual odometry have already approached this challenge. These traditional methods often rely on bundle adjustment techniques to achieve geometric consistency. The core objective of bundle adjustment is to refine initial pose and structure estimates by minimizing the reprojection error across all cameras and 3D points involved [113, 114]. However, it is worth noting that such strategies, while effective, can be computationally demanding and may not meet the requirements of real-time feedback in specific scenarios. Additionally, such approaches only provide sparse 3D reconstructions. Modern depth prediction methodologies have endeavored to incorporate mechanisms that enhance consistency. Some approaches leverage additional ground truth signals, such as velocity data of a moving camera, to constrain their depth predictions [53]. Others harness the capabilities of recurrent neural networks (RNNs) to process sequences of images, thus ensuring temporal consistency across consecutive frames [122]. Additionally, some methods aim to enforce geometric consistency by regularizing depth predictions by geometric constraints between adjacent frames [9].

The domain of depth estimation has witnessed remarkable advancements in recent years, primarily attributed to the adoption of convolutional neural networks (CNNs) for this task. Early explorations into CNN-based depth estimation, exemplified by works like Eigen et al. [34], Laina et al. [91], and Fu et al. [39], have demonstrated the potential of supervised methods in monocular depth estimation. However, a substantial challenge persists in acquiring accurate ground-truth depth datasets, particularly in environments such as outdoor spaces or expansive scenes, where obtaining ground-truth depth data is difficult [48]. The limitations of extensive labeled datasets have given rise to self-supervised methods. Pioneering researchers in this domain, such as Xie et al. [176] and Garg et al. [45], have leveraged stereo imagery during training to propose self-supervised learning techniques, notably employing photometric consistency losses.

Subsequent advancements introduced the concept of left-right consistencies in a fully differentiable manner through the Monodepth framework [50]. The idea was extended to the temporal domain in MonoDepth2 [51], where predicting relative camera poses allows for photo-consistency losses between neighboring frames after projective transformation. While early attempts at simultaneously estimating depth and camera pose showed promise in terms of robustness [6, 190, 194], they lagged in accuracy when compared to traditional methodologies. Integrating optical flow predictions improved depth accuracy, especially for moving objects within scenes, and introduced forward-backward consistency checks. These constraints introduced a novel mechanism for discerning occlusions, proving crucial for depth accuracy at depth discontinuities in the scene [73, 169, 182].

### 2.2.2 Attention for Depth Estimation

Self-attention mechanisms, originating from the domain of natural language processing [160], have become popular in computer vision tasks [104, 192]. Unlike traditional convolutional methods that rely on fixed kernels to process image data [88], self-attention offers dynamic operations tailored to individual image and feature inputs. Mathematically, self-attention can be perceived as a weighted sum of all feature responses, where the weights signify the relevance of other features to the current one. Taking this concept further, Huynh et al. [70] introduced a depth-attention volume that accentuates planar structures, optimizing results for interior environments where such structures are predominant. Sadek et al. [136] leveraged attention gates during the decoding phase of their depth estimation pipeline. In another perspective, Lee et al. [92] employed patch-wise attention to pool information from spatially proximal features in a supervised setting. Yang et al. [178] integrated transformers into an expansive architecture, achieving superior prediction accuracy. However, their methodology primarily focuses on supervised learning. Johnston et al. [76] embarked on a transformative journey by integrating transformers within self-supervised depth estimation, focusing on expansive outdoor scenes. Their proposed mechanism utilizes a self-attention module after a ResNet-based encoder [56], decoding the depth information through a discrete disparity volume. However, a limitation emerged with their approach—their self-attention mechanism, in its current design, struggles to draw meaningful feature correlations, impairing its effectiveness in 3D scene regression tasks.

### 2.2.3 Consistent Depth Estimation

Previous research in the domain of depth consistency metrics inadequately addressed both geometric and temporal consistency. Zhang et al. [190] developed a method to assess the structural similarity between successive depth maps. However, their approach did not include spatial alignment, which is crucial for accurate depth interpretation. Luo et al. [105] incorporated an optical-flow-based KLT tracker to measure the 3D Euclidean distances between photometrically corresponding points. This approach is significantly influenced by the accuracy of the optical flow estimation, which can be a limiting factor in complex scenes.

In self-supervised monocular depth estimation, a predominant focus has been maintaining a constant overall scale of depth predictions. This focus influences the auxiliary pose network, making it more suited for odometry applications. Bian et al. [9] proposed a scale-consistent depth and ego-motion approach by incorporating a depth consistency loss. While this helps reduce scale drift in poses and depth predictions, it may compromise depth accuracy. Zhao et al. [193] presented a method that does not directly regress the 6-DOF camera transformation. Instead, they estimate optical flow between frames to establish correspondences for relative pose estimation using epipolar geometry. Ensuring consistency between triangulated points and the predicted depth is vital to achieving scale consistency. MonoRec [172] also focuses on visual odometry, achieving impressive results by constructing a photometric error cost volume to manage static and dynamic elements in a multi-view stereo setup. However, their approach requires additional supervision on dense stereo depth predictions and a complex training scheme.

Other studies, like those by Luo et al. [105] and NeuralRGBD [100], concentrate on static, small-scale indoor scenes. Luo et al. use learning-based priors and test-time training, optimizing across all pixels in a monocular video for consistent small-scale reconstructions. NeuralRGBD focuses on consistency by integrating multiple depth estimates from video sequences into a probability volume, aggregating consistent 3D scene information for indoor scene reconstruction in a supervised setting. To further exploit input image sequences, Patil et al. [122] employed recurrent units to enhance depth prediction accuracy over multiple frames in a self-supervised approach. However, this method requires long sequences during both the training and testing phases. ManyDepth [171] suggests utilizing adjacent frames from a monocular video sequence during inference by forming a cost volume that aggregates encoded features from multiple frames. This method is more efficient than previous test-time refinement techniques [142] and achieves highly accurate self-supervised depth predictions. However, it also requires predicting relative poses between frames. Our analysis shows that despite improved accuracy and the utilization of multiple consecutive test frames, temporal consistency in depth predictions is not invariably achieved. Our investigations emphasize that high accuracy in monocular depth estimation does not guarantee temporal consistency, accentuating the need for further research and refinement in this domain.

## 2.3 Consistent Self-Supervised Monocular Depth Estimation from Spatial-Temporal Constraints

Depth estimation from monocular image sequences remains a pivotal challenge in computer vision, with the essential issue residing not only in accuracy but also in the consistent prediction of depth information across temporal sequences. To address this, we propose our **Temporally-Consistent Depth** estimation method, short **TC-Depth**. The goal of **TC-Depth** is to attain accurate depth while ensuring temporal consistency, a critical factor in processing sequences of monocular images. This objective is achieved by leveraging a well-established approach that regresses depth and relative camera poses concurrently. The core principle involves minimizing the image reconstruction loss. This minimization is achieved by employing the backwards warping technique, which projectively transforms adjacent frames into a central view using the predicted dense depth and relative pose information, as suggested in prior works [51].

We underscore the significance of temporal consistency in depth predictions, recognizing its critical role in subsequent tasks such as RGB-D reconstruction. To underscore the necessity for consistent depth estimations, we refer to the qualitative 3D reconstructions derived from con-

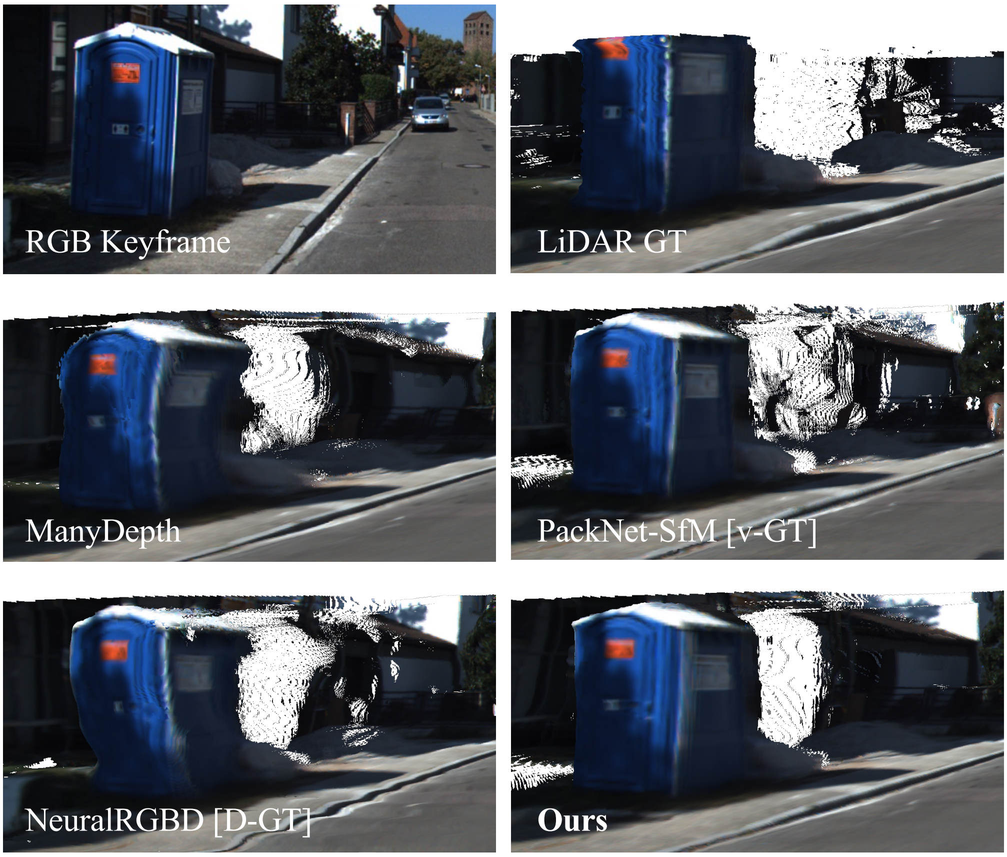


Fig. 2.2 **Qualitative 3D Reconstruction Comparison:** 3D reconstruction from five consecutive depth predictions on Kitti [48]. Our method yields a higher quality reconstruction due to geometrically consistent depth predictions with high accuracy compared against SOTA methods in self-supervised (ManyDepth [171]), semi-supervised (PackNet-SfM [53] with pose velocity [v-GT]), and supervised (NeuralRGBD [100] with depth [D-GT]) methods. Twisted boundaries due to pixel-wise misalignment and "flying pixels" are significantly reduced.

secutive depth predictions, as demonstrated in Fig. 2.2. These reconstructions vividly illustrate the imperative for consistent depth predictions and justify our motivation. While the recently introduced ManyDepth [171] currently leads in terms of quantitative depth accuracy, its depth predictions often exhibit temporal inconsistency, leading to noisy scene reconstructions. Our model, in contrast, not only attains comparable accuracy metrics but, more crucially, delivers temporally consistent depth predictions. This is also evident in handling dynamically moving objects, as shown in Fig. 2.1. The consistency achieved by our model significantly enhances the quality of scene reconstructions, providing a more coherent and reliable understanding of the scene's spatial structure. This emphasis on temporal consistency, especially in the context of dynamic objects, sets our model apart and highlights its potential applicability in a wide range of real-world scenarios where accurate and consistent depth perception is paramount.

Common artifacts, such as twisted boundaries arising from pixel-wise misalignment and the notorious "flying pixels", are markedly reduced. This showcases the geometric and temporal consistency and the high accuracy of the depth predictions offered by our method. A closer look at the reconstructions in Fig. 2.2 reveals:

- **Self-Supervised Comparison (ManyDepth [171]):** When compared against ManyDepth, a leading method in the self-supervised paradigm, our technique consistently producdes a more coherent and detailed 3D reconstruction.

- **Semi-Supervised Comparison (PackNet-SfM [53]):** Even in a semi-supervised setting, our method demonstrates superior performance. While comparing against PackNet-SfM (which incorporates ground-truth pose velocity [v-GT]), our method offers more precise and geometrically consistent depth predictions.

- **Supervised Comparison (NeuralRGBD [100]):** NeuralRGBD, which trains in a supervised setting with depth ground truth [D-GT], is known for its high accuracy in depth estimation. Nevertheless, our method proves to be competitive, offering comparable, if not better, 3D reconstructions.

## 2.3.1 Enabling Spatial-Temporal Constraints

The architecture of our proposed network is depicted in Fig. 2.3. Regarding the regression of relative camera poses, our approach adopts established strategies, aligning with methodologies presented in prior studies [51, 171].

Selecting an appropriate feature encoder is essential to align the resolution of the features for the attention module in the bottleneck. We capitalize on the benefits of dilated convolutions, as presented by Yu et al. [184], which provide expanded receptive fields and results in the desired resolution. The encoder of choice, a DRN-C-26, is similar to the well-known ResNet18 architecture but with dilated strides and the inclusion of de-griding layers. These layers are

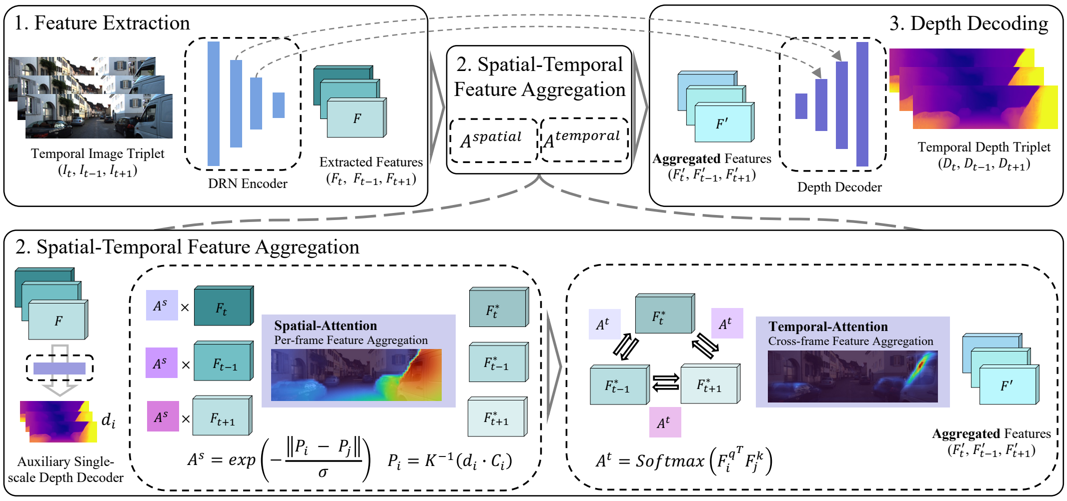

1. Feature Extraction
Temporal Image Triplet
$(I_t, I_{t-1}, I_{t+1})$
DRN Encoder
$F$
Extracted Features
$(F_t,\ F_{t-1}, F_{t+1})$
2. Spatial-Temporal Feature Aggregation
$A^{spatial}$
$A^{temporal}$
3. Depth Decoding
$F'$
Aggregated Features
$(F'_t, F'_{t-1}, F'_{t+1})$
Depth Decoder
Temporal Depth Triplet
$(D_t, D_{t-1}, D_{t+1})$
2. Spatial-Temporal Feature Aggregation
$F$
$d_i$
Auxiliary Single-scale Depth Decoder
$A^s$ × $F_t$
$A^s$ × $F_{t-1}$
$A^s$ × $F_{t+1}$
Spatial-Attention
Per-frame Feature Aggregation
$F^*_t$
$F^*_{t-1}$
$F^*_{t+1}$
$A^s = exp\left(-\frac{\|P_i - P_j\|}{\sigma}\right)$
$P_i = K^{-1}(d_i \cdot C_i)$
$A^t$
$F^*_t$
$A^t$
$F^*_{t-1}$
$F^*_{t+1}$
$A^t$
Temporal-Attention
Cross-frame Feature Aggregation
$F'$
Aggregated Features
$(F'_t, F'_{t-1}, F'_{t+1})$
$A^t = Softmax\left({F_i^q}^T F_j^k\right)$

Fig. 2.3 **Pipeline Overview:** 1. Image features are extracted using a dilated residual network (DRN). 2. An auxiliary low-resolution depth map is predicted by a single-stage reference decoder. This depth map is subsequently passed to the spatial attention module to establish local geometric correlation. The temporal attention module then aggregates these spatially-aware features across frames. 3. Aggregated features are decoded into the final depth predictions, with the assistance of skip connections derived from the encoder.

crucial in reducing the checkerboard artifacts often associated with certain convolutional patterns [184]. As we discuss later in the ablation studies, the influence of the encoder with dilated convolutions on depth accuracy and consistency is only marginal compared to the non-dilated counterpart, i.e., a comparable ResNet.

The encoder's feature embedding fulfills a dual role. Firstly, it contributes to an auxiliary single-scale depth decoder [52, 76], generating an initial coarse depth prediction. This preliminary output then serves as input to the subsequent spatial-temporal attention module. Notably, this attention mechanism is executed at the most reduced resolution level, specifically at a dimension of 24 × 80, corresponding to an eighth of the original input size.

Our temporal attention module integrates the encoded input features by drawing an analogy with optical flow techniques. This integration effectively consolidates temporally consistent scene content in conjunction with spatial attention. The combined information from this process is then fed into the final depth decoder. This approach ensures that the depth predictions are accurate and temporally coherent, enhancing the overall performance of **TC-Depth** in generating consistent and reliable depth estimations from monocular image sequences.

#### 2.3.1.1 Attention Module

Convolutional Neural Networks (CNNs) have inherent limitations due to their restricted receptive fields, hindering their ability to correlate features from spatially distant regions of an input. Transformers, initially conceived for Natural Language Processing (NLP) as described by Vaswani et al. [160], possess the ability to correlate semantically related words regardless of their positional distance within a sentence. This concept has been adapted for computer vision applications, as demonstrated by Dosovitskiy et al. [30], shifting the focus from correlating words to correlating pixels or specific pixel patches.

In the transformer's attention mechanism, inputs are categorized as query (Q), key (K), and value (V). The function of Q is to extract relevant information from V, directed by the attention weight. This process is encapsulated in the equation:

$$\text{Attention}(Q, K, V) = \mathscr{A}(Q, K)V, \tag{2.1}$$

where $\mathscr{A}(\cdot)$ represents a function that calculates a similarity score, serving as the attention weight for aggregating various feature embeddings.

Recent progress in the field [156] has shown that transformer models, particularly those employing self- and cross-attention mechanisms, can outperform fully convolutional networks in tasks that involve finding dense correspondences between image pairs [95]. Inspired by these advancements, we have incorporated a spatial-temporal attention module in our framework. This module leverages the transformer's ability to correlate distant features, thereby enhancing the capability of our system to consistently and accurately interpret spatial and temporal information from monocular image sequences.

**Spatial-Attention Layer** As proposed in Johnston et al. [76], self-attention enables the establishment of correlations within the same image, focusing on regions that display visual similarities. However, a critical aspect to consider in the context of self-attention, especially when applied to dense depth regression tasks, is the potential limitation of the dot-product mechanism used within the attention module. This mechanism can inadequately lead to the aggregation of features from geometrically distant regions of a 3D scene. Such a scenario is often observed at depth discontinuities, where there is a distinct separation between foreground and background objects. In these cases, the aggregation of features across these discontinuities might not be ideal for accurately regressing depth.

This issue arises because the self-attention mechanism, by design, does not inherently account for the geometric relationships between different parts of an image. It focuses more on appearance-based similarities, leading to the possibility of correlating features from different depths within the scene. This characteristic can be a significant drawback for tasks where understanding the depth and spatial arrangement of objects is crucial. It underscores the need for careful consideration and additional mechanisms to ensure that depth regression tasks maintain an awareness of the spatial and geometric properties of the scene, in addition to the visual similarities identified through self-attention.

We propose a model that explicitly integrates self-attention with 3D spatial awareness, leveraging an initially predicted coarse depth estimation. Given the camera intrinsics $\mathbf{K}$ and paired coordinates $\mathbf{C_i} = (u_i, v_i)$ and $\mathbf{C_j} = (u_j, v_j)$, each associated with depths $d_i$ and $d_j$ respectively, our first step is to back-project these pixel coordinates into the 3D space:

$$\begin{aligned} \mathbf{P}_i &= \mathbf{K}^{-1}(d_i \cdot \mathbf{C}_i), \\ \mathbf{P}_j &= \mathbf{K}^{-1}(d_j \cdot \mathbf{C}_j). \end{aligned} \tag{2.2}$$

We then explicitly define the spatial-attention as:

$$\mathbf{A}_{i,j}^{spatial} = \exp\left(-\frac{||\mathbf{P}_i - \mathbf{P}_j||_2}{\sigma}\right), \tag{2.3}$$

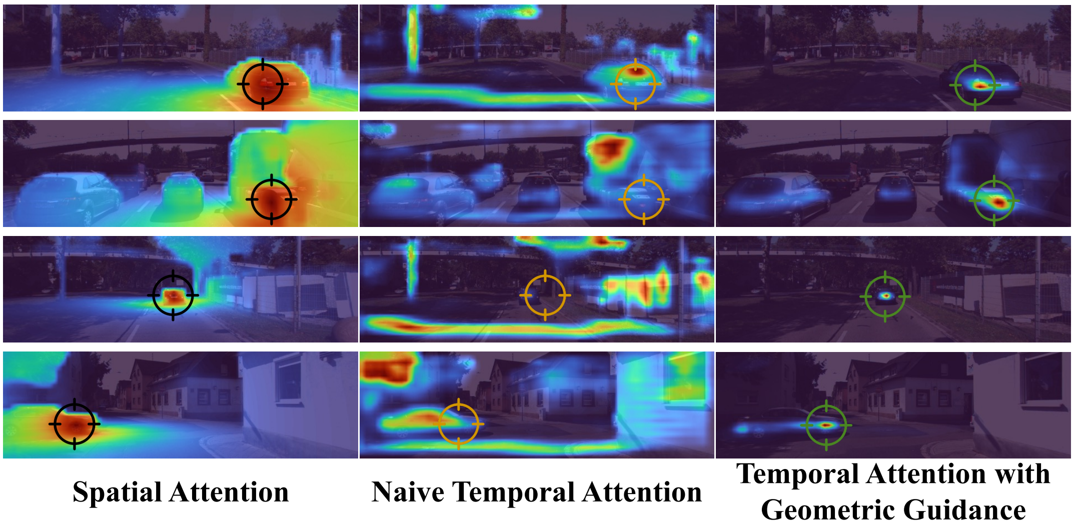
where $\mathbf{P}_i$ and $\mathbf{P}_j$ can be understood as the key and query respectively. This formulation can be interpreted as a 3D positional encoding grounded in 3D spatial correlation.

**Temporal-Attention Layer** Drawing inspiration from the correlation layer used in optical flow studies [71] and recent dense matching approaches [156], we introduce an innovative temporal attention mechanism. This mechanism is specifically tailored to leverage the temporal sequence of images provided by the self-supervised training paradigm.

Given a triplet set of feature maps derived from successive image inputs, we iteratively designate one as the query and use the others as key features. Subsequently, we determine the similarities between these key and query features through the Softmax function. Let $\mathbf{F}_i^q$ represent the query feature and $\mathbf{F}_j^k$ signify the key feature. The temporal-attention can be represented as:

$$\mathbf{A}_{i,j}^{temporal} = \mathrm{Softmax}_j(\mathbf{F}_i^{q\top}\mathbf{F}_j^k). \tag{2.4}$$

**Spatial-Temporal Attention** Our spatial-temporal attention model is distinctively designed to correlate features that are not only geometrically meaningful but also spatially and temporally coherent. This model operates in two phases: initially applying spatial attention to capture geometrically consistent parts of a scene, followed by temporal attention to maintain correlations across successive frames. Fig. 2.4 illustrates this dual attention mechanism, showcasing how spatial and temporal attentions operate for a specific pixel across different frames. In the spatial attention phase, the focus is on aggregating features from geometrically consistent parts of the scene. This is particularly noticeable at the boundaries of objects, where significant attention gradients are often directed towards the background, thereby enhancing geometric consistency. Conversely, temporal attention is oriented towards correlating global information across time. In a naive implementation, this task could introduce challenges and inaccuracies without adequately representing spatial relations in 3D. However, our spatial-temporal attention model mitigates these challenges by incorporating geometric constraints and specific loss formulations, as detailed later and expressed in Eq. 2.13. These additions refine the attention mechanism, ensuring it remains sharply focused and preserves spatial coherence. This is particularly beneficial in handling complex scenarios, such as scenes with thin structures or dynamic objects, where maintaining geometric consistency and temporal correlation is crucial. The integration of spatial and temporal attention in our proposed model thus facilitates a more accurate and coherent understanding of the scene, demonstrating the effectiveness of our approach in addressing the intricacies of dense depth regression in dynamic and geometrically diverse environments.



Spatial Attention

Naive Temporal Attention

Temporal Attention with Geometric Guidance

Fig. 2.4 **Attention Visualisation:** Spatial and temporal attention for a queried pixel (indicated by a cross) between frames: The boundary of the spatial attention aligns with the scene's structure. In contrast, the appearance-based naive temporal attention appears non-specific. Our spatially-aware temporal attention centers on visually similar features, guided by geometric reasoning.

## 2.3.2 Enforcing Spatial-Temporal Constraints

### 2.3.2.1 Regularized Geometric Consistency

**Scale-invariant Consistent Depth Loss** Depth values, when projected into another camera view, are usually affected by the relative pose between the camera views. If the depth or disparity value consistency between these views is directly constrained, the overall scene depth scale can shrink or enlarge. To address these inconsistencies, several scale-invariant formulations have been proposed [9, 105, 193]. While mitigating scale changes between frames, these methods tend to yield suboptimal gradients for depth values characterized by minor alignment discrepancies, especially for close-range depth values. Thus, a more nuanced approach that provides strong gradients, especially for depth values with minor misalignments, is needed. Kopf et al. [87] proposed a formulation that bridges this gap. They offer a method that retains the benefits of scale-invariant formulations while penalizing depth inconsistencies. By integrating additional regularization, as detailed in Eq. 2.13, our approach ensures depth predictions are both consistent and accurate across frames. This regularized approach ensures better depth stability across frames while being depth-scale-agnostic, thus making 3D reconstructions more reliable and robust.

**Cycle-Mask from Photometric Consistency** Estimating depth from monocular images presents multiple challenges, one is dealing with the inherent ambiguity and inconsistencies arising from the scene's structure while the camera is moving. Aggregating geometric loss on a pixel-wise basis over different views can be problematic due to the dynamics of the camera. For instance, occlusions can lead to certain regions of the scene being visible in one frame but hidden in another. When these occluded regions contribute to the loss computation, the resultant depth maps often suffer from blurred boundaries and a noticeable decrease in depth accuracy [9].

Recognizing the issues arising from directly aggregating the pixel-wise mean geometric loss, researchers have explored alternative strategies. One such method is computing the pixel-wise minimum depth error [43, 189], similar to the pixel-wise minimum reprojection error [51]. While this approach addresses the problems posed by occlusions to some extent, it has some drawbacks. Our quantitative and qualitative evaluations suggest that this strategy can inadequately exclude significant parts of the scene. Specifically, regions exhibiting large inconsistencies due to inaccurate transformations between adjacent depth maps can be masked. Such exclusion is not desired, especially when these regions carry essential information about the scene structure. When the minimum operator is applied, it can sometimes lead to the inadvertent omission of large parts of the scene, which, as visualized in Fig. 2.5, severely harms the training signal.

Considering these challenges, a more sophisticated approach to handle depth inconsistencies and occlusions is required. Therefore, our proposition leans towards exploiting images' inherent properties rather than just the depth values. We introduce a novel masking scheme grounded in the principle of photo-consistency. Leveraging this assumption, we projectively transform the central target image $I_t$ to match the perspective of an adjacent source frame, denoted as $I_{t \to s}$. In a subsequent step, the transformed image is reverted back to the original viewpoint, resulting in $I_{t\to s\to t}$. This two-step transformation ensures that the regions that remain consistent across these transformations are reliable and can be trusted more during the depth estimation process, thus providing a more robust training signal. By incorporating this photo-consistency-based masking scheme, our method aims to produce depth maps that are both accurate and consistent while adequately handling occlusions and scene inconsistencies.

Our cycle-masking approach is concisely formulated as:

$$\mathscr{M}_{\text{cycle}} = \left[E_{\text{pe}}(I_t, I_{t\to s\to t}) < \gamma\right], \tag{2.5}$$

where $[\cdot]$ is the Iverson bracket. The term $E_{\text{pe}}$ represents the photometric error between the target image $I_t$ and the image that undergoes a two-step transformation: first, it is projected to an adjacent frame and then back to the target's perspective, denoted by $I_{t\to s\to t}$. This photometric error quantifies the visual consistency between these images and is detailed further in Eq. 2.9.

Rather than setting a static threshold for masking, we employ an adaptive thresholding mechanism. Specifically, the threshold $\gamma$ is defined as the value below which 70% of the photometric errors fall, known as percentile, when computed among all pixels of $I_s$, and serves for binarization of the mask. This adaptive approach ensures that the threshold is flexible and can be adjusted based on the scene's characteristics. It adeptly eliminates occluded regions from contributing to the loss and ensures that a majority of the non-occluded regions remain, thereby

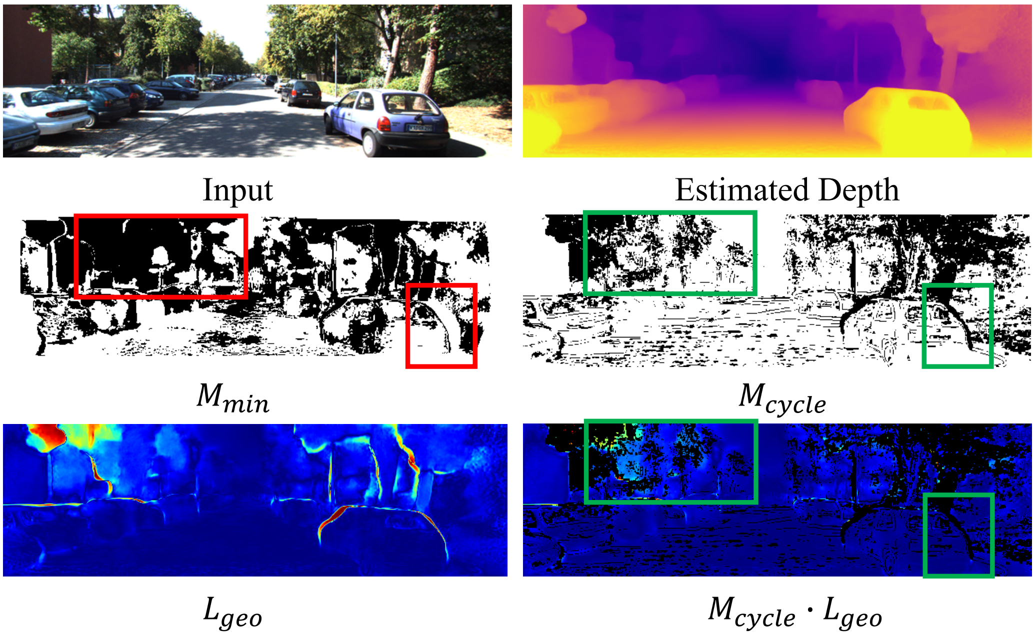


Fig. 2.5 **Visualization of Occlusion Handling:** We illustrate the proficiency of occlusion handling in $\mathscr{L}_{\text{geo}}$ using $\mathscr{M}_{\text{cycle}}$ when compared against a pixel-wise minimum approach. It becomes evident that $\mathscr{M}_{\text{min}}$ fails to effectively address all occlusions, leading to the erroneous masking of extensive image regions. Conversely, $\mathscr{M}_{\text{cycle}}$ demonstrates a superior capability in handling such scenarios. This proficiency ultimately yields improved gradient results during training when deploying $\mathscr{M}_{\text{cycle}} \cdot \mathscr{L}_{\text{geo}}$.

providing a comprehensive and robust consistency check. This strategy's effectiveness and its clear distinction from other methods can be visualized in Fig. 2.5.

#### 2.3.2.2 Loss Formulation

We train our model using a combination of loss terms that are derived from image content reconstruction and the geometric characteristics of our depth map, which can be expressed as:

$$\mathscr{L} = \mathscr{L}_{\text{photo}} + \lambda_{\text{s}}\mathscr{L}_{\text{s}} + \lambda_{\text{geo}}\mathscr{L}_{\text{geo}} + \lambda_{\text{m}}\mathscr{L}_{\text{m}} + \mathscr{L}_{\text{ref}}, \tag{2.6}$$

where $\mathscr{L}_{\text{photo}}$ and $\mathscr{L}_{\text{s}}$ are derived from well-established methodologies as presented in [51, 171]. We will only provide a concise overview of these but will elaborate on the other components in the subsequent sections.

**Motion Consistency Loss $\mathscr{L}_{\text{m}}$** Drawing inspiration from the knowledge distillation approach described in [125], we concurrently train a streamlined self-supervised depth prediction network, specifically MonoDepth2 [51] as referenced in Table 2.1, to act as a weak teacher. In alignment with [171], we introduce a mask. This mask highlights significant discrepancies between our depth prediction $D_t$ and the teacher's $\hat{D}_t$, potentially signaling moving objects. This mask is subsequently employed in the photometric loss computation, defined as:

$$\mathscr{M}_{\text{m}} = \max\left(\frac{D_t - \hat{D}_t}{\hat{D}_t}, \frac{\hat{D}_t - D_t}{D_t}\right) < 0.6. \tag{2.7}$$

This results in our motion consistency loss term, aiding the student in assimilating knowledge from the weak teacher, expressed as:

$$\mathscr{L}_{\text{m}} = (1 - \mathscr{M}_{\text{m}}) \cdot \|D_t - \hat{D}_t\|_1. \tag{2.8}$$

**Photometric Loss $\mathscr{L}_{\text{photo}}$** The photometric reconstruction error [51, 171] between image $I_x$ and $I_y$ is given by:

$$E_{\text{pe}}(I_x, I_y) = \alpha\frac{1-\text{SSIM}(I_x,I_y)}{2} + (1-\alpha)\left\|I_x - I_y\right\|_1, \tag{2.9}$$

and is evaluated between the reference frame $I_t$ and every associated source frame $I_s$, where $s \in S$, subsequently selecting the minimum error at each pixel location. An auto-mask accommodates for objects in the scene moving with identical velocity and direction as the camera-ego motion as:

$$\mathscr{M}_{\text{auto}} = \left[\min_{s\in S} E_{\text{pe}}(I_t, I_{s\to t}) < \min_{s\in S} E_{\text{pe}}(I_t, I_s)\right]. \tag{2.10}$$

$\mathscr{L}_{\text{photo}}$ is finally defined over $S \in \{t-1, t+1\}$ as

$$\mathscr{L}_{\text{photo}} = \mathscr{M}_{\text{m}} \cdot \mathscr{M}_{\text{auto}} \cdot \min_{s\in S} E_{\text{pe}}(I_t, I_{s\to t}). \tag{2.11}$$

**Edge-aware Smoothness Loss $\mathscr{L}_s$** Similar to earlier studies [50, 51], we employ edge-aware smoothness to promote depth predictions that are smooth in local regions, utilizing the mean-normalized inverse depth $\overline{d_t}$, as:

$$\mathscr{L}_s = \left|\partial_x \overline{d_t}\right| e^{-|\partial_x I_t|} + \left|\partial_y \overline{d_t}\right| e^{-|\partial_y I_t|}. \tag{2.12}$$

**Geometric Loss $\mathscr{L}_{geo}$** In light of the discussion and motivation presented before, we propose a geometric loss that aims to maintain depth prediction consistency between frames. Besides mitigating the challenges of depth scale penalization, this loss also integrates cycle consistency (Eq. 2.5) to adeptly address occlusions as:

$$\mathscr{L}_{\text{geo}} = \mathscr{M}_{\text{m}} \cdot \mathscr{M}_{\text{auto}} \cdot \mathscr{M}_{\text{cycle}} \cdot \Big(1 - \frac{\min(D_{s\to t}, D_t')}{\max(D_{s\to t}, D_t')}\Big), \tag{2.13}$$

where $D_{s\to t}$ represents the depth map transformed from the neighboring source frame to the target frame, and $D_t'$ signifies the interpolated depth map of the target as presented in [9, 43].

**Reference Loss $\mathscr{L}_{ref}$** In order to train the single-stage auxiliary depth decoder $D_{ref}$ for spatial attention acquisition, we aim to reduce the discrepancy between it and the (detached) final depth prediction from our complete pipeline, denoted as $D_t$:

$$\mathscr{L}_{\text{ref}} = \left\|D_t - D_{ref}\right\|_1. \tag{2.14}$$

## 2.4 Quantifying Monocular Depth Consistency

In depth prediction from monocular video, particularly in dynamic outdoor driving environments, a key challenge is maintaining consistency across temporally adjacent frames. To address and quantify this, we propose a method for directly measuring consistency in the predicted depth output. This is achieved by aligning a set of $k$ frames in 3D through projective transformation. This approach ensures a consistent reference scale across all comparisons rather than a per-frame scale to account for quantifying scale drift.

Our Temporal Consistency Metric (TCM) is designed to measure the discrepancy between estimated pixel-wise depth and ground truth across multiple frames. A visual representation of TCM is provided in Fig. 2.6. This metric is beneficial for evaluating the performance of depth estimation methods in terms of their temporal consistency. Given artifacts by the sensor, the potential for errors arising from interpolated ground-truth LiDAR data, and the presence of moving objects, we implement a filtering mechanism to enhance the robustness of our metric [103]. Specifically, we discard the 20% most significant outliers in our analysis. This step is crucial for ensuring that our comparisons remain fair and unbiased, especially considering the inherent challenges and noise in the datasets typically used for outdoor driving scenarios. Applying this filtering mechanism allows us to assess the temporal consistency of depth predic-

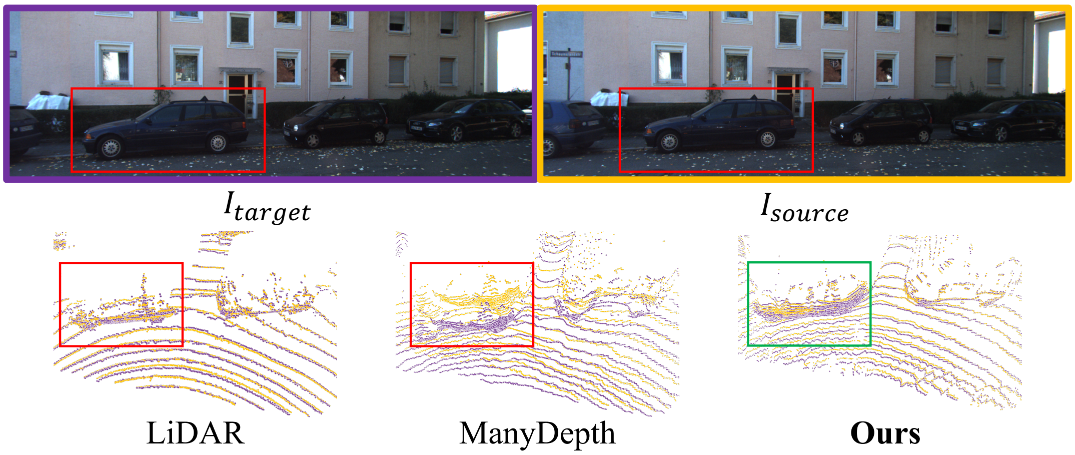


**Fig. 2.6** **Illustration of Depth Consistency:** Successive depth estimations are aligned in a 3D space and assessed on a pixel-wise basis over several frames. With our approach, the alignment of the car across frames is markedly improved.

tions more accurately, providing a clearer understanding of the performance of various depth estimation methods in dynamic environments.

### 2.4.1 Temporal Consistency Metric (TCM)

The Temporal Consistency Metric (TCM) is designed to assess the consistency of depth predictions across consecutive frames in monocular image sequences. Traditional standard accuracy metrics, which compare predicted depths against ground truth on a per-frame basis, fail to capture this aspect of temporal consistency. This is primarily due to their reliance on aligning and evaluating each frame individually.

To evaluate temporal consistency, we focus on the alignment of multiple consecutive depth predictions within a sequence in 3D. This alignment is facilitated by warping consecutive predictions into the same camera view, using the flow generated by the ground truth depth and pose. Within a short sequence of predictions, we designate the central frame's depth as the target depth $D_t$, and the depths of other frames in the sequence as source depths $D_s$. The length of these short sequences is selected to be $k = \{3, 5, 7\}$. We opt for these specific sequence lengths based on the observation that longer sequences are generally not suitable for outdoor driving scenarios. This is due to the reduced visual overlap between images in such contexts, especially considering the typical forward motion speed and frame rate (e.g., $10fps$ in the KITTI dataset [48]). By using TCM, we can more effectively quantify the degree of consistency in depth predictions over time, providing a more comprehensive assessment of performance in dynamic environments. This metric is particularly valuable in scenarios such as autonomous driving, where understanding the temporal evolution of the scene is crucial for safe and accurate navigation.

We define a new terminology of *track*. Conceptually, *track* signifies the point-wise Euclidean distance between the depth of a target frame and the depths of its corresponding source frames in a 3D space, post their alignment within the same camera viewpoint. Mathematically, it can be represented as:

$$track = \left\| T_{t \longrightarrow s} \pi^{-1}(D_t) - \pi^{-1}(D_s^{'}) \right\|_2, \tag{2.15}$$

where:

- $D_t$ stands for the depth originating from the target frame.
- $T_{t \longrightarrow s}$ denotes the ground-truth pose between the target and source frames.
- $D_s^{'}$ is the depth from the source frame, transformed and aligned through the warping flow, derived from both ground truth pose and depth.
- $\pi^{-1}(\cdot)$ symbolizes the projective transformation function for 3D lifting.

For both the ground-truth and predicted depths, we compute the term *track*, leading to $track^{GT}$ and $track^{pred}$. It is crucial to note that for monocular techniques affected by scale ambiguity, every frame in a sequence is uniformly scaled with the same common scaling ratio. This scale is determined by median-aligning the target frame with its respective ground truth. In conclusion, with the obtained *track*, we define metrics such as absolute error ($TCM_{\text{abs}}$), square relative error ($TCM_{\text{sq}}$), and root mean square error ($TCM_{\text{RMSE}}$) to assess the depth consistency across inputs:

$$TCM_{\text{abs}} = \frac{1}{H} \sum_{j=1}^{H} \left| track_j^{GT} - track_j^{pred} \right|, \tag{2.16}$$

$$TCM_{\text{sq}} = \frac{1}{H} \sum_{j=1}^{H} \left( track_j^{GT} - track_j^{pred} \right)^2, \tag{2.17}$$

$$TCM_{\text{RMSE}} = \sqrt{\frac{1}{H} \sum_{j=1}^{H} \left( track_j^{GT} - track_j^{pred} \right)^2}. \tag{2.18}$$

Here, $H$ represents the total amount of valid tracks in the current input, after filtering out large outliers, which accounts for 20%. To encapsulate the complete TCM evaluation, we aggregate these metrics by averaging over every input set presented during testing. In essence, TCM metrics encapsulate the error derived from comparing the Euclidean distances of sequential 3D predictions against the analogous distances of their corresponding ground truths, all post the necessary camera view alignment.

## 2.5 Experimental Results

Our model's performance is assessed against recent state-of-the-art methods, both in terms of our introduced Temporal Consistency Metric (TCM) and well established depth accuracy benchmarks [51]. Aligning with past self-supervised depth estimation research [51, 171], our experiments extensively utilize the Eigen split [34] of the Kitti dataset [48], and we also present findings on Cityscapes [23].

### 2.5.1 Depth Accuracy

The depth accuracy results are presented in Tab. 2.1. In comparison to notable self-supervised models like MonoDepth2 [51], our model showcases superior performance. It even surpasses models with more extensive backbones such as FeatDepth [142], those that incorporate consistency constraints like SC-SfMLearner [9], and semi-supervised strategies like PackNet-SfM [53]. Incorporating the test time refinement approach (notated as TTR in Table 2.1) from [109], our method distinctly excels over ManyDepth [171]. Moreover, our technique registers the highest accuracy scores on the demanding Cityscapes dataset [23].

Tab. 2.1 **Depth Accuracy:** Performance metrics for self-supervised monocular techniques on the Kitti [48] Eigen test set [34] are presented. The symbol * denotes semi-supervised approaches. In the middle section, we incorporate test time refinement (TTR) [142]. The lower section addresses the Cityscape dataset [23]. The symbols † and ‡ indicate fresh results sourced from GitHub with standard image resolution for fair comparison. Rankings are emphasized as: **best**, 2nd best, and 3rd best.

| Method | | Abs Rel | Sq Rel | RMSE | $\sigma < 1.25$ | $\sigma < 1.25^3$ |
|---|---|---|---|---|---|---|
| Monodepth2 [51] | | 0.115 | 0.903 | 4.863 | 0.877 | 0.981 |
| SC-SfMLearner [9] † | | 0.119 | 0.857 | 4.950 | 0.863 | 0.981 |
| TrianFlow [193] | | 0.113 | **0.704** | 4.581 | 0.871 | **0.984** |
| PackNet-SfM[53]* | | 0.111 | 0.829 | 4.788 | 0.864 | 0.980 |
| FeatDepth[142] ‡ | | 0.109 | 0.923 | 4.819 | 0.886 | 0.981 |
| ManyDepth [171] | | **0.098** | 0.770 | **4.459** | **0.900** | 0.983 |
| **Ours (DRN-C-26)** | | 0.106 | 0.770 | 4.558 | 0.890 | 0.983 |
| **Ours (DRN-D-54)** | | 0.103 | 0.746 | 4.483 | 0.894 | 0.983 |
| ManyDepth[171] | TTR | 0.090 | 0.713 | 4.137 | 0.914 | **0.997** |
| **Ours (DRN-C-26)** | | **0.082** | **0.667** | **4.104** | **0.921** | **0.997** |
| Monodepth2 [51] | CS | 0.129 | 1.569 | 6.876 | 0.849 | 0.983 |
| ManyDepth [171] | CS | 0.114 | 1.193 | 6.223 | **0.875** | 0.989 |
| **Ours (DRN-C-26)** | CS | **0.110** | **0.958** | **5.820** | 0.867 | **0.991** |

### 2.5.2 Depth Consistency

We utilize a subset of Kitti odometry data for our TCM evaluations. Due to the abundance of moving objects, Cityscapes is not considered in the TCM assessment. In the inference stage, we process image triplets, as illustrated in Fig. 2.3, mirroring the approach of ManyDepth [171] that also employs successive images. Unlike ManyDepth [171], our approach does not necessitate predicting relative camera positions between adjacent frames for depth deduction. Comprehensive ablation tests reveal that our model's effectiveness is nearly unchanged, regardless of whether we employ an encoder with dilated convolutions or a traditional ResNet, as used in [51]. This is evident in both depth predictions and accuracy evaluations. The qualitative results further highlight our model's advantages, especially its capability to deliver temporally consistent 3D reconstructions using sequential depth estimations as illustrated in Fig. 2.7).

In Tab. 2.2, we summarize the relative TCM results, evaluating depth consistency across an increasing count of test frames. Notably, **TC-Depth** excels over prominent self-supervised benchmarks such as MonoDepth2 [51]. It also surpasses SC-SfMLearner [9], known for emphasizing temporal consistency, and ManyDepth [171], which distinctively leverages adjacent frames during inference. Furthermore, our approach demonstrates superiority over the semi-supervised technique from PackNet-SfM [53] and even outperforms NeuralRGBD [100], which operates under full supervision and employs GT poses in its testing phase.

Figs. 2.8 and 2.9 provide additional qualitative insights into 3D reconstructions, complementing those in Fig. 2.7. The significance of having temporally coherent depth predictions becomes evident in these reconstructions. While a standalone depth map might not reveal inconsistencies, a composite of depth maps viewed from different perspectives showcases them. For instance, even though ManyDepth [171] sets a high standard in accuracy, it still has issues with object deformations, ghosting effects, and "flying pixels". Similar visual artifacts are noticeable in the semi-supervised PackNet-SfM* [53]. In contrast, our approach consistently

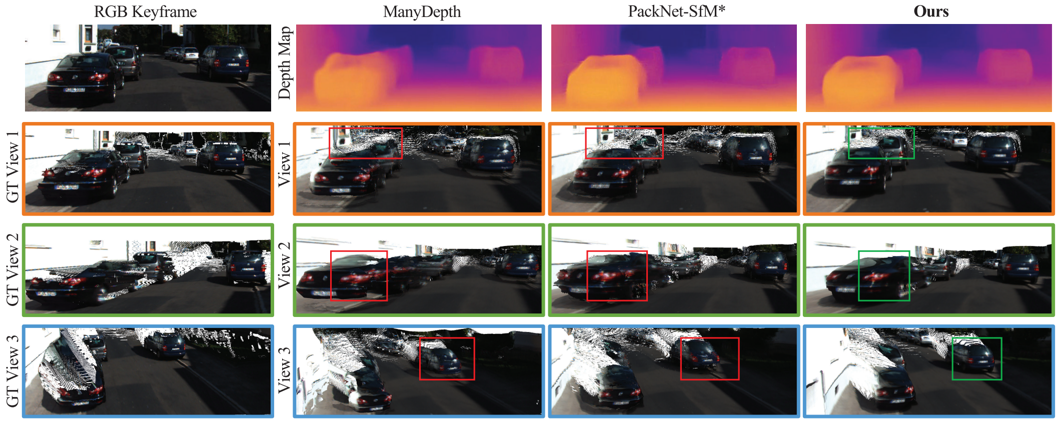


Fig. 2.7 **Qualitative Depth Consistency:** Qualitative reconstructions using five sequential depth predictions reveal distinct differences. Both ManyDepth [171] and PackNet-SfM* [53] with velocity semi-supervision, exhibit artifacts like "flying pixels" in View 1, ghosting effects in View 2, and distorted objects in View 3. These issues stem from temporal inconsistencies. While these may not be evident in single-frame depth predictions, they become prominent when the viewpoint shifts. Our approach substantially reduces these visual discrepancies.

Tab. 2.2 **Depth Consistency:** Evaluation of the Temporal Consistency Metric (TCM) over an increasing count of test frames [3, 5, 7]. *: Incorporates semi-supervision with velocity ground-truth. **: Utilizes supervision with ground-truth depth and conducts inference with ground-truth pose. Our self-supervised approach enhances TCM by approximately 60% across all metrics when contrasted with powerful benchmarks utilizing temporal frames [171]. Moreover, it surpasses semi-supervised [53] and fully supervised [100] methods designed to estimate depth with temporal coherence.

| Method | Abs Err | | | Sq Err | | | RMSE | | |
|---|---|---|---|---|---|---|---|---|---|
| # Test Frames | 3 | 5 | 7 | 3 | 5 | 7 | 3 | 5 | 7 |
| ManyDepth [171] | 0.204 | 0.260 | 0.307 | 0.087 | 0.147 | 0.206 | 0.256 | 0.319 | 0.373 |
| MonoDepth2 [51] | 0.137 | 0.177 | 0.215 | 0.039 | 0.068 | 0.104 | 0.176 | 0.223 | 0.268 |
| SC-SfMLearner [9] | 0.126 | 0.170 | 0.211 | 0.032 | 0.062 | 0.099 | 0.159 | 0.210 | 0.259 |
| PackNet-SfM [53] | 0.141 | 0.196 | 0.247 | 0.044 | 0.090 | 0.147 | 0.177 | 0.240 | 0.299 |
| PackNet-SfM [53]* | 0.118 | 0.154 | 0.190 | 0.030 | 0.052 | 0.083 | 0.154 | 0.197 | 0.240 |
| NeuralRGBD [100]** | 0.116 | 0.148 | 0.179 | 0.024 | 0.044 | 0.066 | 0.147 | 0.186 | 0.222 |
| **Ours DRN-C-26** | **0.079** | **0.111** | **0.147** | **0.011** | **0.025** | **0.047** | **0.099** | **0.139** | **0.184** |
| **Ours DRN-D-54** | **0.076** | **0.106** | **0.138** | **0.010** | **0.022** | **0.041** | **0.095** | **0.131** | **0.172** |

offers the most accurate reconstructions across sequential temporal depth maps and their 3D reconstructions.

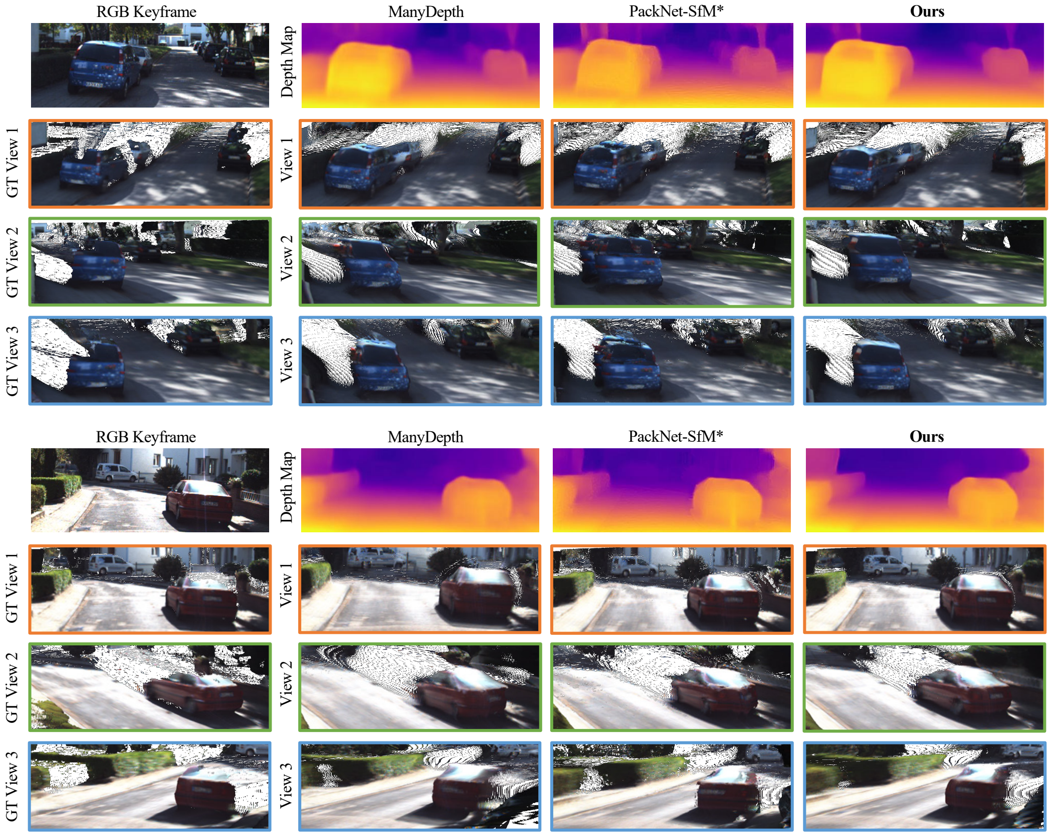


Fig. 2.8 **Qualitative Depth Consistency 2:** Additional qualitative reconstructions using five sequential depth predictions reveal distinct differences (cf. Fig. 2.7).

In-depth statistical analyses of the absolute TCM metric for various sequence lengths $k$ are available in Figs. 2.10, 2.11, and 2.12 (left). Across all sequence lengths $k = \{3, 5, 7\}$, our

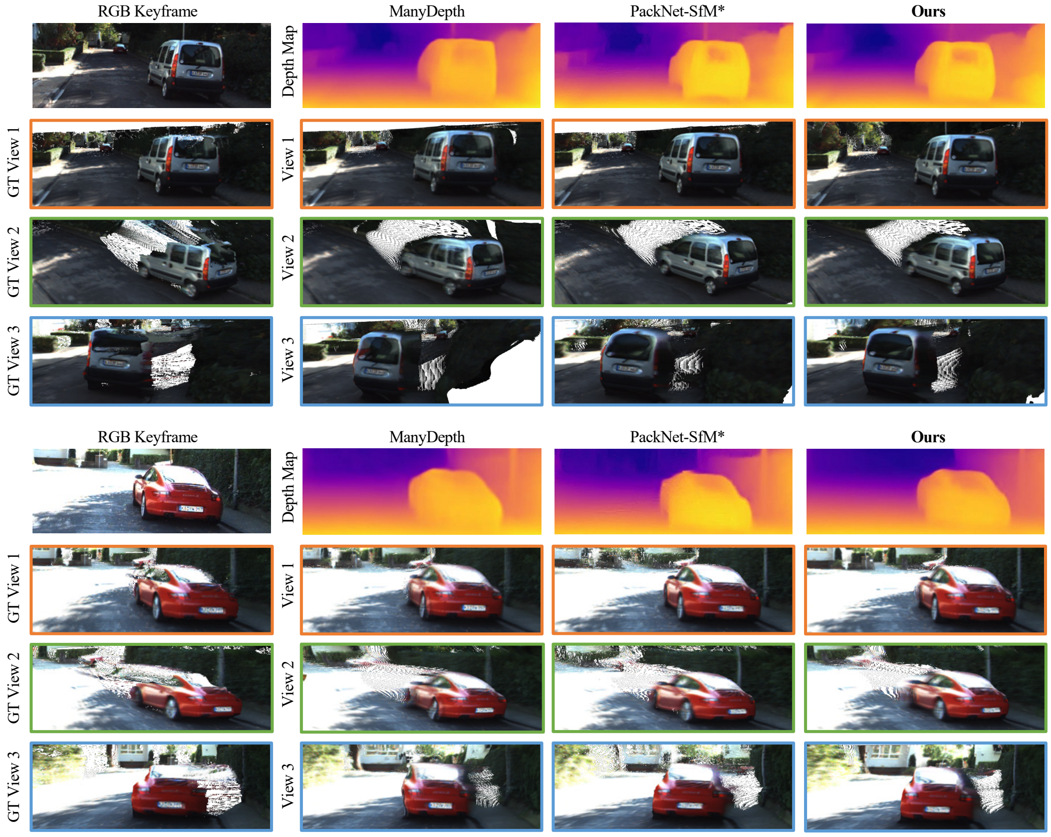


Fig. 2.9 **Qualitative Depth Consistency 3:** Additional qualitative reconstructions using five sequential depth predictions reveal distinct differences (cf. Fig. 2.7).

approach consistently showcases the lowest mean and median absolute TCM errors, along with fewer outliers.

To ensure fair comparisons between methodologies and taking into account errors from interpolated ground-truth LiDAR and moving objects, we set a threshold to eliminate the top 20% of outliers. Here, we additionally provide TCM results for different outlier sampling rates for detailed analysis in Figs. 2.10, 2.11, and 2.12 (right). Our method remains consistently superior across these various sampling rates compared to other methods.

### 2.5.3 Ablation Studies

For a quantitative assessment of the impact of each component in our pipeline, we conduct a thorough ablation study, presenting both TCM outcomes and depth accuracy as shown in Tab. 2.3. The selection of the backbone, whether ResNet18 as in MonoDepth2 (MD2) [51] or DRN-C-26 in our baseline, exerts only a minimal influence on accuracy and TCM.

Our ablation study offers insightful revelations about the impact of spatial-temporal attention (ST-A) on both accuracy and the Temporal Consistency Metric (TCM) results. It is observed that while spatial attention (S-A) enhances accuracy, it exhibits negligible influence on TCM, as

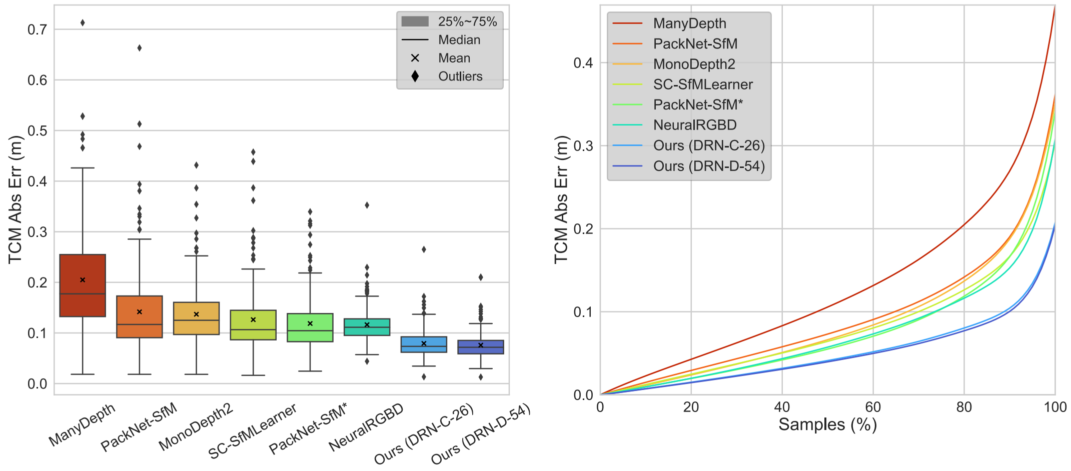


**Fig. 2.10** **TCM Statistics:** Detailed statistical analysis using 3-frame-TCM. On the left, we present the distribution of absolute errors from a 3-frame TCM test. The right side illustrates the absolute errors of TCM evaluations, taking into account varying sampling rates for outlier handling.

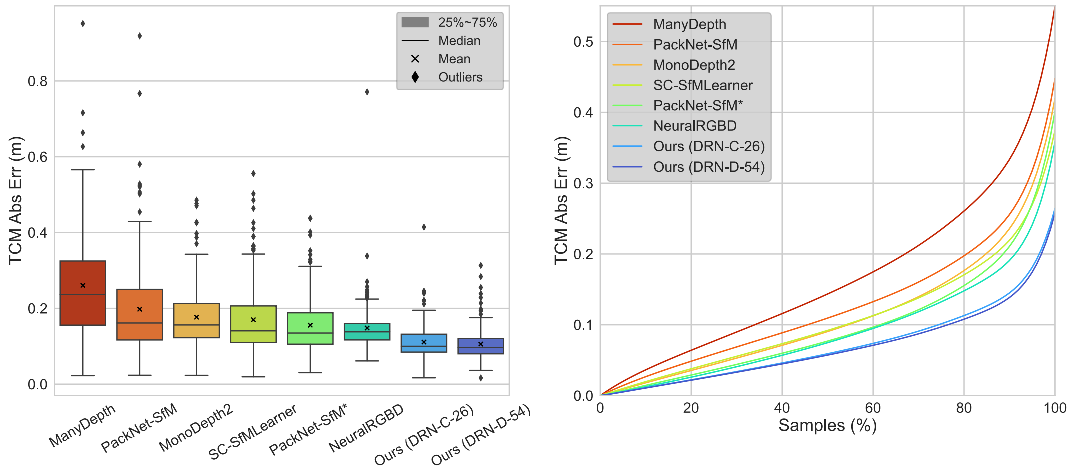


**Fig. 2.11** **TCM Statistics:** Detailed statistical analysis using 5-frame-TCM. On the left, we present the distribution of absolute errors from a 5-frame TCM test. The right side illustrates the absolute errors of TCM evaluations, taking into account varying sampling rates for outlier handling.

indicated in Tab. 2.3. On the other hand, temporal attention (T-A) alone does not improve TCM and can even detrimentally affect accuracy. This is attributed to the potential for highly noisy and imprecise feature aggregation in the absence of positional information [171]. To address this, we introduce S-A with correlated 3D information, serving as a form of 3D positional encoding. This addition ensures that the temporal feature aggregation in T-A is spatially aware, thereby preventing degradation in accuracy.

The consistency loss $\mathscr{L}_\mathrm{m}$ plays a pivotal role in enforcing congruence between the weak teacher's predictions and our model's predictions, particularly in regions with significant deviations (i.e., $1-\mathscr{M}_\mathrm{m}$), such as those caused by e.g. moving objects. Employing only $\mathscr{L}_\mathrm{m}$ in training (without

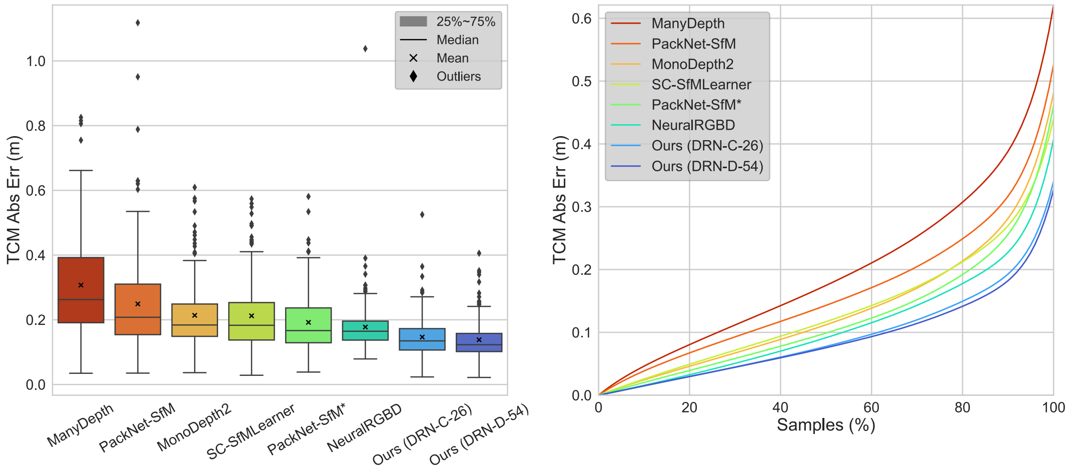


**Fig. 2.12** **TCM Statistics:** Detailed statistical analysis using 7-frame-TCM. On the left, we present the distribution of absolute errors from a 7-frame TCM test. The right side illustrates the absolute errors of TCM evaluations, taking into account varying sampling rates for outlier handling.

geometric guidance) already facilitates the identification of regions with large deviations, thereby improving accuracy. This is in line with the findings of Watson et al. [171].

From the ablation results, we discern that the geometric loss $\mathscr{L}_{\text{geo}}$ is critical for achieving highly consistent depth predictions. Its effectiveness in respecting occlusions is complemented by the cycle mask $\mathscr{M}_{\text{cycle}}$, while potential dynamic objects are filtered using $\mathscr{M}_{\text{m}} \cdot \mathscr{M}_{\text{auto}}$. This comprehensive masking ensures that dynamic objects, which violate the static assumption inherent in $\mathscr{L}_{\text{geo}}$, are accurately accounted for, thereby enhancing consistency performance.

Interestingly, using $\mathscr{L}_{\text{geo}}$ with only $\mathscr{M}_{\text{min}}$ marginally reduces accuracy for the $\sigma < 1.25$ accuracy measure, in line with observations from SC-SfMLearner [9]. However, the addition of $\mathscr{M}_{\text{cycle}}$ mitigates this issue by better handling occluded regions through photometric cues. The combination of $\mathscr{L}_{\text{geo}}$ and $\mathscr{M}_{\text{cycle}}$ significantly improves TCM, with $\mathscr{L}_{\text{m}}$ further reducing outlier rates as indicated by the Sq.Rel. error metric.

Integrating spatial-temporal attention with $\mathscr{L}_{\text{geo}}$ and $\mathscr{M}_{\text{cycle}}$ allows the additional loss function and appropriate regularization to guide the attention module towards more geometrically consistent aggregation of temporal information. This integration markedly enhances both TCM and depth accuracy. The complete model, incorporating all these elements, achieves the best results. Moreover, employing a larger encoder can further improve these outcomes.

Qualitative results in Fig. 2.13 validate our findings. The baseline model, without our contributions, exhibits compromised 3D reconstruction capabilities. In contrast, incorporating our proposed geometric constraints and the novel spatial-temporal attention module independently and collectively enhances the quality of 3D reconstructions.

Tab. 2.3 **Ablation Study:** We conduct an ablation study on depth accuracy and consistency (TCM). Individually enabling individual components of our pipeline enhances the overall performance. The full model with all components achieves the best results.

| Model | Ablations | | | | Accuracy | | | | | | | TCM (3 Frames) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\mathscr{L}_{geo}$ | $\mathscr{M}_{cycle}$ | Attention | $\mathscr{L}_{m}$ | Abs Rel | Sq Rel | RMSE | RMSE log | $\sigma < 1.25$ | $\sigma < 1.25^2$ | $\sigma < 1.25^3$ | Abs Err | Sq Err | RMSE |
| MD2 [51] | | | | | 0.115 | 0.903 | 4.863 | 0.193 | 0.877 | 0.959 | 0.981 | 0.137 | 0.039 | 0.176 |
| DRN-C-26 | | | | | 0.115 | 1.027 | 5.004 | 0.197 | 0.879 | 0.958 | 0.979 | 0.136 | 0.039 | 0.173 |
| | ✓ | | | | 0.113 | 0.904 | 4.773 | 0.193 | 0.877 | 0.959 | 0.980 | 0.124 | 0.032 | 0.157 |
| | ✓ | ✓ | | | 0.111 | 0.878 | 4.761 | 0.190 | 0.882 | 0.961 | 0.981 | 0.113 | 0.026 | 0.141 |
| | | | S-A | | 0.113 | 0.958 | 4.861 | 0.192 | 0.882 | 0.960 | 0.980 | 0.134 | 0.038 | 0.172 |
| | | | T-A | | 0.116 | 1.028 | 5.024 | 0.197 | 0.879 | 0.957 | 0.979 | 0.133 | 0.037 | 0.171 |
| | | | ST-A | | 0.112 | 0.974 | 4.921 | 0.194 | 0.882 | 0.960 | 0.980 | 0.130 | 0.035 | 0.165 |
| | | | | ✓ | 0.112 | 0.840 | 4.683 | 0.189 | 0.880 | 0.961 | 0.982 | 0.132 | 0.036 | 0.169 |
| | ✓ | ✓ | ST-A | | 0.108 | 0.819 | 4.655 | 0.186 | 0.886 | 0.962 | 0.982 | 0.105 | 0.022 | 0.133 |
| | ✓ | ✓ | ST-A | ✓ | **0.106** | **0.770** | **4.558** | **0.182** | **0.890** | **0.964** | **0.983** | **0.079** | **0.011** | **0.099** |
| DRN-D-54 | ✓ | ✓ | ST-A | ✓ | 0.103 | 0.746 | 4.483 | 0.180 | 0.894 | 0.965 | 0.983 | 0.076 | 0.010 | 0.095 |

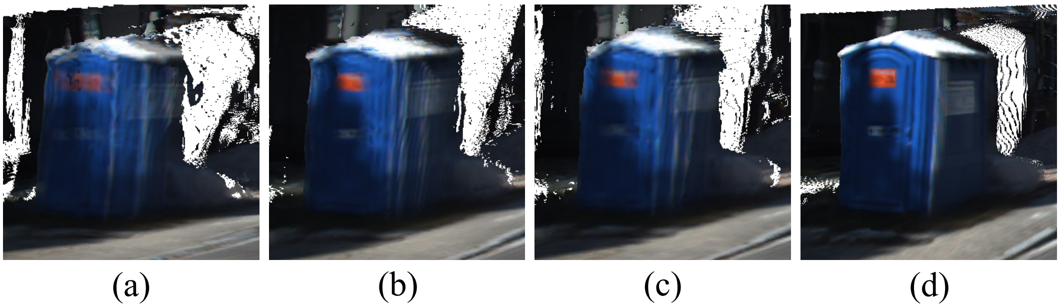


Fig. 2.13 **Qualitative Ablation Results:** The baseline model without our contributions (a) exhibits pronounced ghosting effects resulting from incorrect pixel-wise alignment. By integrating the constraint $\mathscr{M}_{\text{cycle}} \cdot \mathscr{L}_{\text{geo}}$ (b) or implementing spatial-temporal attention (c), such issues are substantially reduced. Employing our complete model (d) delivers the most refined output.

#### 2.5.3.1 Static Camera Performance

We feed just a single static image into our methodology to simulate a stationary camera scenario without sequential images and scene content changes. In this case, we exclusively report on accuracy metrics, as employing TCM metrics would be irrelevant in a scenario with a non-moving camera. Given that ManyDepth [171] also capitalizes on sequential frame inputs, we designate this method as our reference. Tab. 2.4 summarizes the accuracy results. Although our approach with a single static frame input yields slightly suboptimal outcomes in contrast to those with temporal images, the degradation in accuracy is less pronounced, as observed with ManyDepth [171].

#### 2.5.3.2 Attention Mechanism

In Fig. 2.14 (top), we can observe that the spatial attention ball-query effectively correlates spatially proximate structures. In contrast, the temporal attention cannot always yield a single distinct maximum attention value for the queried pixel. This is particularly evident when multiple non-identical objects with similar appearances are present, resulting in ambiguous attention patterns, such as when there are multiple similarly-looking pedestrians or cars. Furthermore, it is worth noting that the temporal attention predominantly correlates objects within a close depth layer while potentially disregarding similar objects located at a greater

Tab. 2.4 **Emulating a Stationary Camera Setting:** Accuracy outcomes on the Kitti Eigen test split [34] contrasting standard temporal frame inputs with a singular static frame input.

| Method | Test-time input | Abs Rel | Sq Rel | RMSE | RMSE log | $\sigma < 1.25$ | $\sigma < 1.25^2$ | $\sigma < 1.25^3$ |
|---|---|---|---|---|---|---|---|---|
| ManyDepth [171] | Temporal Frames (Standard) | **0.098** | 0.770 | **4.459** | **0.176** | **0.900** | **0.965** | **0.983** |
| **Ours (DRN-C-26)** | Temporal Frames (Standard) | 0.106 | 0.770 | 4.558 | 0.182 | 0.890 | 0.964 | **0.983** |
| **Ours (DRN-D-54)** | Temporal Frames (Standard) | 0.103 | **0.746** | 4.483 | 0.180 | 0.894 | **0.965** | **0.983** |
| ManyDepth [171] | Single Frame (Static) | 0.117 | 0.886 | 4.754 | 0.191 | 0.872 | 0.959 | 0.982 |
| **Ours (DRN-C-26)** | Single Frame (Static) | 0.107 | 0.784 | 4.596 | 0.184 | 0.888 | 0.963 | **0.983** |
| **Ours (DRN-D-54)** | Single Frame (Static) | **0.104** | **0.760** | **4.515** | **0.181** | **0.982** | **0.964** | **0.983** |

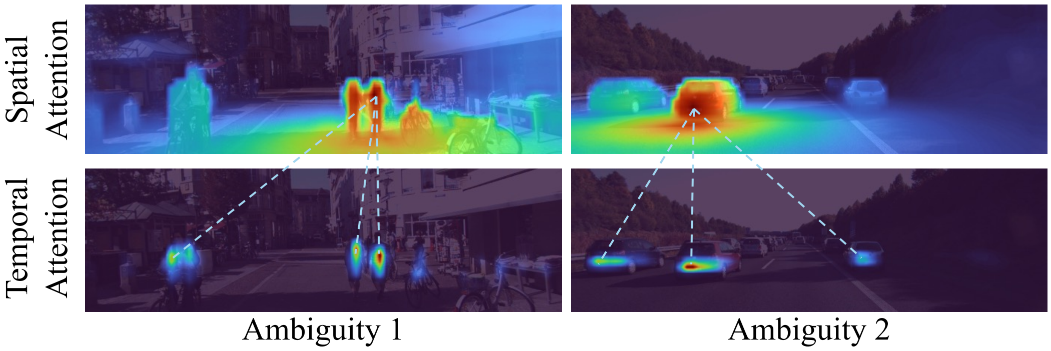


Fig. 2.14 **Limitations of Attention:** Visualization of spatial and temporal attention mechanisms in a challenging scene containing multiple similar objects.

distance (e.g., cars in the background). This observation aligns with our initial hypothesis that spatial attention and geometric constraints are crucial in guiding temporal attention toward geometry-aware feature aggregation for enhanced consistency.

## 2.6 Conclusion

This chapter introduced **TC-Depth**, a novel self-supervised monocular depth estimation approach focusing on accuracy and temporal consistency for robust 3D perception. It features a spatial-temporal attention mechanism with geometric guidance, enhancing depth prediction robustness and accuracy in challenging environments and achieving superior consistency of predictions across temporal frames, as validated by the introduced Temporal Consistency Metric (TCM).

**TC-Depth** demonstrates remarkable robustness against various camera movements and environmental conditions, benefiting from its adaptive aggregation of spatial-temporal features. Ablation studies highlighted the importance of each component, particularly the consistency loss for geometric guidance and spatial-temporal attention. The results and experiments revealed **TC-Depth**'s superior performance in maintaining depth consistency, particularly in dynamic environments, and its ability to maintain or improve depth accuracy. The spatial-temporal attention mechanism significantly boosts prediction accuracy and temporal coherence. Moreover, the model's robustness in stationary camera scenarios and the integration of geometric loss functions and cycle masks further improve predictions. Qualitative evaluations, particularly in 3D reconstructions from multiple fused predicted depth maps, highlight **TC-Depth**'s practical superiority, showcasing reduced artifacts and enhanced depth prediction quality.

# Part III

## Sensor Characteristics and Dense 3D Perception

**How can the characteristics of active depth sensors and their associated artifacts be systematically analyzed, also considering challenging scenarios?**

3

# Sensor Characteristics and Dense 3D Perception

*“Geometry existed before the creation”*

**– Plato**

## 3.1 Introduction

The previous chapter proposed a novel self-supervised monocular depth estimation technique, focusing on a temporally consistent depth estimation model and the Temporal Consistency Metric (TCM). These advancements enhanced the robustness and temporal stability of depth predictions and allowed objectively quantifying the temporal depth consistency across consecutive frames. However, noisy depth sensor data, used for ground-truth generation in established benchmarks like the ones used in the previous chapter, forces the mask out of the data with the largest outliers from the evaluation with TCM due to errors for moving objects or other artifacts from the sensor [103]. This motivates the research question of how accurate the depth ground truth data is and which kind of sensor-specific artifacts lead to specific noisy or corrupt depth measurements, especially when we consider challenging cases with reflections, textureless surfaces, or transparent material.

In response to the need for high-quality depth data in 3D perception, this section delves into creating a comprehensive multi-modal dataset. This dataset, integrating a variety of depth sensors and a polarization camera, is designed to analyze the challenges posed by sensor-specific artifacts in depth estimation. Later, this will lead to exciting observations for other 3D perception tasks like 6D object pose estimation, where corrupt depth estimates can lead to noisy and incorrect predictions. However, the passively captured polarization of light yields valuable properties that can be leveraged and integrated for such 3D perception tasks.

First, we critically examine and provide a comprehensive analysis of different depth sensors, comparing their performance against dense depth ground truth and discussing their impact as supervision signals for depth estimation and reconstruction tasks in dense 3D vision. After analyzing the sensors, we also extend the spatial-temporal conditions in the previous chapter to a multi-view scenario with abundant camera views for analyzing dense 3D vision tasks, like depth estimation or novel view synthesis through implicit scene representations. For both tasks, we can rely only on RGB-only information as before or integrate supervision signals from the depth sensors. We can then analyze the results in detail, compare them against the sensor depth itself and accurate ground truth, and thus get an understanding of each sensor. This detailed understanding of sensor characteristics and how they influence dense 3D vision tasks is essential for further developing accurate and robust 3D perception methods, especially for photometrically challenging scenarios.

### 3.1.1 Motivation

In our three-dimensional world, accurate distance measurements enable machines to understand and interact with the environment spatially. This necessity is particularly pronounced in applications like autonomous vehicles [49, 86, 132, 153], robot vision systems for 6D object pose estimation and manipulation [35, 168, 188], and augmented reality (AR) for enhanced realism [14, 87]. The field of computer vision benefits from a diverse range of sensor modalities and depth prediction pipelines supported by publicly available datasets [49, 139, 145, 152, 161, 167, 175]. These resources allow for comprehensive evaluation of depth estimation methods. Each sensor type used for ground truth depth mapping has unique advantages and limitations. However, often, pipelines are trained without fully considering the characteristics and confidence levels of these sensors.

Popular passive sensor setups, such as multi-view stereo cameras, rely on photometric or structural associations to triangulate points within their field of view [139]. However, these methods falter in low-textured areas or under poor lighting conditions. Active sensing, including active stereo and Time-of-Flight (ToF) technologies (both D-ToF and I-ToF) [54], addresses these limitations. Despite their advantages, active sensors can produce artifacts like multi-reflection errors, especially with reflective and translucent materials. LiDAR sensors, often providing sparse distance measurements, are another alternative, often used in outdoor driving benchmarks like KITTI [49]. Radar offers a more affordable, albeit sparser, solution [47]. Fusing multiple modalities can enhance distance estimates, but challenges arise in aligning data to a common reference frame [77, 103]. Multi-modal approaches have improved monocular depth estimation using self-supervision from stereo and temporal cues [161], yet their performance analysis is typically limited by the sensors used.

The first part of this chapter details the dataset acquisition process, emphasizing the novel methods and procedures employed to attain high accuracy in our data. Following this, we conduct an in-depth analysis of different depth sensors, comparing their performance against dense depth ground truth. This comparison is crucial in understanding each sensor's unique characteristics and limitations. Subsequently, we train depth prediction and view synthesis methods using different supervision signals. We aim to extract maximum insights into how these methods perform under varied conditions by focusing on these dense tasks. While our dataset also supports the analysis of other 3D perception tasks like 6D object pose estimation, we initially concentrate on dense tasks for a more detailed understanding. Our findings reveal that depth sensors often exhibit specific artifacts, particularly when dealing with photometrically challenging objects. In the subsequent chapters, we also delve into how polarization information from our multi-modal dataset can be effectively incorporated into other learning-based 3D perception systems, like 6D object pose estimation, to enhance the accuracy and robustness, especially in scenarios involving challenging photometric conditions.

### 3.1.2 Contributions

Our research aims to address key questions regarding the drawbacks of current depth-sensing modalities and their impact on training pipelines for 3D perception tasks. To facilitate this analysis, we provide multi-modal sensor data alongside highly accurate annotated depth, enabling the examination of popular depth estimation, novel view synthesis, and 3D reconstruction methods on diverse photometric complexities and material properties (see Fig. 3.1).

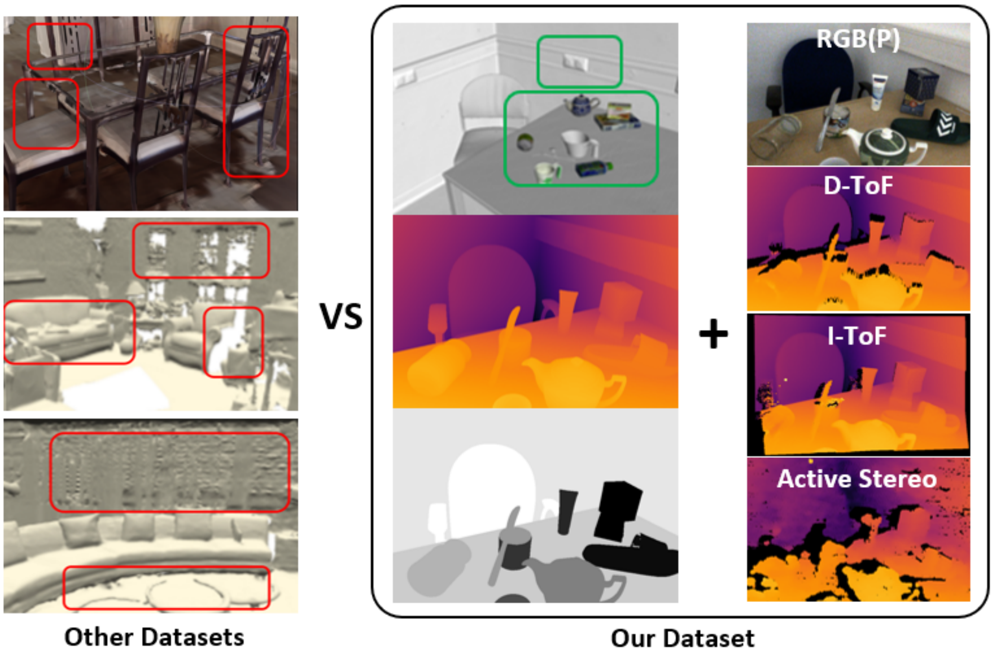


Fig. 3.1 **Dataset Comparison:** While existing datasets for dense 3D vision tasks typically reconstruct scenes in a single pass, leading to compromised quality and accuracy as highlighted by the red boxes in examples from Replica [151], ScanNet [27], and Matterport3D [20], our dataset takes a different approach. We scan each object in the scene and the background separately before accurately annotating them to create dense, high-quality 3D meshes. Combined with precise camera extrinsics obtained from robotic forward-kinematics, our dataset provides fully dense rendered depth maps that serve as highly accurate pixel-wise ground truth. The inclusion of multimodal sensor data, including RGB with polarization, D-ToF, I-ToF, and Active Stereo, makes our dataset a comprehensive tool for quantifying performance across various downstream 3D vision tasks, such as monocular depth estimation, novel view synthesis, and 6D object pose estimation.

The primary contributions in this chapter are in summary:

## Contributions

1. Analyzing the **measurement quality of various depth sensor modalities** and their impact as supervision signals for 3D vision tasks.
2. Investigating the **performance** on materials with varying textures and **photometrically challenging** areas, such as reflective, translucent, and transparent surfaces.
3. Providing a **comprehensive indoor dataset** that combines **multi-modal sensors** (I-ToF, D-ToF, monocular RGB+P, monochrome stereo, and active light stereo) with highly accurate ground truth, for evaluating and quantifying the impact of different **sensor characteristics on 3D perception**.

# 3.2 Related Work

## 3.2.1 Geometry from X

A variety of sensor modalities have been utilized to acquire depth maps. Typically, datasets consist of a single ground truth sensor used consistently for all acquisitions, and it is assumed to provide accurate enough data for model training, testing, and validation.

### 3.2.1.1 Stereo Vision

In stereo vision, initial methodologies [139] primarily utilized a pair of passive cameras and focused on scenes with piecewise planar objects to facilitate triangulation. These early approaches laid the groundwork for understanding stereo depth perception but were limited in handling complex scenes. Complex setups involving industrial robots and structured light have been employed for more precise ground truth depth in stereo images [1]. Additionally, robots have been used to annotate keypoints on transparent household objects, as explored KeyPose [102]. These methods, however, face challenges in textureless areas where stereo matching is less effective. To overcome this, active sensors have been used to project patterns onto scenes, artificially creating structures and enhancing depth retrieval in these challenging regions. The advent of active stereo sensors has also enabled the acquisition of real indoor environments [145]. In such setups, depth data at missing pixels is often inpainted to create a more complete depth map. Structure from Motion (SfM) techniques have been applied to generate depth maps, such as in the Sun3D dataset by Xiao et al. [175]. In this approach, a moving camera captures the scenes, and the data is fused post-capture. Furthermore, a temporally tracked handheld active sensor has been utilized for depth mapping and SLAM evaluation, notably in the pioneering dataset by Sturm et al. [152]. Despite advancing the field, the depth maps in this dataset are limited to the active IR pattern used by the RGB-D sensor. These methods in the stereo literature reflect the evolution and diversification of techniques for depth perception, each with its specific applications, advantages, and limitations. They highlight the ongoing pursuit of more accurate, reliable, and comprehensive depth mapping solutions in various environments.

### 3.2.1.2 Time-of-Flight Sensors

Recent advancements in active depth sensing have increasingly focused on Time-of-Flight (ToF) technology. Early explorations in this area concentrated on simulated data [54], and experiments in controlled environments with minimal ambient noise [148]. The growing presence of ToF sensors in commercial products, such as the Microsoft Kinect series and modern smartphones (including the Indirect Time-of-Flight (I-ToF) in Huawei's P30 Pro and Direct Time-of-Flight (D-ToF) in Apple's iPhone 12), has spurred research aimed at addressing common sensor-related errors. These challenges include multi-path interference (MPI), motion artifacts, and issues of sparsity and shot noise, as discussed by Jung et al. [77]. Given the prevalence and significance of these ToF modalities in contemporary depth sensing, our experimental framework encompasses both classical active and passive stereo techniques and D-ToF and I-ToF technologies. This inclusive approach allows for a comprehensive evaluation and comparison of these modalities, considering their respective strengths and limitations in various depth-sensing scenarios. The inclusion of these diverse modalities in our experiments is crucial for understanding the full spectrum of depth-sensing capabilities and challenges in real-world applications. It also provides insights into how different technologies can be optimized or combined to improve the accuracy and reliability of depth measurements, particularly in environments where traditional stereo vision approaches may be less effective.

#### 3.2.1.3 Polarimetric Cues

In addition to depth sensing, recent research has explored the use of properties of light to infer surface characteristics of scenes, especially for the estimation of surface normals. This is achieved by examining the amount of linearly polarized light and its direction of polarization. Such techniques are particularly informative for surfaces that are highly reflective or transparent [82]. Early research in shape-from-polarization primarily focused on controlled environments [4, 44, 146, 185]. Recent advancements have expanded these methods to more complex scenarios. For instance, sensor fusion methods [81], and applications in environments with strong ambient light, as explored by Verdie et al. [161], demonstrate the versatility and potential of polarization-based techniques. Given these developments, our research also incorporates the acquisition of RGB+Polarization (RGB+P) data for all scenes. This approach allows us to capture the traditional color information of the scenes and gather additional data regarding surface properties through polarization analysis. Integrating RGB+P data enriches our understanding of the scene's geometry and material characteristics, offering a more comprehensive and nuanced view of the environment.

#### 3.2.1.4 Synthetic Renderings

To achieve pixel-perfect ground truth for depth maps, researchers create synthetic scenes [108]. While this approach yields the most precise depth maps possible, it comes with a drawback — synthetic scenes lack realism. Consequently, pipelines trained on datasets like Sintel [17] or SceneFlow [108] face challenges arising from the synthetic-to-real domain gap. In contrast to this purely synthetic approach, our methodology follows a hybrid path. We capitalize on pixel-perfect synthetic data generated using modern 3D engines and combine it with highly accurate 3D models adjusted to real-world captures.

### 3.2.2 Monocular Depth Estimation

When trained with supervision, deep learning networks have demonstrated the capability to predict depth from single images. This development in monocular depth estimation was initially pioneered by Eigen et al. [34], who introduced a two-stage network that predicts coarse depth maps and then refines them with a second network. Following this, Laina et al. [91] enhanced the approach by employing a single CNN composed entirely of convolutional layers, streamlining the depth estimation process.

However, these supervised methods' dependence on ground truth depth data for training is a significant limitation. This requirement often confines the applicability of such methods to outdoor scenarios with available datasets, such as the one provided by Geiger et al. [48]. To circumvent the need for real-world ground truth data, some approaches have utilized synthetic data for training [107]. While effective, these methods often face a domain gap when applied to real-world scenarios, though strategies to narrow this gap have been proposed [54].

MiDaS [128] represents a significant stride in generalizing depth estimation to unknown scenes by leveraging a diverse dataset that includes data from 3D movies. This approach enhances the model's ability to adapt to various environments and scenarios. For predicting high-resolution depth, many methods incorporate multi-scale features or post-processing techniques [110, 181]. While these strategies can improve depth estimation accuracy, they often complicate the learning process. Furthermore, if not trained on an extensive and varied dataset, these methods can exhibit limited generalization capabilities, underscoring the need for large and diverse datasets to train robust depth estimation models. These developments highlight the evolution of depth estimation techniques, showcasing the balance between model complexity, data requirements, and the ability to generalize across different environments and scenarios.

Self-supervised monocular methods have been developed to address the limitations of supervised depth estimation methods, particularly the reliance on ground truth depth data. Early self-supervised approaches [45, 176] utilized stereo images to train a network for depth prediction. In these methods, the left image is warped to match the right image, using photometric consistency as the training signal. Monodepth by Godard et al. [50] introduced a left-right consistency loss, enhancing depth estimation by leveraging the mutual warping of stereo image pairs. While this method showed improved depth quality, it still relied on synchronized stereo image pairs. Monocular training methods were developed to overcome the dependency on stereo images. These methods utilize single-camera video frames, employing estimated poses between frames for image warping. This approach is inherently more challenging but has seen significant advancements. Monodepth2 by Godard et al. [51] narrowed the accuracy gap between stereo and monocular training through techniques like automasking and a minimum reprojection loss. Subsequent research has continued to refine these techniques, with a notable focus on improving accuracy in various scenarios [22, 92, 127, 128, 150, 180]. Additionally, studies by Luo et al. [105] and Watson et al. [171], have specifically investigated temporal consistency in depth estimation.

In our work, to assess the impact of different training signals, either supervised or self-supervised, on monocular depth estimation, we utilize the ResNet backbone from the well established Monodepth2 [51] along with its varied training strategies, as detailed later. This approach allows us to compare and evaluate the effectiveness of these different methodologies in achieving accurate depth predictions from monocular video sequences.

### 3.2.3 Reconstruction and Novel View Synthesis from Multi-View Consistency

The reconstruction of the 3D geometry of a scene from two-dimensional 2D images, optionally aided by depth maps [115], is a fundamental task in computer vision and 3D modeling. In this process, scenes can be represented either explicitly or implicitly. Typical explicit representations include point clouds or meshes [26], where the scene is represented as a set of discrete points or vertices connected to form a mesh. This representation offers a direct and detailed depiction of the scene's geometry. On the other hand, implicit representations, such as distance fields [187], describe the scene as a level set of a specified mathematical function. Distance fields provide a more compact and continuous representation of the scene's geometry, enabling efficient storage and processing. A recent advancement in implicit representation involves neural fields, where the scene is encoded within the weights of the network [177].

#### 3.2.3.1 Neural Radiance Fields

Neural radiance fields (NeRF), particularly since the seminal work by Mildenhall et al. [111], have seen significant attention due to their ability to synthesize novel photorealistic views of scenes. In a typical NeRF setup, a neural network is trained on images with known poses to represent a scene. This method optimizes for predicting volume density and view-dependent emitted radiance within a volume. Novel views are synthesized by integrating along query rays, also for deformable scenes, as demonstrated in NeRFies [118].

Recent advancements in this field have further enhanced the capabilities of NeRF. Barron et al. [7] extended the original NeRF concept to unbounded scenes with Mip-NeRF 360, which achieves higher-quality scene representations. Chen et al. [21] introduced TensoRF, which factors the scene representation into low-rank components, allowing for faster and more efficient usage. There have also been developments in improving pose estimates and calibration robustness, as explored in BARF [99] and in NeRF [170].

While initial NeRF training was computationally demanding, techniques such as Plenoxels [38], which leverage spherical harmonics within a voxel grid structure, have accelerated processes. Additionally, interpolation techniques proposed by Sun et al. [155] have further accelerated training. Incorporating geometric priors, such as sparse and dense depth maps, can further regularize convergence, enhance quality, and reduce training time [28, 131]. Moreover, recent research has focused not only on methodological advancements but also on leveraging real-world data [130], like constructing a dataset for evaluating novel view synthesis and category-centric 3D reconstruction methods using crowd-sourced videos of real-world objects.

In our work, we incorporate recent advances in NeRF, particularly analyzing the impact of sensor-specific depth priors as discussed in Roessle et al. [131], for the task of implicit scene reconstruction. We utilize the ground truth robotic pose data from our dataset to minimize the influence of pose estimation errors and ensure highly accurate data. This approach allows us to explore the full potential of NeRF-based reconstruction methods in producing detailed and accurate 3D representations of scenes.

## 3.3 Accurate Data for Evaluation of Sensor Characteristics and 3D Vision Tasks

The computer vision community heavily relies on publicly available datasets to evaluate 3D vision tasks. In depth estimation, early datasets [139] predominantly relied on passive multi-view stereo cameras. However, these datasets exhibited limitations in textureless regions and constrained scenarios. In response to these limitations, active sensor configurations, such as active stereo and pattern projection sensors, were introduced to address these challenges by introducing artificial patterns and extending their utility to unconstrained scenarios [1, 145, 152, 161].

Nevertheless, it is worth noting that these sensors are not devoid of artifacts, which include bias, jitter [191], blurriness [79], and the inability to accurately estimate depth on certain surfaces, often necessitating post-processing with human annotation [145, 151]. Additionally, Time of Flight (ToF) sensors rely on measuring the time light takes to travel for distance measurement. However, they can introduce artifacts like phase wrapping [77] for I-ToF, multi-path interference (MPI) [40, 41, 75], and material-dependent artifacts [79]. Despite the presence of these artifacts, datasets derived from these sensors are frequently utilized in research without rigorous evaluation of depth quality.

In the domain of 6D pose estimation, several datasets have been developed for research and evaluation purposes. Notable among these are commonly used datasets such as LineMOD [59], YCB [174], and NOCS [166], which provide images with annotated object poses. These annotations are typically obtained using checkerboards, RGB-D cameras, or a combination of both. However, it is essential to acknowledge that the annotation quality of these datasets has been reported as inaccurate. This inaccuracy can be attributed to the limitations of checkerboard-based localization and errors introduced by depth sensors [15, 167].

Efforts have been made to improve the accuracy of pose annotation methods. One such approach involves utilizing multiview keypoints localized by checkerboards [101, 102]. These methods have demonstrated a significant enhancement in annotation accuracy compared to depth-based annotation. Notably, they have even succeeded in annotating objects with complex photometric properties, such as glasses [102]. However, several limitations in the acquisition pipeline persist. This pipeline involves scene scanning using a robotic arm, while object annotation relies on 2D keypoint annotation, resulting in notable annotation errors, with deviations reported as significant as 3.4mm [102].

We have constructed scenes of multiple objects with various shapes and materials. These scenes were specifically designed to facilitate an in-depth analysis of sensor characteristics. We acquired 3D models of objects with complex photometric properties, including reflective or transparent surfaces, with a priori high-quality capturing and then aligned them with the scenes. We employ a synchronized multi-modal custom sensor configuration mounted at a robot end-effector to capture images for our analysis. This setup ensures precise measurement of camera poses [167]. Subsequently, we extract high-quality rendered depth information a

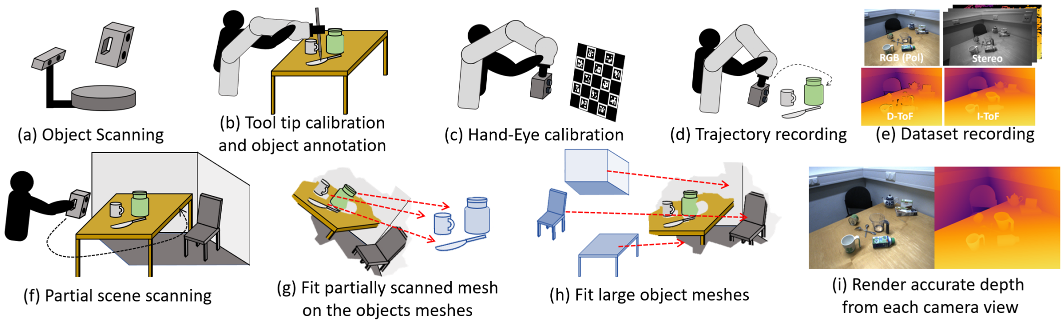

RGB (Pol)
Stereo
D-ToF
I-ToF
(a) Object Scanning
(b) Tool tip calibration and object annotation
(c) Hand-Eye calibration
(d) Trajectory recording
(e) Dataset recording
(f) Partial scene scanning
(g) Fit partially scanned mesh on the objects meshes
(h) Fit large object meshes
(i) Render accurate depth from each camera view

Fig. 3.2 **Scanning Process Overview:** We have developed a multi-stage acquisition process to achieve highly accurate scene geometry. Initially, 3D models are obtained using structured light 3D scanners (a). Subsequently, a calibration procedure is carried out to align the scene objects (b) and the mounted sensor rig (c) with the robotic platform, a process facilitated by the use of the PhoCal calibration framework [167]. The acquisition process continues with recording a motion trajectory in gravity compensation mode (d). This trajectory is repeated to capture synchronized images from all sensors (e). Utilizing this data, we create a partial digital twin of the 3D scene (f), which is then aligned with both smaller (g) and larger (h) objects within the scene. This process generates a complete virtual twin to render views from the perspective of each sensor employed (i). These rendered views provide highly accurate dense depth maps, enabling detailed investigations of individual sensor components.

posteriori for the viewpoint of each sensor. An overview of the acquisition pipeline is presented in Fig. 3.2.

In contrast to prior 3D and depth acquisition setups [20, 27, 151], which typically scan the entire scene as a whole, thus constraining the quality by the capabilities of the utilized sensor, our approach adopts a different strategy. We opt to independently scan each object, encompassing objects such as chairs, background elements, and smaller household items, using two high-quality structured light object scanners in advance. This methodology significantly enhances the annotation quality for the scenes, with the robotic 3D labeling process exhibiting only a point RMSE of 0.80 mm [167]. For perspective, as a point of comparison, a Kinect Azure camera introduces a standard deviation of 17 mm within its working range [101]. This elevated level of accuracy empowers us to methodically investigate depth errors objectively that originate from sensor noise and related artifacts, as depicted in Fig. 3.3. Simultaneously, it resolves prevalent issues related to imperfect meshes often encountered in available datasets (cf. Fig. 3.1, left).

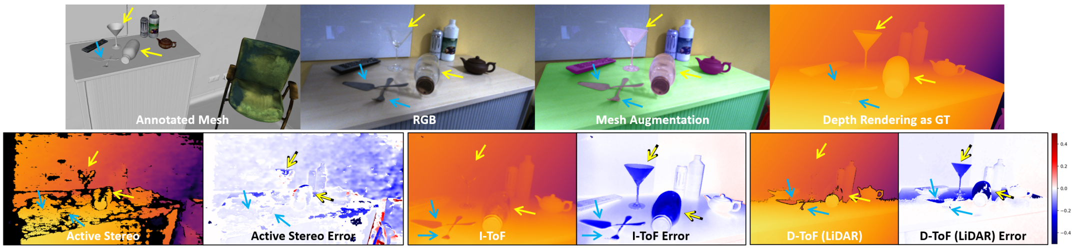


Fig. 3.3 **Data Composition:** The comprehensive annotation of the scene with a mesh enables the generation of exact depth maps from any viewpoint. These depth maps serve as ground truth data, facilitating the analysis of sensor errors under various scene structures. For example, when considering Time of Flight (ToF) sensors, it becomes evident that transparent objects like glass (highlighted in yellow) remain undetectable, and reflective objects (highlighted in cyan) introduce errors due to reflection-induced effects inherent to the measurement principle.

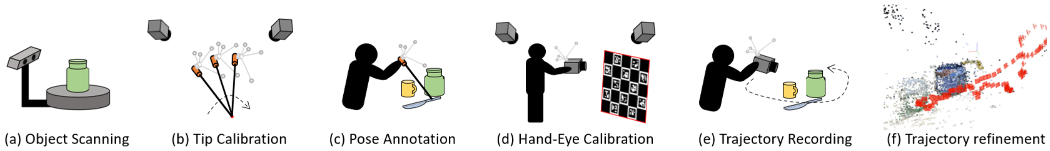


Fig. 3.4 **Data Acquisition Pipeline for Free-hand Camera Rig:** (a) Pre-scanning 3D models. (b) Pivot calibration to measure the tip. (c) Object pose annotation using the measuring tip. (d) Hand-Eye-Calibration for camera center calibration. (e) Recording of camera trajectory. (f) Post-processing steps to minimize synchronization errors.

To address the limitations of established datasets in the context of 3D vision tasks, we propose novel paradigms for acquiring high-quality and multi-modal datasets. Our approach includes a unique multi-modal sensor rig that incorporates a variety of depth modalities. Leveraging this advanced setup, we achieve precise surface measurements of objects and comprehensive annotations for objects and scenes, even in scenarios with challenging photometric conditions. Our solution introduces a robotic setup to annotate 6D object poses and camera poses. By harnessing the precision of forward kinematics in the robotic arm, we attain exceptionally accurate annotations for depth and 6D pose datasets. This approach is illustrated in detail in Fig. 3.2.

While our robotic setup provides precise annotations, its limitations include a restricted working range (with a maximum radius of 80cm) and inherent joint limitations. These limitations can affect camera pose distribution within a 6D pose dataset. We introduce a freehand data acquisition procedure to address these challenges, as illustrated in Fig. 3.4. This procedure ensures accurate data recording by utilizing an infrared tracking system and subsequent post-processing techniques. These post-processing steps encompass multiple calibrations and trajectory refinements. Our freehand data acquisition method extends the viewpoint coverage and enhances the accuracy of object pose annotations compared to existing datasets. Importantly, it achieves this while maintaining superior overall annotation quality.

Both of our proposed dataset annotation methods, robotic and freehand, adhere to a common underlying principle for obtaining high-quality 3D data, comprising four key steps:

1. Object or scene scanning.

2. Measurement of 20-30 accurate surface points on objects using a tracked tool tip.

3. Annotation of the object pose achieved through point correspondence followed by Iterative Closest Point (ICP) refinement.

4. Recording the scene using a tracked camera.

While these methods share a common principle, each step necessitates distinct calibration and post-processing procedures to ensure the overall quality of the acquired data.

## 3.3.1 Robotic Approach

In our robotic approach, we employ the KUKA LBR iiwa 7 R800 robot, known for its exceptional position accuracy of ±0.1 mm. We utilize the EinScan-SP table-top scanner to scan smaller objects, while larger objects are scanned using the Artec Eva hand-held scanner. To overcome challenges posed by challenging materials, we apply the AESUB Blue 3D scanning spray prior to scanning.

### 3.3.1.1 Object Pose Annotation

After acquiring object meshes, we affix a measuring tool tip to the robot's End-Effector (EE) and calibrate the tool tip. Subsequently, we carefully capture precise surface points on the objects using the tool tip. These points are the basis for annotating object poses relative to the robot's base, accomplished through point correspondence followed by Iterative Closest Point (ICP) alignment with the scanned object meshes. The pose error for the object pose annotation step is quantified as 0.20 mm (RMSE) and 0.38° [167]. The procedure is illustrated in Fig. 3.2 (a)-(e).

In the context of our depth dataset, we extend our annotation efforts to encompass background elements such as walls and tables, facilitating the rendering of complete scenes (refer to Fig. 3.5). Due to the constrained working range of the robot, annotating the background with the robotic arm is impractical. Consequently, we initiate the process by scanning the scene using the hand-held scanner to acquire a partial mesh representation. Subsequently, we align this partially scanned mesh with the objects annotated from the robot's base. Following this alignment, we adapt the background meshes to correspond with the robot's base, ensuring a coherent scene annotation. The additional background annotation step is outlined in Fig. 3.2 (f)-(i).

### 3.3.1.2 Camera Pose Annotation

Once all the objects and the background are thoroughly annotated, the scene is captured using the camera rig. In the robotic approach, the camera rig is attached to the robot end-effector (EE), and we perform hand-eye calibration for each sensor to determine the transformation

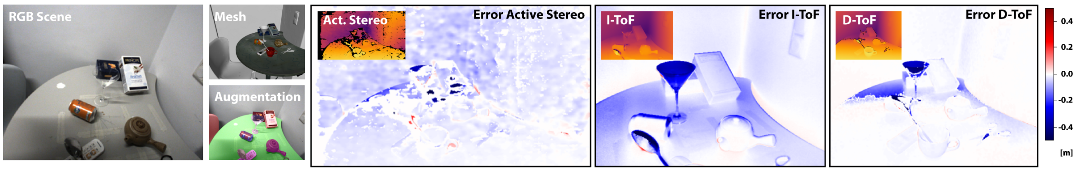


Fig. 3.5 **Data Quality:** A complete 3D reconstruction of the RGB scene (left) enables the generating of exact depth maps from arbitrary viewpoints. These depth maps serve as ground truth data, facilitating an in-depth exploration of depth errors associated with different sensors and various scene structures (right). For instance, owing to the measurement principle, translucent glass objects may become invisible to Time-of-Flight (ToF) sensors.

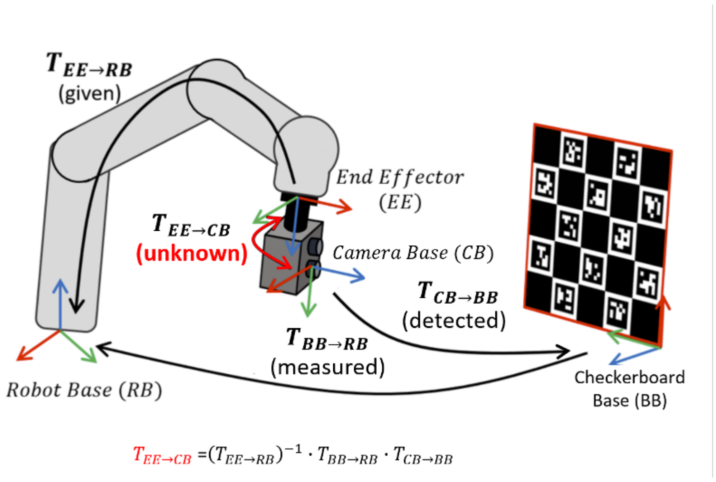


Fig. 3.6 **Hand-Eye-Calibration Robot:** We employ a closed-form approach for hand-eye calibration in our robotic approach.

matrix from the EE. For this purpose, we employ a closed-form solution, as depicted in Fig. 3.6 (a). To obtain the transformation $T_{BB\to RB}$ of the checkerboard via the robotic tip and $T_{CB\to BB}$ when the checkerboard is detected by camera, the hand-eye calibration $T_{MB\to CB}$ can be calculated through matrix multiplication involving $T_{RB\to EE}$ (derived from forward kinematics), $T_{BB\to RB}$, and $T_{CB\to BB}$. To ensure perfect synchronization between the cameras and the robot during trajectory capture, we pause the robot on each frame before triggering the cameras. We obtain an average RMSE of 0.86 mm by accumulating all calibration errors. Considering object and camera pose annotation, the overall error for the entire pipeline is measured as 0.8 mm, as also observed in [167].

## 3.3.2 External Tracker Approach

To address the limitation of pose coverage posed by the robotic arm-based annotation, we introduce a free-hand approach for acquiring high-quality datasets with more diverse viewpoints and scene coverage. In this extended approach, we replace the robotic arm with an external tracking camera (ARTTRACK2), with an accuracy of 0.67 mm/0.12° in static scenarios and 0.92 mm/0.16° in dynamic cases. Remarkably, this approach yields accuracy comparable to that achieved through the robot-based annotation. The dataset acquisition pipeline for this approach is detailed in Fig. 3.4.

### 3.3.2.1 Object Pose Annotation

In the free-hand approach, object pose annotation follows a procedure similar to the robotic approach, with a notable distinction being replacing the end-effector (EE) with an infrared (IR) tracking body. This IR tracking body is capable of being tracked by the external tracker. The process remains consistent with measuring sparse surface points on the object using the calibrated tip and subsequently conducting point correspondence with ICP for annotating the

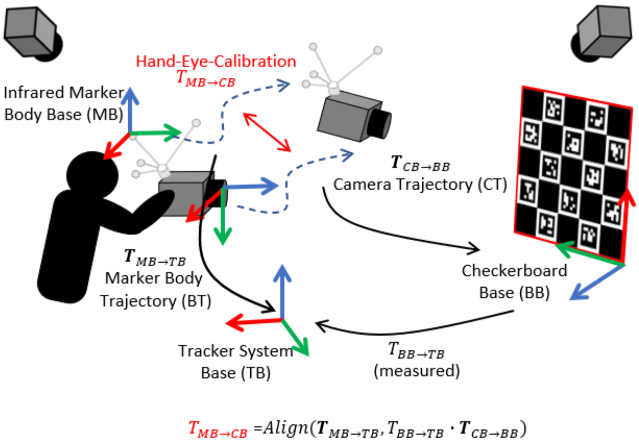


**Fig. 3.7** **Hand-Eye-Calibration Tracker:** We employ a trajectory alignment approach specifically suited for the external tracker approach for hand-eye calibration. This method enhances robustness by mitigating errors.

object from the tracker base. The tracker error observed in object annotation averages an RMSE of 0.32 mm in translation and 0.43° in rotation, a level of accuracy on par with the robotic annotation method.

#### 3.3.2.2 Camera Pose Annotation

Substituting the end-effector (EE) with a tracking body attached to the camera and achieving high-quality pose annotation introduces two noteworthy challenges. Firstly, the hand-eye calibration process encounters increased error due to using less precise hardware. Additionally, the camera trajectory involves more rotations, increasing error propagation. Secondly, the accuracy of the tracking system exhibits a decline in dynamic scenarios due to synchronization issues.

To achieve a more robust and error-resistant hand-eye calibration method, we employ a trajectory-based approach, deviating from the reliance on a closed-form solution (as depicted in Fig. 3.7, (b)). Initially, we determine the pose of the checkerboard using the calibrated tooltip, denoted as $T_{BB\to TB}$. Subsequently, the trajectory of the camera can be accurately localized from the tracker system when the checkerboard is detected ($T_{CB\to BB}$). This is achieved through the multiplication of $T_{CB\to BB}$ and $T_{BB\to TB}$. Given that the camera's tracking marker trajectory is already localized from the tracker base ($T_{MB\to TB}$), aligning the camera trajectory with the marker trajectory provides the offset pose, constituting the hand-eye calibration matrix ($T_{MB\to CB}$). We assess the calibration quality by measuring the pose error between the aligned trajectories, resulting in a measurement of 0.27 mm for translation accuracy and 0.42° for rotation accuracy.

We initiate a preliminary synchronization of all hardware components through a hardware signal to ensure accurate camera synchronization. Subsequently, we refine the time offset by employing the ICP on the hardware trajectory. The refinement process involves visualizing the trajectory as a 2D distance graph with time on the x-axis and distance on the y-axis, a method-

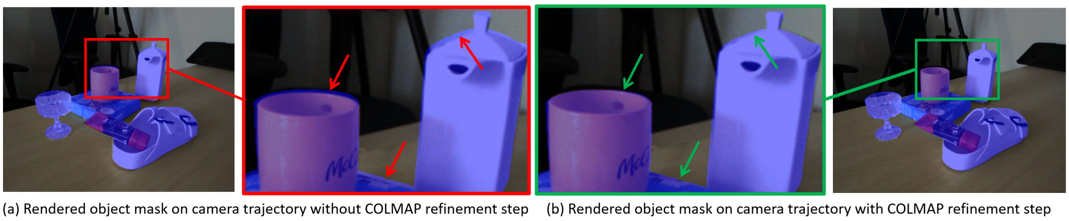

(a) Rendered object mask on camera trajectory without COLMAP refinement step (b) Rendered object mask on camera trajectory with COLMAP refinement step

Fig. 3.8 **Example of SfM-based Refinement:** The refinement process based on Structure-from-Motion (SfM) effectively mitigates subtle errors that may persist during abrupt camera movements.

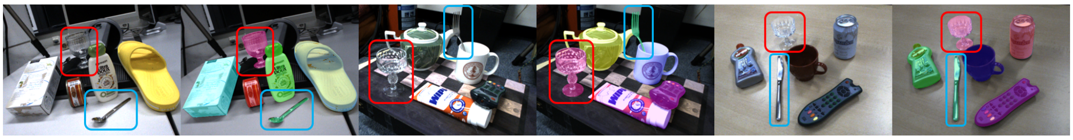


Fig. 3.9 **Qualitative Example of External Tracker-based Annotation:** The red box highlights the annotation on glass objects, while the cyan box highlights the annotation on reflective objects.

ology akin to previous works [33, 68]. The ICP is then utilized to align the trajectory points, and the time offset is determined by evaluating the displacement along the x-axis. Despite these refinements, we have observed minor pose offsets during abrupt camera movements (as depicted in Fig. 3.8 (a)). To address this challenge, we employ refinement techniques based on Structure-from-Motion (SfM) principles [140, 141]. This SfM-based refinement enhances the camera trajectory by incorporating hand-selected fixed poses from a subset of frames. The results of this refinement are particularly notable during sudden movements, as illustrated in Fig. 3.8 (b). Quantifying the direct improvement brought about by SfM is a complex task. Therefore, we evaluate camera pose errors using upper and lower bounds. The upper bound takes into account dynamic errors introduced by the tracking system. The lower bound assumes that dynamic errors have been resolved and utilizes tracking errors from static scenarios. To evaluate the quality of our annotation, we propagate the pose annotation error in conjunction with the camera pose error. Our assessment yields an annotation quality range from 1.35 mm to 1.73 mm in terms of RMSE. Fig. 3.9 provides a qualitative evaluation using object mask rendering, particularly emphasizing the photometrically challenging objects.

### 3.3.3 Sensor Setup & Hardware Description

The two 3D scanners used in our dataset acquisition serve distinct purposes. The table-top scanner, EinScan-SP, by SHINING 3D Tech. Co., Ltd., Hangzhou, China, is equipped with a rotating table primarily designed to scan small objects. In contrast, the hand-held scanner, Artec Eva by Artec 3D, Luxembourg, is employed for larger objects and background scanning. For objects and surfaces with challenging materials, we apply a self-vanishing 3D scanning spray (AESUB Blue). In cases where larger, texture-less areas like tables and walls are encountered, small markers [46] are temporarily affixed to the surface. These markers enable the relocalization of the 3D scanner during scanning sessions. The robotic manipulator utilized in our setup is the KUKA LBR iiwa 7 R800, manufactured by KUKA Roboter GmbH, Germany. It achieves a position accuracy of ±0.1 mm. We rigorously validated this accuracy during

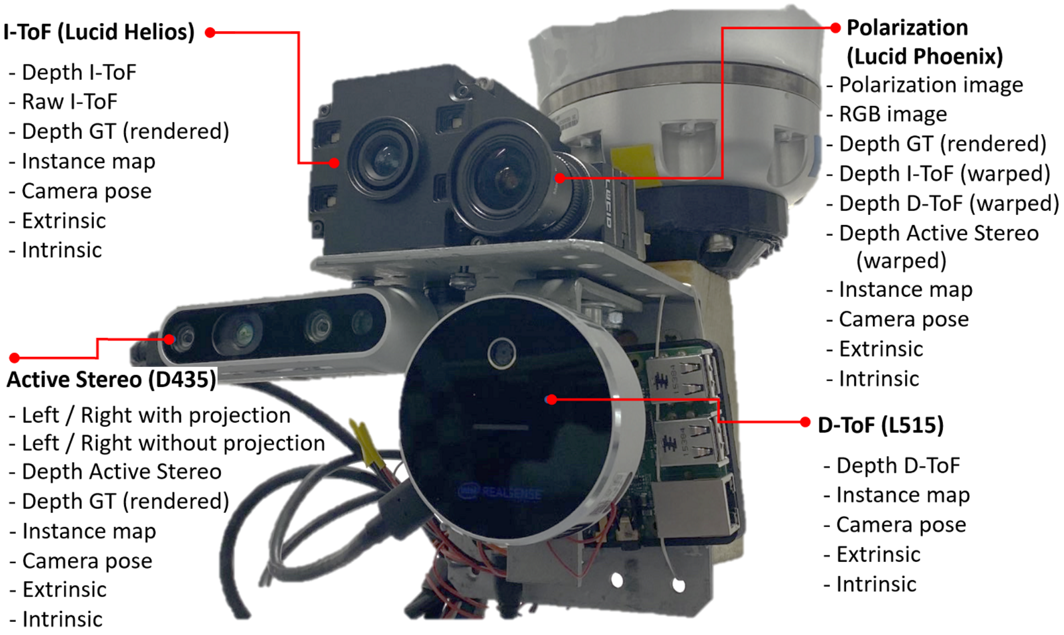


Fig. 3.10 **Multi-Modal Camera Rig:** The custom multi-modal sensor rig incorporates a range of depth sensors, including I-ToF (top left), Stereo (lower left), D-ToF (lower right), and RGB-P (Polarization, top right). This rig is securely affixed to a robot end-effector (top), with a Raspberry Pi (right) serving as the trigger mechanism for acquisition.

the pivot calibration stage by calculating the 3D location of the tool tip, leveraging forward kinematics and hand-tip calibration. This validation process revealed positional variations within the range of $[-0.158, 0.125]$ mm, aligning with the specified accuracy. Our dataset is distinguished by a unique multi-modal setup comprising four different cameras, each providing distinct types of input images, namely RGB, polarization, stereo, Indirect Time of Flight (I-ToF) correlation and depth, Direct Time of Flight (D-ToF) depth, and Active Stereo depth. RGB and polarization images are acquired with a Phoenix 5.0 MP Polarization camera (PHX050S1-QC) equipped with a Sony Polarsens sensor (IMX264MYR CMOS, Sony, Japan). For stereo image acquisition, we employ an Intel RealSense D435 camera from Intel, USA, with the infrared projector disabled. Depth information is acquired from an Intel RealSense L515 D-ToF sensor, an Intel RealSense D435 active stereo sensor with infrared pattern projection, and a Lucid Helios (HLS003S-001) I-ToF sensor by LUCID Vision Labs, Canada. Each camera is triggered separately by a Raspberry Pi to eliminate interference effects arising from the infrared signals of the depth sensors. For the robotic approach, the hardware is rigidly mounted at the robot's end-effector (refer to Fig. 3.10), allowing precise frame-by-frame synchronization to acquire a pre-recorded trajectory.

#### 3.3.3.1 Polarization Camera

Figure 3.11 displays sample images from the polarization camera. The polarization camera captures images with varying polarization angles, enabling the extraction of surface normals based on the physical material properties of objects in the scene. The polarization camera used in our dataset provides polarized images at four different angles (0, 90, 180, and 270

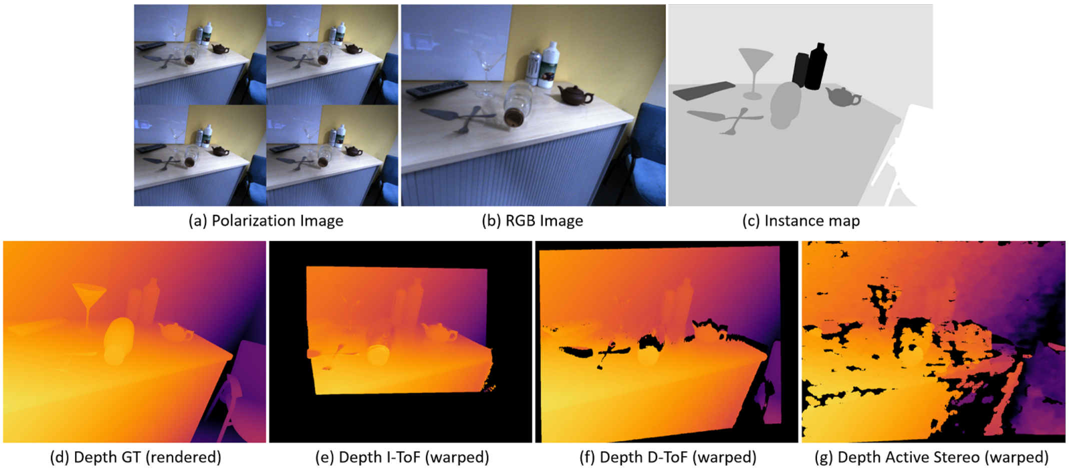


**Fig. 3.11 Polarization Image Example:** Examples showcasing the images included for the polarization camera input (top), complemented by an instance label map and depth measurements from the other sensors accurately transformed to polarization reference frame.

degrees), which are arranged in a single 2x2 image (Figure 3.11, (a)). A regular RGB image is generated by averaging these four images (Figure 3.11, (b)).

We have included warped depth images from each depth camera into the polarization camera's coordinates to showcase the results of depth maps from different sensors. This transformation is achieved using the extrinsic parameters between the cameras (Figures 3.11, (d-g)). Additionally, we provide supplementary information, including an instance map (Figure 3.11, (c)), and 6D object poses to support the training and validation of pipelines for other tasks. We also include precise 6D camera pose information in the form of 4x4 matrices obtained from the robotic arm, extrinsic transformations between cameras as 4x4 matrices, and camera intrinsics as a 3x3 matrix.

#### 3.3.3.2 D-ToF Camera

The Direct ToF (D-ToF) camera operates by emitting an infrared signal and measuring the return time of the signal, providing depth information of its surroundings. Signal reflections influence the quality of this modality. It can be susceptible to specific physical noise sources, such

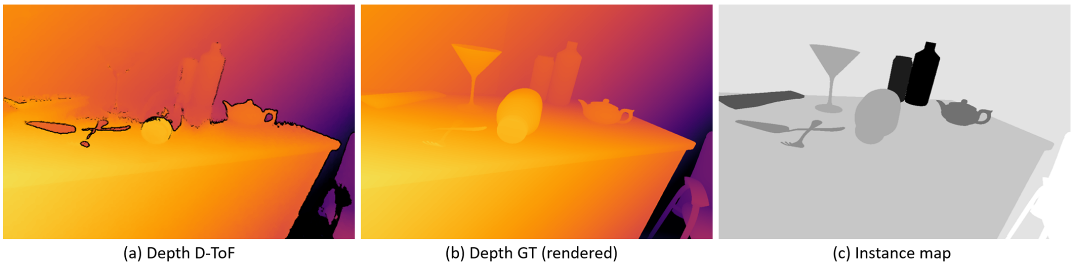


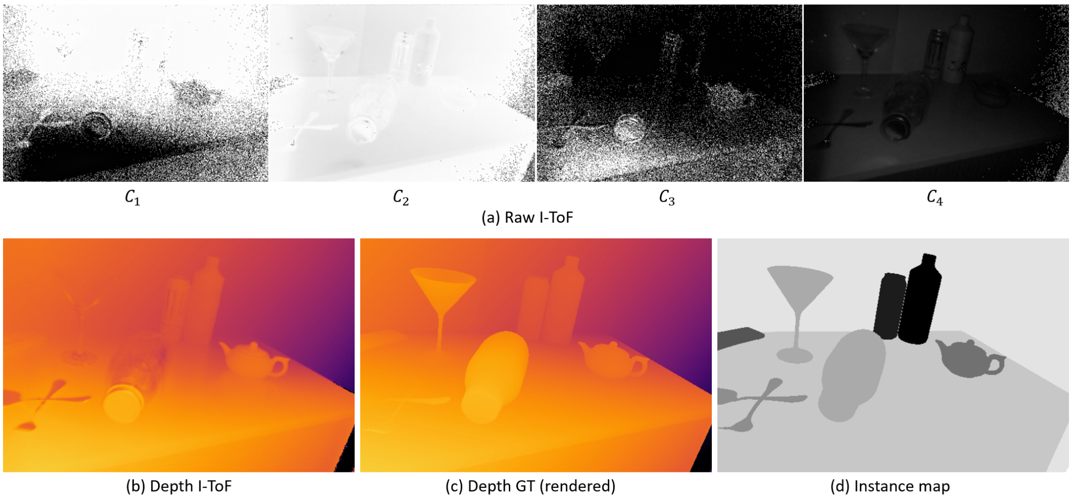
**Fig. 3.12 D-ToF Example:** Example of the images provided for the D-ToF camera: its depth map (left), corresponding ground truth depth (center), and an object instance label map (right).

as Multi-Path Interference (MPI) and material-dependent artifacts (cf. Fig. 3.15 as detailed later). In our dataset, we offer both the depth map captured by the D-ToF camera (Fig. 3.12, (a)) and its corresponding ground truth depth map (Fig. 3.12, (b)). This allows for research into D-ToF refinement techniques aimed at reducing such errors.

#### 3.3.3.3 I-ToF Camera

Indirect ToF (I-ToF) cameras perceive depth information of their surroundings by emitting a frequency-modulated signal and measuring the returning signal. Unlike Direct ToF (D-ToF), I-ToF cameras do not determine depth based on time differences; instead, they correlate the returning signal with phase-shifted emitting signals, resulting in four distinct measurements known as correlation images. These correlations are represented as sinusoidal functions of distance ($(\sin(d), \cos(d), -\sin(d), -\cos(d)) = \left(c_1, c_2, c_3, c_4\right)$ in Fig. 3.13, (a)). Depth information can be extracted from the correlation images using the arctangent formula or convolutional neural networks.

Similar to D-ToF, the I-ToF modality relies on signal reflections, making it susceptible to artifacts such as Multi-Path Interference (MPI) and material-dependent effects (compare qualitative results in test scenes in Figs. 3.22, 3.23, and 3.24). In our dataset, we provide raw correlation images and the depth map captured by the I-ToF camera (see Fig. 3.13, (a, b)), along with their corresponding ground truth depth map (Fig. 3.13, (c)). This enables researchers to train depth improvement pipelines for I-ToF, either from raw signals or starting with the initial I-ToF depth data.



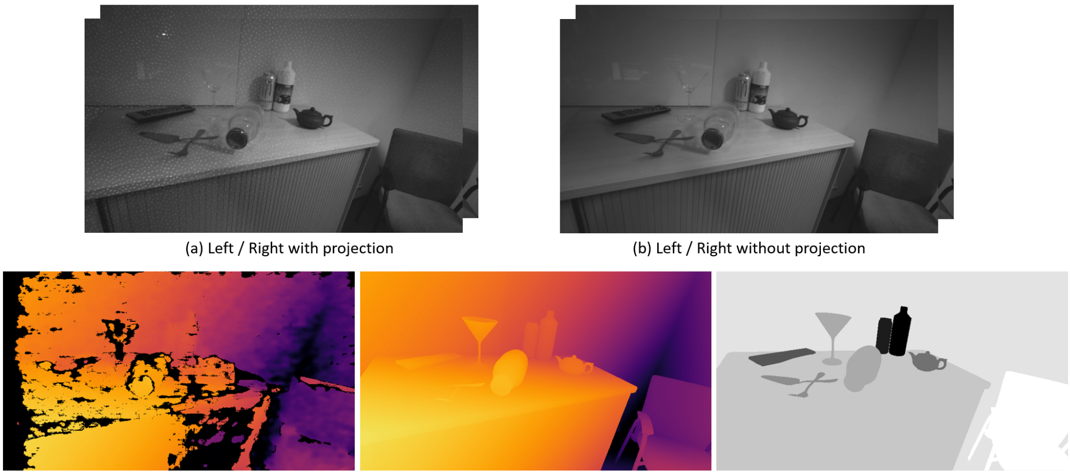
(a) Raw I-ToF

(b) Depth I-ToF

(c) Depth GT (rendered)

(d) Instance map

Fig. 3.13 **I-ToF Example:** Example of the images provided for the I-ToF camera, including raw correlation images, the computed depth map, depth ground truth, and an object instance label map.



(a) Left / Right with projection

(b) Left / Right without projection

(c) Depth Active Stereo

(d) Depth GT (rendered)

(e) Instance map

Fig. 3.14 **Active Stereo Example:** Example of the images provided for the Active Stereo camera, comprising left and right images (with/without projection), the computed depth map, depth ground truth and an instance map.

#### 3.3.3.4 Active Stereo Camera

Stereo depth estimation relies on photometric consistency and geometrical constraints derived from epipolar geometry to triangulate depth maps from disparities between left and right cameras. While stereo depth estimation methods are less sensitive to specific materials, they face challenges related to stereo occlusion and large texture-less regions. Active projection, such as Active Stereo, is employed to mitigate these issues.

In our dataset, we provide both active and passive stereo left/right images (Fig. 3.14, (a), (b)), as well as raw depth data from the camera (active, Fig. 3.14, (c)), and their corresponding ground truth depth maps (Fig. 3.14, (d)). This comprehensive data enables researchers to enhance stereo methods, whether focused on passive or active stereo, as well as depth refinement pipelines.

## 3.4 Learning 3D Perception in Multi-View Conditions

Due to its unique characteristics, our dataset enables comprehensive and thorough analysis of various depth sensor modalities. It also offers a detailed quantitative assessment of learning-based dense scene regression techniques when trained with diverse supervision signals and for 6D object pose estimation. Our primary focus revolves around two widely recognized tasks to better understand and study the characteristics of the individual sensors: monocular depth estimation and implicit 3D reconstruction, with a particular emphasis on novel view synthesis.

### 3.4.1 Depth Estimation

We adopt the widely used architecture introduced in [51] to train monocular depth estimation models. Our approach involves training an encoder-decoder network incorporating a ResNet18 encoder and skip connections to predict dense depth maps. Using different supervision signals from various depth modalities allows us to investigate the influence and characteristics of different 3D sensors. Additionally, we explore the potential of leveraging complementary semi-supervision, using information for the relative pose between monocular acquisitions and consecutive image data from a moving camera.

**Dense Supervision** In the fully supervised configuration, we utilize depth modalities from the dataset as supervision signals to guide the prediction of the four pyramid-level outputs. These predictions are upsampled to match the original input resolution. The loss function for this setup is defined as:

$$\mathscr{L}_{\text{supervised}} = \sum_{i=1}^{4} \left\| \widetilde{D}_i - D \right\|_1 . \tag{3.1}$$

Here, $D$ represents the supervision signal corresponding to valid pixels of the depth map, and $\widetilde{D}_i$ signifies the predicted depth at the $i$-th pyramid scale.

**Self-Supervision** Predicting depth and relative pose between consecutive frames captured by a moving camera can be viewed as a coupled optimization problem. We adopt established methods that formulate a dense image reconstruction loss through projective geometric warping [51]. In this process, we projectively transform a temporal image $I_{t'}$ at time $t'$ to the frame at time $t$ using the following equation:

$$I_{t'\to t} = I_{t'} \Big\langle \operatorname{proj}(D_t, T_{t\to t'}, K) \Big\rangle, \tag{3.2}$$

where $D_t$ represents the predicted depth for frame $t$, $T_{t\to t'}$ is the relative camera pose, and $K$ denotes the camera intrinsics. The photometric reconstruction error [51, 132, 171] between images $I_x$ and $I_y$, given by:

$$E_{\text{pe}}(I_x, I_y) = \alpha \tfrac{1-\text{SSIM}(I_x, I_y)}{2} + (1-\alpha) \left\| I_x - I_y \right\|_1 , \tag{3.3}$$

is computed between the target frame $I_t$ and each source frame $I_s$ with $s \in S$. The pixel-wise minimum error is selected to define $\mathscr{L}_{\text{photo}}$ over $S = [t-F, t+F]$ as:

$$\mathscr{L}_{\text{photo}} = \min_{s\in S} E_{\text{pe}}(I_t, I_{s\to t}). \tag{3.4}$$

Edge-aware smoothness, denoted as $\mathscr{L}_{\text{s}}$, is applied [51] to encourage locally smooth depth estimations, with the mean-normalized inverse depth $\overline{d_t}$ defined as:

$$\mathscr{L}_{\text{s}} = \left| \partial_x \overline{d_t} \right| e^{-|\partial_x I_t|} + \left| \partial_y \overline{d_t} \right| e^{-|\partial_y I_t|}. \tag{3.5}$$

The final training loss for the self-supervised setup is composed of both photometric loss ($\mathscr{L}_{\text{photo}}$) and edge-aware smoothness loss ($\mathscr{L}_{\text{s}}$), weighted by $\lambda_{\text{s}}$:

$$\mathscr{L}_{\text{self-supervised}} = \mathscr{L}_{\text{photo}} + \lambda_{\text{s}} \cdot \mathscr{L}_{\text{s}}. \tag{3.6}$$

**Semi-Supervision** In the case of semi-supervised training, we utilize the ground truth relative camera pose. The predicted depth estimate is employed to formulate the photometric image reconstruction loss, and we also incorporate the smoothness loss as previously described.

**Implementation Details** For all our depth estimation experiments, we utilize the PyTorch framework [120] and conduct training for 20 epochs to ensure comparability across experiments. We employ the Adam optimizer [85] for optimization. Monocular approaches are trained using a batch size of 12 on a single NVIDIA RTX-3090 GPU. We set $\lambda_{\text{s}}$ to $10^{-3}$ and sample $S$ with $T = 10$ frames offset due to the small relative camera movement between frames and the high frame rate. The RGB inputs are resized to dimensions of $480 \times 320$ for supervised training and $320 \times 160$ for self-supervised training, respectively. The depth network produces dense depth predictions across four pyramid levels, each with half the resolution of the previous level. The pose network and augmentations are consistent with the methodology outlined in [51]. We initiate training with an initial learning rate of $1 \times 10^{-4}$ for 15 epochs, which is then reduced to $1 \times 10^{-5}$ after 15 epochs in the self-supervised setting. For the supervised case, we begin with a learning rate of $1 \times 10^{-3}$, and every five epochs, we decrease it by a factor of ten.

## 3.4.2 Implicit 3D Reconstruction

Recent advancements in implicit 3D scene reconstruction have introduced neural radiance fields (NeRF) [111]. This technique excels in novel view synthesis, allowing the rendering of scene geometry or RGB views from unobserved viewpoints. Introducing additional depth supervision regularizes the problem, reducing the required number of views and improving training efficiency [28, 131].

Following the motivation of Roessle et al. [131], we utilize various depth modalities as additional depth supervision for novel view synthesis. Consistent with NeRF literature [111, 131], we employ an MLP $F_\theta$ to encode the radiance field for a scene, predicting color $\mathbf{C} = [r, g, b]$ and volume density $\sigma$ for a 3D position $\mathbf{x} \in \mathbb{R}^3$ and viewing direction $\mathbf{d} \in \mathbb{S}^2$. We apply positional encoding as introduced in [131]. For each pixel, we sample a ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ originating from the camera's origin $\mathbf{o}$. This ray travels through the volume at locations $t_k \in [t_n, t_f]$ between the near and far planes. By querying $F_\theta$, we obtain color and density information:

$$\hat{\mathbf{C}}(\mathbf{r}) = \sum_{k=1}^{K} w_k \mathbf{c}_k \text{ with } w_k = T_k \left(1 - \exp(-\sigma_k \delta_k)\right), \tag{3.7}$$

$$T_k = \exp\left(-\sum_{k'=1}^{k} \sigma_{k'} \delta_{k'}\right) \text{ and } \delta_k = t_{k+1} - t_k. \tag{3.8}$$

The NeRF depth $\hat{z}(\mathbf{r})$ is computed by:

$$\hat{z}(\mathbf{r}) = \sum_{k=1}^{K} w_k t_k, \tag{3.9}$$

and the depth regularization for an image with rays $\mathscr{R}$ is:

$$\mathscr{L}_{\mathrm{D}} = \sum_{\mathbf{r} \in \mathscr{R}} \frac{|\hat{z}(\mathbf{r}) - z(\mathbf{r})|}{\hat{z}(\mathbf{r}) + z(\mathbf{r})}, \tag{3.10}$$

where $z(\mathbf{r})$ is the depth of the sensor. Using the mean squared error (MSE) loss:

$$\mathscr{L}_{\text{color}} = \mathrm{MSE}(\hat{\mathbf{C}}, \mathbf{C}) \tag{3.11}$$

for synthesized colors, the final training loss is:

$$\mathscr{L}_{\text{NeRF}} = \mathscr{L}_{\text{color}} + \lambda_{\mathrm{D}} \cdot \mathscr{L}_{\mathrm{D}}. \tag{3.12}$$

**Implementation Details** We adhere to the NeRF framework of Mildenhall et al. [111] and extend upon the approach presented in [131], omitting a depth completion network. Instead, we utilize the depth information from the respective sensors and apply a scale-invariant depth loss $\mathscr{L}_{\mathrm{D}}$. Our image resolutions are set to 640 × 480, and we process batches of 1024 rays. We configure $\lambda_{\mathrm{D}}$ as 0.1 and the learning rate as $5 \times 10^{-4}$. The optimization process runs for 100,000 iterations using the Adam optimizer [85].

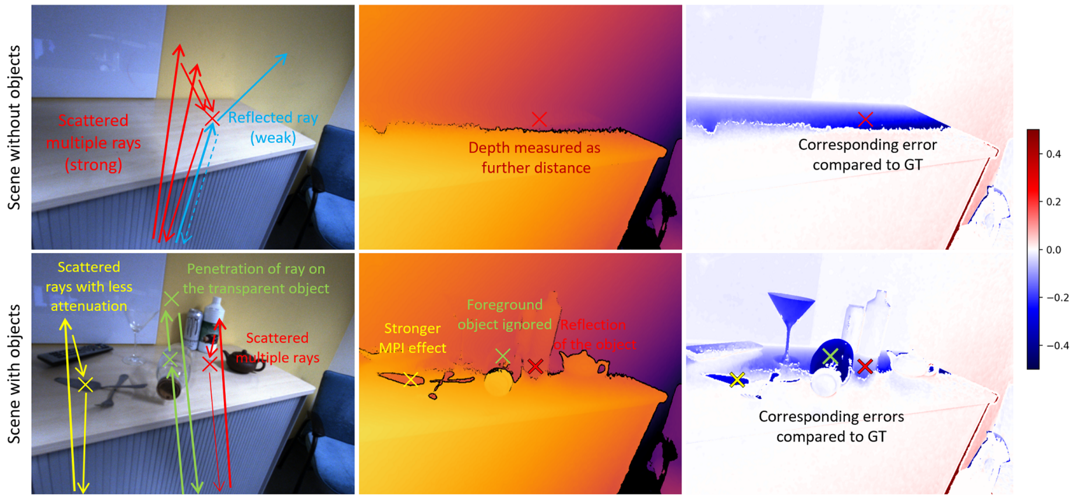
## 3.5 Experimental Results

### 3.5.1 Error Analysis of Different Sensor Modalities

This section focuses on analyzing specific errors associated with different depth sensor modalities. Our objective is to highlight how the quality of depth information is affected when a particular modality is used as ground truth for training or evaluation. Additionally, we emphasize the advantages of employing our rendered depth as the ground truth for various applications.

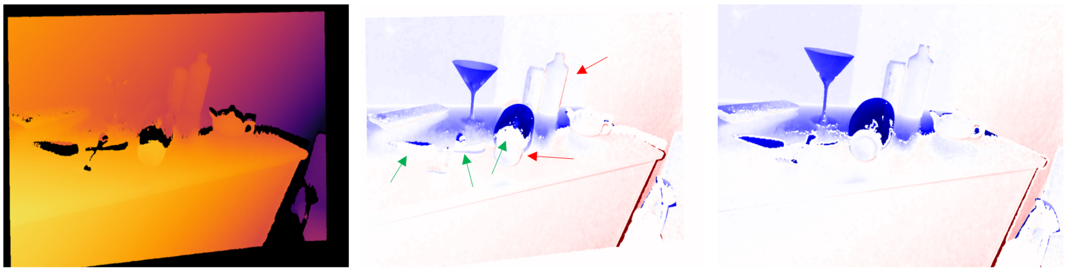
#### 3.5.1.1 D-ToF Camera

The D-ToF modality exhibits issues related to its reflection-based nature, including Multi-Path Interference (MPI) and material-dependent artifacts. When the angle of the surface normal of the scene closely aligns with the incident angle of the infrared signal, the reflected signal's strength weakens due to scattering effects (Fig. 3.15, (a), blue arrow). Meanwhile, multiple scattered signals from other surfaces, which have a longer travel distance, are received with stronger signals (Fig. 3.15, (a), red arrow) and interfere with the original signal, resulting in MPI. This produces incorrect depth measurements in areas with greater distance, which may appear as reflections or shadows of the object on the surface (Fig. 3.15, (b), red marker). This effect can be intensified when the surface material is reflective, as it reflects even weaker and noisier signals with less attenuation (Fig. 3.15, (a,b), yellow arrow and marker). Conversely, when the surface material is transparent, the emitted infrared signal tends to pass through the object in both directions (Fig. 3.15, (a), green arrow), effectively ignoring the object and causing the sensor to produce depth values similar to the background (Fig. 3.15, (b), green marker - material-dependent artifact). The quality of the depth map may degrade slightly around certain boundaries after warping it into the RGB frame (Fig. 3.16, (b), red). However, the invalid regions can be advantageous in invalidating more areas with incorrect depth, especially on reflective objects (Fig. 3.16, (b), green), which can be beneficial when used in training.


Scene without objects
Scattered multiple rays (strong)
Reflected ray (weak)
Depth measured as further distance
Corresponding error compared to GT
0.4
0.2
0.0
−0.2
−0.4
Scene with objects
Scattered rays with less attenuation
Penetration of ray on the transparent object
Scattered multiple rays
Stronger MPI effect
Foreground object ignored
Reflection of the object
Corresponding errors compared to GT

(a) Path of the rays (b) Depth value on MPI effected region (c) Error map on the effected region (m)

**Fig. 3.15** **D-ToF Analysis:** In-depth analysis of ray paths illustrating the impact of MPI and surface material-induced errors on the D-ToF modality. While D-ToF provides dense and sharp depth information, its quality is significantly influenced by surface material properties and incident angles.

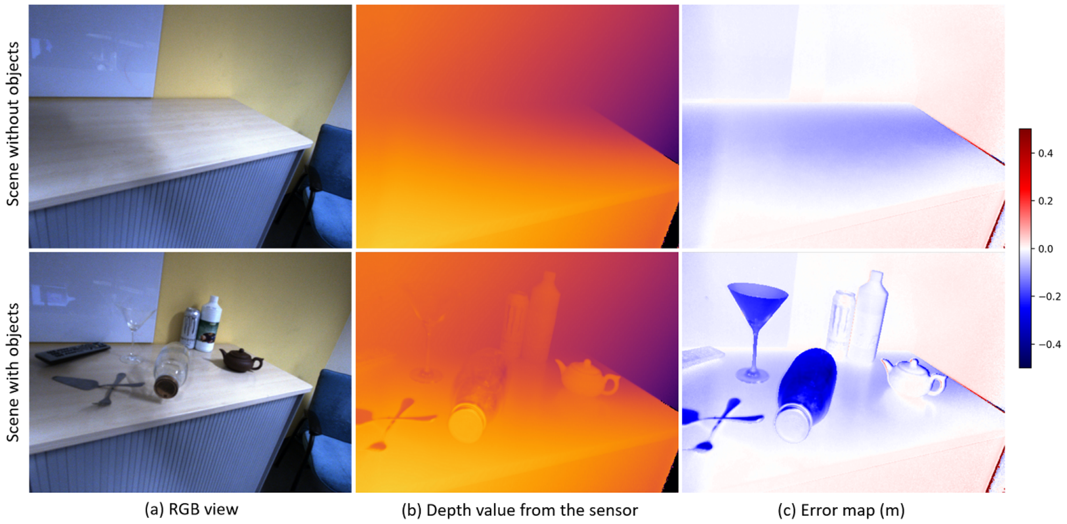


(a) D-ToF depth warped on RGB view (b) Error map on RGB view (c) Error map original view

**Fig. 3.16** **D-ToF Analysis after Alignment:** Error introduced after projectively transforming the D-ToF depth map into the RGB view. Minor errors are noticeable along some edges (highlighted in red), but the expansion of the invalid area contributes to the proper invalidation of depth values on reflective objects (highlighted in green).

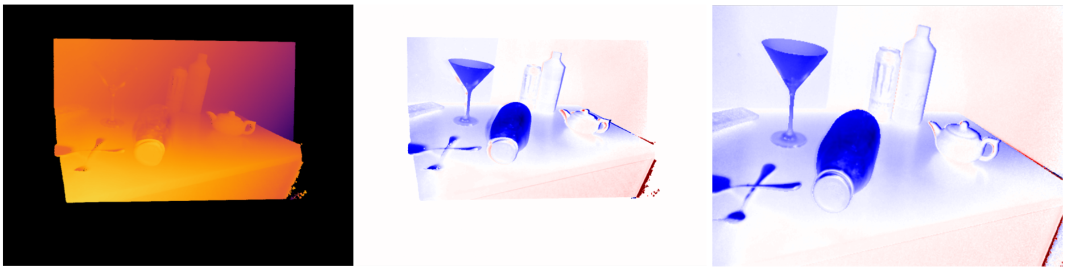
#### 3.5.1.2 I-ToF Camera

The I-ToF modality faces similar challenges to the D-ToF modality, such as Multi-Path Interference (MPI) and material-dependent artifacts (Fig. 3.17). While the quality of the depth itself may appear to be better, with denser depth maps (less invalid regions) and fewer artifacts, it is essential to note that I-ToF and D-ToF cameras belong to different price ranges and power levels. Comparing them directly can be challenging, and their suitability depends on specific application requirements. Unlike the D-ToF case, having fewer invalid areas but instead having areas with incorrect depth does not help invalidate depth information (Fig. 3.18). This difference can potentially lead to artifacts in predictions when using I-ToF data as ground truth during training.



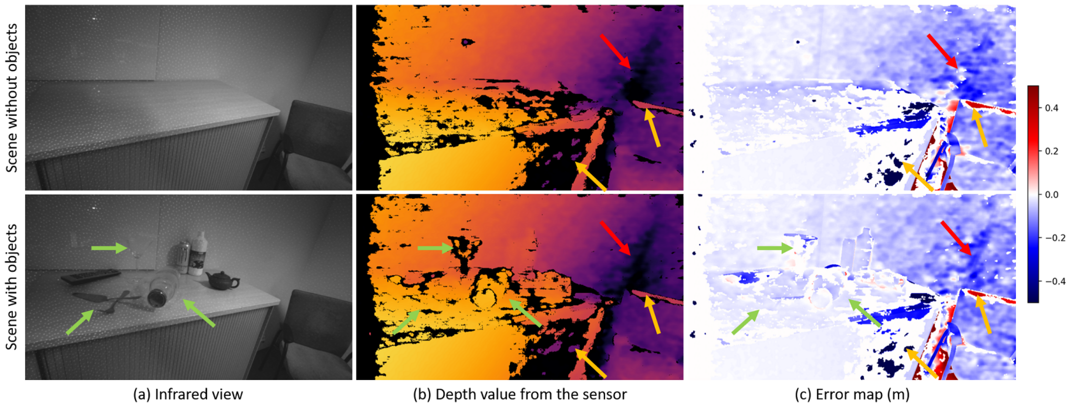
(a) RGB view (b) Depth value from the sensor (c) Error map (m)

Fig. 3.17 **I-ToF Analysis:** Depth quality from the I-ToF camera. The I-ToF modality exhibits similar artifacts to the D-ToF, but the depth map itself is denser and experiences fewer MPI artifacts on the table surface.



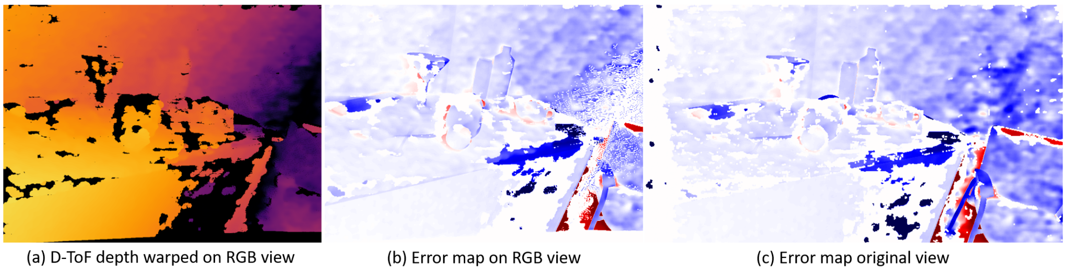
(a) D-ToF depth warped on RGB view (b) Error map on RGB view (c) Error map original view

Fig. 3.18 **I-ToF Analysis after Alignment:** Error after projectively transforming the I-ToF depth map into the RGB view. Unlike D-ToF, most of the depth errors remain without being invalidated, which could potentially introduce more errors when used as ground truth during training.



(a) Infrared view (b) Depth value from the sensor (c) Error map (m)

**Fig. 3.19** **Active Stereo Analysis:** Depth quality from the Active Stereo camera. Although the depth map is less affected by challenging materials, its overall quality falls behind both ToF modalities in various aspects, including sharpness, variance, and sparsity.



(a) D-ToF depth warped on RGB view (b) Error map on RGB view (c) Error map original view

**Fig. 3.20** **Active Stereo Analysis after Alignment:** Error after projectively transforming Active Stereo into RGB view. It is worth noting that there is not a significant change in the depth quality after the warping process.

### 3.5.1.3 Active Stereo Camera

The stereo camera, which relies on left and right matching with photometric cues, tends to produce depth maps less affected by challenging materials. This is because the stereo system can use the visible projections of the active patter projector on the surfaces and perform left-right consistency checks to invalidate regions with incorrect depth. As a result, the depth estimation for materials like glass or reflective surfaces is significantly more accurate compared to either of the ToF modalities (Fig. 3.19, green arrow).

However, the stereo camera has its limitations. For scenes that are further away, the quality of the depth map tends to degrade (Fig. 3.19, red arrow). This degradation occurs because the projection pattern becomes attenuated and spread out in the far distance. Additionally, the depth map from the stereo camera can be more blurry, jittery, sparse, and may contain incorrect values in some regions without being invalidated (Fig. 3.19, orange arrow). When used as ground truth, these issues can introduce negative influences, such as blurriness and depth jittering. However, the errors introduced by warping are relatively minor (Fig. 3.20) because the original depth map from the stereo camera is already blurry and sparse.

### 3.5.2 Sensor Impact for Dense 3D Vision Tasks

We train a set of neural networks for monocular depth estimation and implicit scene reconstruction tasks.

### 3.5.3 Depth Estimation

The results for monocular depth estimation with varying training signals are summarized in Table 3.1 and visualized in Fig. 3.21. We present average results for entire scenes as well as separate performance evaluations for background, objects, and materials with different photometric complexities.

The observed errors vary across different scenes and are influenced by the photometric complexity of the scene components. It is noteworthy that depth estimation with ToF training is mainly affected by reflective and transparent object materials, where the active stereo camera can project patterns onto diffusely reflective surfaces. Both self-supervised and semi-supervised setups can recover information in these challenging scenarios. In these cases, they even outperform ToF supervision for photometrically complex objects. On the other hand, simpler structures like the background benefit from ToF supervision. These findings suggest that sensor-specific noise is learned by the models, emphasizing the importance of critically analyzing systematic errors in learning approaches. It also highlights that using 3D devices for ground truth evaluation can lead to incorrect result interpretations, especially when evaluating self-supervised approaches against co-modality sensor data. Furthermore, the results reveal that accurately predicting inter-frame poses in self-supervised indoor setups can be challenging, and precise pose labels can significantly impact depth estimation results (Pose vs. M).

In Table 3.2, we present an extensive quantitative evaluation of supervised training with different depth modalities as supervision signals for different challenging and unseen test

Tab. 3.1 **Depth Prediction Results for Different Training Signals:** Top: Dense depth supervision from different depth modalities. Bottom: Assessment of training with semi-supervised (pose GT) and self-supervised (mono M and mono+stereo M+S) approaches. The evaluation considers the entire scene (Full), background (BG), and objects (Obj) separately. Object materials are further categorized into textured, reflective, and transparent. The **best** and 2nd best RMSE values in millimeters (mm) are highlighted.

| | Training Signal | Full | BG | Obj | Text. | Refl. | Transp. |
|---|---|---|---|---|---|---|---|
| Sup. | I-ToF | 113.29 | 111.13 | 119.72 | 54.45 | 87.84 | 207.89 |
| | D-ToF | 77.97 | **69.87** | 112.83 | **37.88** | 71.59 | 207.85 |
| | Active Stereo | **72.20** | 71.94 | **61.13** | 50.90 | **52.43** | **87.24** |
| Sel/Sem | Pose | **154.87** | **158.67** | **65.42** | **57.22** | **37.78** | 61.86 |
| | M | 180.34 | 183.65 | 85.51 | 84.26 | 48.80 | **49.62** |
| | M+S | 159.80 | 161.65 | 82.16 | 71.24 | 63.92 | 66.48 |

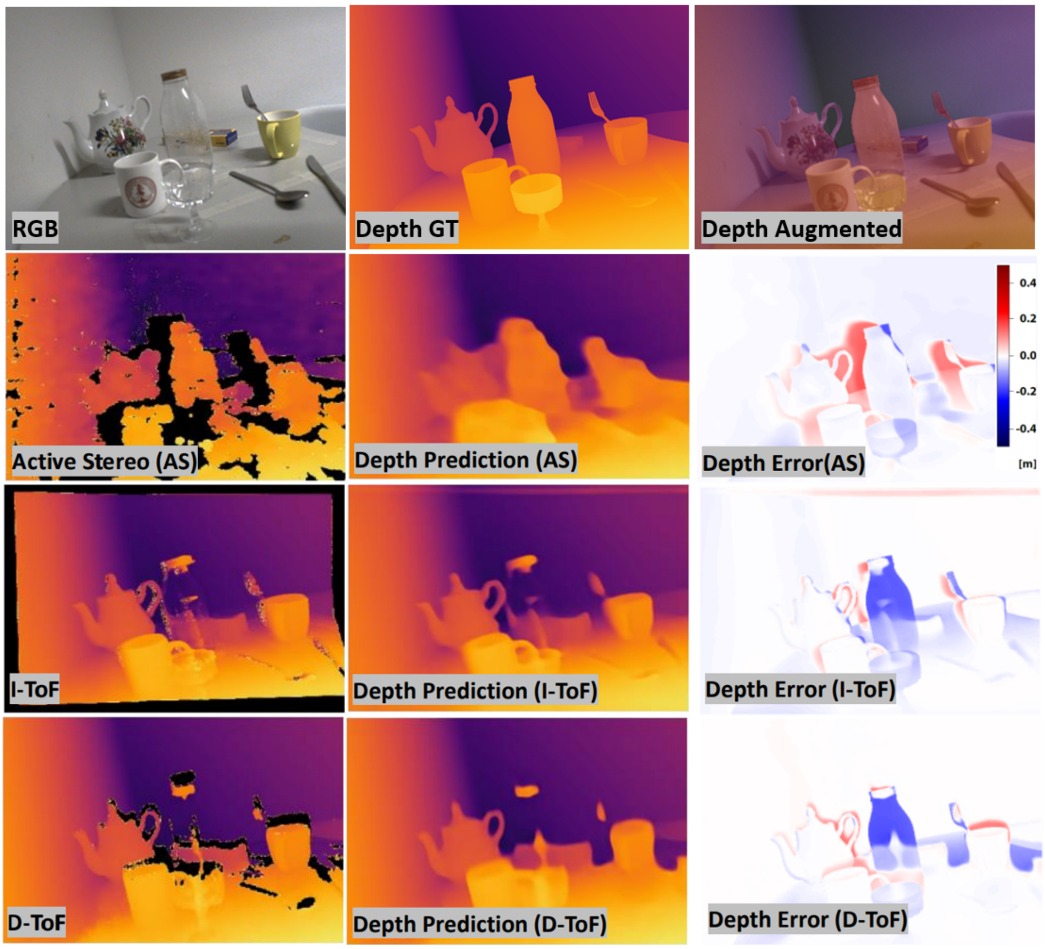


Fig. 3.21 **Fully Supervised Monocular Depth Analysis:** Monocular depth estimation models often exhibit overfitting to the unique noise characteristics of the training sensor. Predictions derived from Active Stereo demonstrate robustness regardless of the material but result in somewhat blurred depth maps. Conversely, both Indirect Time-of-Flight (I-ToF) and Direct Time-of-Flight (D-ToF) models show pronounced material-dependent artifacts but maintain better sharpness along edges.

scenes. These test scenes assess the generalization capability of the trained models to unseen scenarios. We provide an overview of the test scenes and their characteristics:

**Test Scene 1:** This scene shares similarities with the training scenes in terms of its background but introduces additional unseen objects. Moreover, it is observed from viewing angles that are significantly different from the training data.

**Test Scene 2:** In contrast, this scene's background is only partially observed in the training data and primarily includes unseen objects.

**Test Scene 3:** Test Scene 3 is similar to Test Scene 2 but features a modified object layout and challenging lighting conditions due to an additional bright light source above the scene.

Additionally, we introduce an extra test set comprising (partly) seen scenes, consisting of the first 10 frames of each training sequence. These frames have not been utilized during the training process. For the evaluation, we first assess the depth predictions against the rendered

**Tab. 3.2 Comparison of Depth Predictions:** We assess the performance of depth predictions when trained on various modalities and tested on both unseen (Test 1-3) and familiar scenes (Test Seen). (Top) Evaluation against ground truth depth predictions on the test set with dense supervision from different depth modalities. (Bottom) Predictions assessed on the respective modality. Errors are reported as Squared Relative Error (Sq.Rel.) and Root Mean Square Error (RMSE) in millimeters (mm).

| | Mask | Full Scene | | Background | | All Objects | | Textured | | Reflective | | Transparent | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Metric | Sq.Rel. | RMSE | Sq.Rel. | RMSE | Sq.Rel. | RMSE | Sq.Rel. | RMSE | Sq.Rel. | RMSE | Sq.Rel. | RMSE |
| Test 1 | I-ToF | 24.78 | 148.09 | 22.25 | 151.07 | 29.62 | 123.19 | 16.47 | 99.08 | 102.79 | 214.60 | 44.29 | 134.44 |
| | D-ToF | 24.23 | 151.72 | 23.74 | 159.28 | 22.85 | 110.88 | 16.22 | 101.12 | 57.14 | 148.61 | 30.23 | 107.23 |
| | AS | 32.15 | 173.72 | 33.84 | 184.16 | 22.23 | 116.57 | 19.55 | 114.07 | 64.27 | 167.71 | 12.92 | 69.49 |
| Test 2 | I-ToF | 27.42 | 123.79 | 22.66 | 116.86 | 39.85 | 139.67 | 48.66 | 144.92 | 16.15 | 99.44 | 25.15 | 122.25 |
| | D-ToF | 23.00 | 115.40 | 21.18 | 113.27 | 27.89 | 119.59 | 30.00 | 112.92 | 15.81 | 90.89 | 23.73 | 117.72 |
| | AS | 25.94 | 124.17 | 25.50 | 126.28 | 27.18 | 117.04 | 32.81 | 121.24 | 16.40 | 101.86 | 15.73 | 95.27 |
| Test 3 | I-ToF | 36.82 | 152.51 | 35.92 | 153.26 | 38.75 | 147.14 | 34.09 | 127.51 | 20.21 | 110.85 | 55.09 | 183.14 |
| | D-ToF | 32.99 | 145.50 | 35.64 | 153.07 | 25.90 | 120.35 | 19.92 | 96.01 | 21.59 | 105.41 | 37.26 | 149.66 |
| | AS | 31.63 | 141.77 | 35.24 | 151.37 | 22.44 | 110.42 | 23.47 | 106.63 | 14.49 | 94.51 | 21.21 | 109.53 |
| T. Seen | I-ToF | 9.87 | 77.99 | 4.62 | 57.10 | 33.91 | 133.46 | 6.18 | 60.48 | 35.65 | 119.76 | 91.30 | 224.27 |
| | D-ToF | 15.43 | 93.31 | 11.62 | 79.89 | 31.12 | 123.97 | 4.40 | 51.91 | 17.42 | 82.29 | 89.19 | 212.55 |
| | AS | 9.43 | 88.30 | 9.28 | 88.24 | 9.11 | 75.21 | 6.32 | 65.54 | 12.98 | 65.73 | 16.62 | 98.75 |

Tested on Modality:

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Test Seen | I-ToF | 8.34 | 52.29 | 8.57 | 50.00 | 7.01 | 58.85 | 3.80 | 43.44 | 23.28 | 95.38 | 13.69 | 65.41 |
| | D-ToF | 8.05 | 50.43 | 6.82 | 45.50 | 13.52 | 66.34 | 9.00 | 54.15 | 30.91 | 87.71 | 27.92 | 87.32 |
| | AS | 39.25 | 101.76 | 40.87 | 102.29 | 30.32 | 90.00 | 32.24 | 90.49 | 23.36 | 72.21 | 37.25 | 101.23 |
| | GT | 1.12 | 28.81 | 0.71 | 24.41 | 2.65 | 40.41 | 1.83 | 34.89 | 2.16 | 29.55 | 5.02 | 52.43 |

ground truth (Top) and then separately evaluate the predictions using each respective modality (Bottom). This approach helps highlight potential issues related to overfitting to invalid ground truth from individual modalities. Our results reveal that supervision with accurately rendered ground truth yields the best generalization performance for (mostly) unknown scenes. Notably, the active stereo approach produces accurate predictions for transparent objects and performs well with reflective surfaces. In contrast, predictions from I-ToF and D-ToF modalities suffer from incorrect ground truth values for such objects.

In our evaluation, we observe that the (partly) seen scenes generally exhibit lower overall errors across all modalities when compared to the (mostly) unseen test scenes 1, 2, and 3. This suggests that prior exposure to certain scene elements can aid in depth estimation.

Once again, the active stereo approach demonstrates its ability to provide reliable depth supervision for reflective and transparent objects, areas where ToF sensors struggle to provide valid depth information. However, it is worth noting that the active stereo system performs less effectively when predicting the background of the scene. This performance decline may be due to the challenges posed by textureless walls, which remain problematic for the sensor.

When evaluating predictions on the respective modalities themselves, we encounter an overfitting issue stemming from incorrect depth values provided by the sensors. Specifically, in cases where a particular sensor fails to provide accurate depth values (e.g., transparent objects for

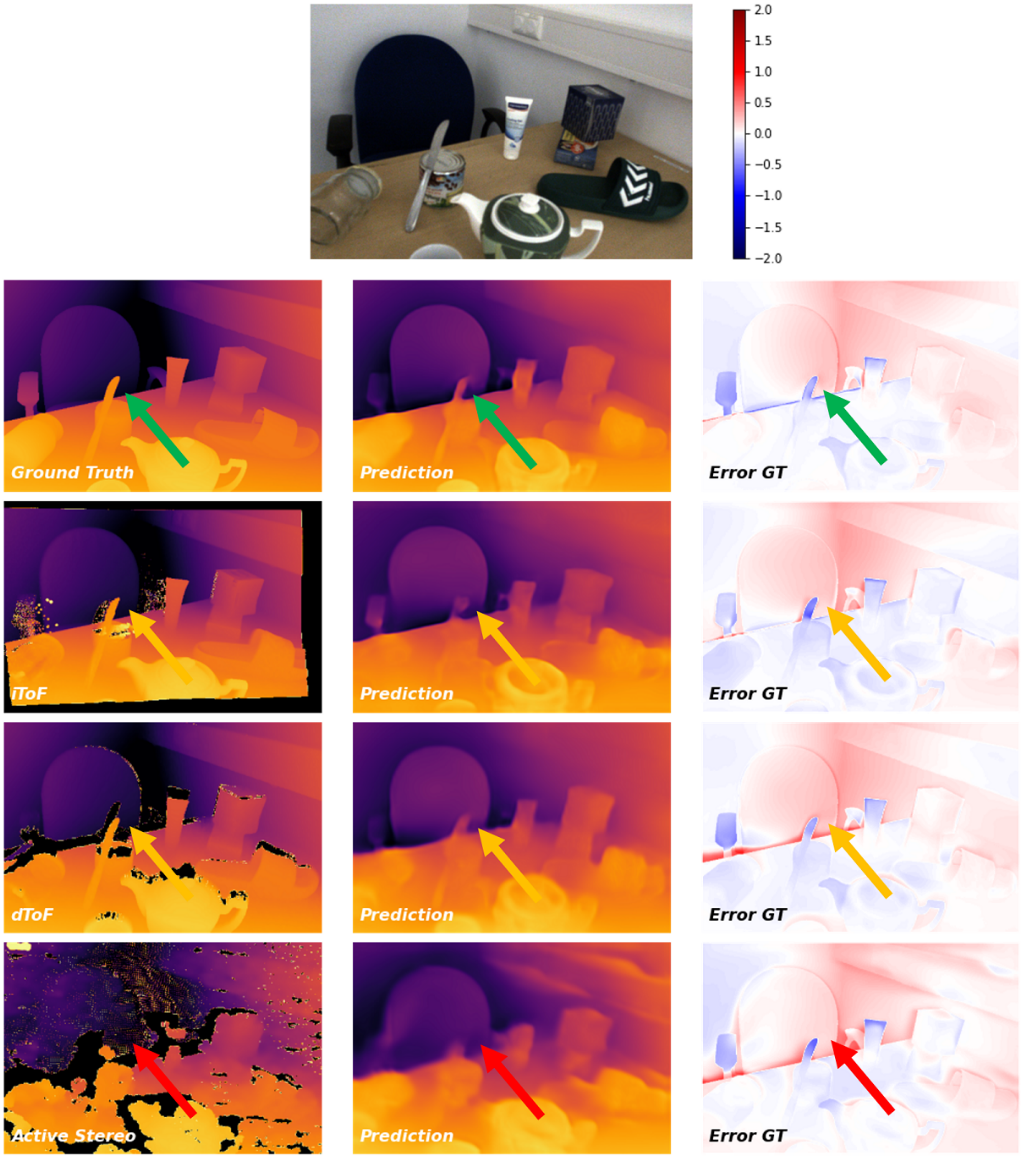


Fig. 3.22 **Qualitative Evaluation on Test Scene 1:** This figure presents a qualitative comparison involving various depth modalities. It showcases the network's depth predictions trained under the supervision of each modality, accompanied by a visual representation of the corresponding errors for a comprehensive evaluation.

I-ToF or reflective objects for D-ToF), we observe significantly lower errors. This phenomenon suggests that the models tend to overfit to the characteristics of the specific sensor modality.

Figs. 3.22, 3.23, and 3.24 present predictions on exemplary frames from test scenes 1, 2, and 3, respectively. These figures include various sensor modalities and error plots that compare the predictions against the ground truth. Training with rendered ground truth generally leads to the best performance. Both ToF sensors exhibit incorrect depth values for reflective or transparent objects, resulting in inaccurate predictions within these regions (see Fig. 3.22).

In the case of training with active stereo as supervision, the predictions tend to appear blurrier and exhibit less distinct edges at depth boundaries when compared to other modalities.

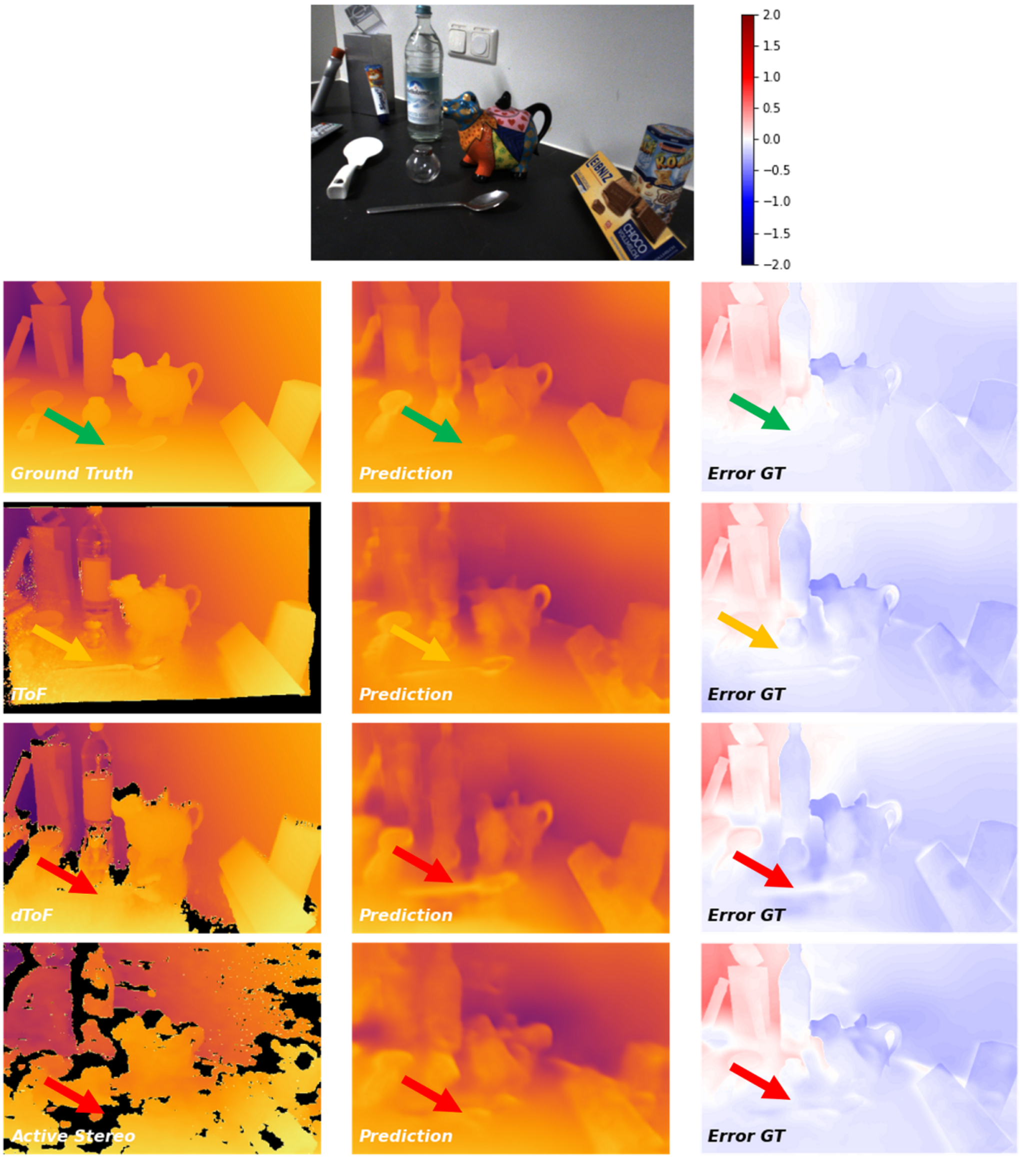


Fig. 3.23 **Qualitative Evaluation on Test Scene 2:** This figure presents a qualitative comparison involving various depth modalities. It showcases the network's depth predictions trained under the supervision of each modality, accompanied by a visual representation of the corresponding errors for a comprehensive evaluation.

This phenomenon may arise from the sensors' invalidation of many depth pixels near such boundaries (see Fig. 3.23).

Test scene 3, with its challenging conditions such as bright lighting and numerous unseen objects, is difficult to predict for all training setups (see Fig. 3.24). Similar artifacts, as described previously, are observed. Additionally, the unseen trophy object, which features reflective and transparent materials, exhibits substantial errors in sensor inputs and predictions. The desk surface is also inaccurately captured by the D-ToF sensor due to significant reflections and MPI originating from the background.

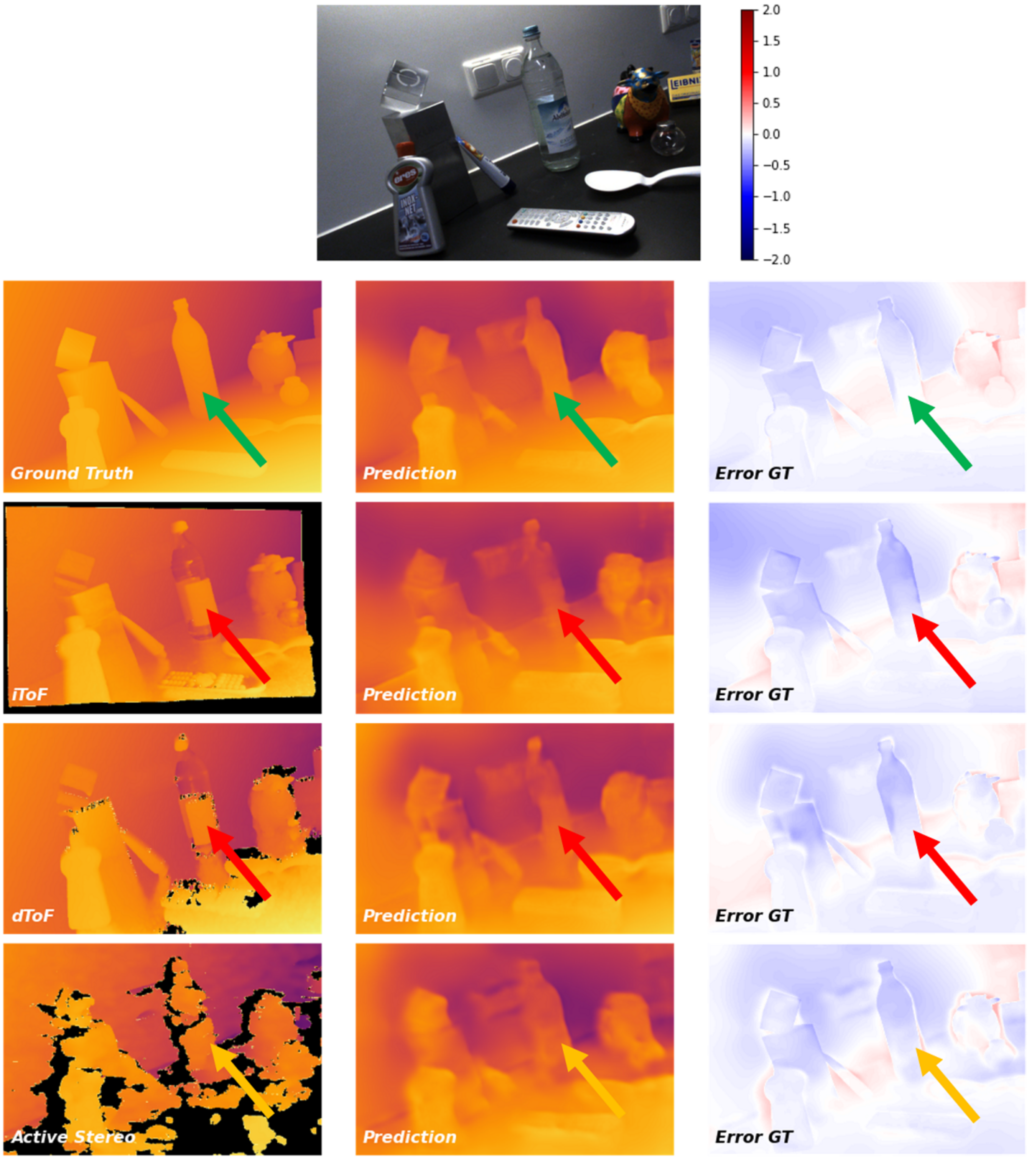


Fig. 3.24 **Qualitative Evaluation on Test Scene 3:** This figure presents a qualitative comparison involving various depth modalities. It showcases the network's depth predictions trained under the supervision of each modality, accompanied by a visual representation of the corresponding errors for a comprehensive evaluation.

### 3.5.4 Implicit 3D Reconstruction & View Synthesis

Our implicit 3D reconstruction method generates novel views for depth, surface normals, and RGB, each with varying levels of quality. When trained solely on color information, the NeRF model produces RGB views that are visually convincing and achieve the highest PSNR (cf. Fig. 3.25 and Tab. 3.3). However, the reconstruction of the 3D scene geometry is suboptimal. Consistent with existing literature [28, 131], depth regularization improves the reconstruction quality, particularly in texture-less regions.

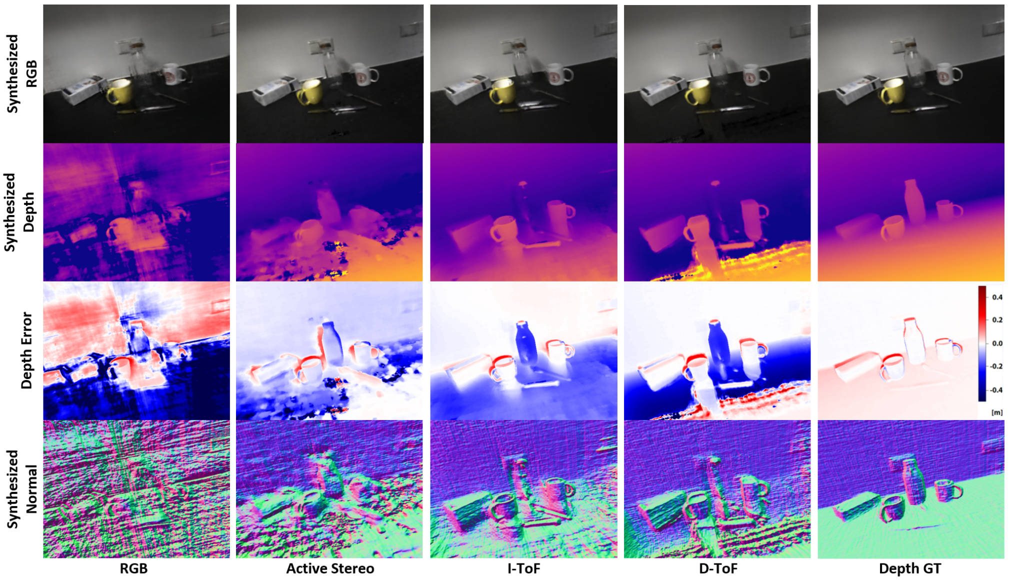


Fig. 3.25 **Implicit Scene Reconstruction with Neural Radiance Field (NeRF):** The results depict the synthesis for depth, surface normals, and RGB images for an unseen view. These synthesized images are presented alongside prediction error visualizations. The columns enable a comparison of different training methods: the first column illustrates a NeRF [111] trained solely on RGB data, and the other columns shows results when various depth maps are used for regularization, as proposed in [131], and the last column demonstrates synthesized results from training with ground truth (GT) depth for comparison. Notable differences are visible, particularly in regions such as the partly reflective table edges, the translucent bottle, and around depth discontinuities.

The application of regularization using different depth modalities reveals the inherent sensor noise in I-ToF, Active Stereo (AS), and D-ToF depth data. While RMSE values show a similar trend to the monocular depth prediction results, with AS performing the best, followed by D-ToF and I-ToF, the cosine similarity metrics for surface normal estimates also confirm this trend. The overall depth and normal reconstructions obtained from AS data appear pretty noisy. However, the depth error metrics are more sensitive to significant errors in estimating depth, especially in the case of reflective and translucent objects. Previous artifacts from the respective depth sensors can influence the NeRF model and result in incorrect scene reconstructions. These include errors caused by D-ToF and I-ToF for translucent materials, noisy backgrounds, and

Tab. 3.3 **Evaluation of Novel View Synthesis from Implicit 3D Reconstruction:** Comparison of RGB, depth, and surface normal estimates using various optimization strategies. These strategies include training with RGB-only supervision and incorporating respective sensor depth information. The results are evaluated against ground truth, with the **best**, 2nd best, and 3rd best performance indicated. Depth metrics are reported in millimeters (mm).

| | RGB | | Depth | | | | Normal |
|---|---|---|---|---|---|---|---|
| Modality | PSNR ↑ | SSIM ↑ | Abs.Rel.↓ | Sq.Rel.↓ | RMSE ↓ | $\sigma < 1.25$ ↑ | Cos.Sim.↓ |
| RGB Only | **32.406** | 0.889 | 0.328 | 111.229 | 226.187 | 0.631 | 0.084 |
| + AS | 17.570 | 0.656 | 0.113 | 16.050 | 94.520 | 0.853 | 0.071 |
| + I-ToF | 18.042 | 0.653 | 0.296 | 91.426 | 217.334 | 0.520 | 0.102 |
| + D-ToF | 31.812 | 0.888 | 0.112 | 24.988 | 119.455 | 0.882 | 0.031 |
| + Syn. | 32.082 | **0.894** | **0.001** | **0.049** | **3.520** | **1.000** | **0.001** |

inaccurate depth discontinuities at object edges for AS data. Interestingly, leveraging D-ToF data as a depth prior can improve the overall scene reconstruction in most parts of the scene but falls short for the bottle object, where AS provides better depth priors. This difference is also evident in the synthesized depth maps. Many of these issues are mitigated when utilizing synthetic depth ground truth (as shown in the last row), resulting in improved view synthesis quality, as indicated by higher SSIM values.

## 3.6 Conclusion

We introduced an innovative annotation and data acquisition pipeline that significantly enhances the precision and realism of 3D vision datasets. This approach involves the integration of robotic forward-kinematics or external infrared trackers, along with improved calibration and annotation techniques, resulting in valuable tools for dataset generation. Incorporating external infrared trackers is especially noteworthy as it expands the coverage of camera poses and viewpoints. This improvement addresses limitations observed in existing datasets, eliminating the requirement for checkerboards or reference objects to be present in the scene. These principles hold significant promise, particularly when creating datasets that involve objects with high photometric complexity, such as those made of glass, reflective materials, or textureless surfaces (as illustrated in Figs. 3.3 and 3.9).

This detailed analysis of sensor artifacts and the findings of 3D vision tasks trained on the data underscore the importance of questioning and thoroughly investigating commonly used 3D sensors to gain a deeper understanding of their impact. For the first time, we have created a framework that enables the systematic study of how sensor characteristics influence the learning process in these domains, providing objective insights into their effects. We have quantified the impact of various photometric challenges, such as translucency and reflectivity, on tasks like depth estimation, reconstruction, and novel view synthesis. We have also introduced a unique dataset that serves as a valuable resource to stimulate further research in this and other directions. While it may not surprise that sensor noise affects these tasks, our dataset allows for the first quantification of this impact. Notably, we have observed that D-ToF supervision outperforms AS by a significant margin (13.02 mm) for textured objects, while AS, in turn, surpasses I-ToF by 3.55 mm RMSE (as shown in Table 3.1). This trend holds even for mostly texture-less backgrounds, where D-ToF exhibits a 37% higher accuracy than I-ToF.

In addition to our investigations and the evaluation of sensor signals for standard 3D vision tasks, our dataset has the potential to open up new avenues for exploring cross-modal fusion pipelines. Specifically for the detailed and robust 3D understanding of scenes with multiple objects, such data is invaluable for 6D object pose estimation methods.

# Part IV

## Polarization Properties for 6D Object Pose Estimation

**How can we integrate the physical properties of polarized light into a learning pipeline for robust 3D perception tasks like 6D object pose estimation?**

**Can we avoid the need for annotated real data - potentially by leveraging polarization for self-supervision?**

# 4 Physical Priors in 3D Perception for Robust 6D Object Pose Estimation

> *“fiat lux”*
>
> – **"let there be light"**
> 1. Book of Moses

## 4.1 Introduction

Besides estimating the pixel-wise depth of a scene, 3D perception also includes accurately determining the position and orientation of specific objects within the scene, known as 6D object pose estimation, which is essential for applications ranging from AR/VR to robotics. Current methods usually rely on RGB-only or RGB-D data, which often struggle with photometrically challenging objects such as those with reflective, transparent, or textureless surfaces due to artifacts in the depth information, as discussed in the previous chapter.

To address these challenges, we propose the integration of polarimetric imaging as a novel modality in 6D object pose estimation. Polarimetric images capture the polarization state of light, which encodes robust surface and shape information. This multi-modal approach can benefit this task, especially when considering photometrically challenging objects.

### 4.1.1 Motivation

Current RGB-D and RGB-only 6D object pose estimation techniques show limited accuracy and robustness when confronted with photometrically challenging objects. RGB-D methods utilize depth information and often struggle with such objects due to transparent or reflective surfaces. While avoiding some of these issues, RGB-only methods typically have difficulties in accurately estimating the pose of textureless objects. We aim to overcome these challenges by exploring a novel approach that leverages polarization images and robust physical cues, thereby enhancing the accuracy and robustness of pose estimation for objects that are otherwise difficult to handle.

Light has captivated humanity for centuries and has been at the heart of numerous significant scientific discoveries. In the context of computer vision, typical light sensors measure the wavelength and energy of light to determine color and intensity within a specific spectrum. However, these are not the only attributes of an electromagnetic (EM) wave. Polarization, defined as the oscillation direction of the EM field relative to the direction of the light ray, is another crucial property of light. Most natural light sources, like the sun or artificial lights, emit unpolarized light, where the light wave oscillates in multiple directions. When light reflects off a surface, it becomes partially or fully polarized. This phenomenon means that polarization carries valuable information about the surface structure, material properties, and the angle of reflection [82]. This information is beneficial for dealing with photometrically challenging objects made of metallic, reflective, or transparent materials. These objects often pose significant challenges to standard vision pipelines, limiting their effectiveness and robustness in applications that require precise pose estimation. By incorporating polarization data into our pipeline, we aim to extract this additional information from scenes and objects, particularly those involving photometrically challenging surfaces or materials.

### 4.1.2 Contributions

While several 6D object pose estimation pipelines have been developed [15, 29, 60, 106], also targeting texture-less objects [31, 62], the challenge of photometrically complex objects, characterized by high reflectance and partial transparency, has only recently gained attention in research [101]. These objects present unique challenges to RGB-D sensing [101]. In response, our approach extends beyond conventional methods that rely on RGB or depth information. We leverage the polarization properties of light as an additional data source, yielding surface normal estimations. By doing so, we create a hybrid method that combines a physical model with a data-driven learning approach, significantly enhancing 6D pose estimation capabilities. This method proves particularly effective not only for photometrically challenging objects but also in improving pose accuracy and robustness for more traditional object types.

The primary contributions in this chapter are in summary:

Contributions

1. We propose the use of **polarization** as a novel modality for **6D object pose estimation**, exploring its benefits over RGB-only and RGB-D modalities.
2. We design a **hybrid pipeline** for **instance-level** 6D pose estimation that integrates polarization cues. This combination of a **physical model with a learning-based approach** shows significant improvements, especially for **photometrically challenging objects** with high reflectance and translucency.
3. We have constructed the first **polarimetric instance-level 6D object pose estimation benchmark**.

## 4.2 Related Work

We review the literature on polarimetric imaging and 6D object pose estimation and cover datasets to offer a comprehensive insight into the research domain.

### 4.2.1 Polarimetric Imaging

**Polarization for 2D** Polarization cues provide a valuable source of complementary information in 2D computer vision, particularly for tasks involving photometrically challenging objects, such as those that are reflective or transparent. This utility has inspired various research efforts, including segmentation tasks [82] to handle reflective and transparent objects effectively.

Another significant advantage of polarization is its ability to mitigate glare. Lei et al. [93] demonstrated how specific polarization filters can effectively remove reflections from images, enhancing the clarity and quality of the captured data. While the deployment of a single polarization camera can already substantially improve upon traditional photometric acquisition setups, the integration of multispectral polarimetric light fields, as investigated by Islam et al. [72], takes this a step further. This advanced approach combines polarization with multispectral imaging and light field technology, leading to even more significant enhancements in performance.

**Polarization for 3D** Previous research in shape from polarization (SfP) investigated to extract surface normals and depth information from polarimetric data, given its intrinsic link to the object's surface properties. However, early works in this area faced limitations due to model ambiguities and often relied on controlled setups. Classical SfP methods typically used an orthographic camera model and were constrained to lab scenarios with controlled environmental conditions [4, 44, 146, 185]. Yu et al. [185] mathematically related polarization intensity to surface height, optimizing for depth in controlled scenarios. In contrast, Atkinson et al. [4] focused on recovering surface orientation for fully diffuse surfaces. While these methods predominantly utilized monocular polarization, combining more than one view with physical models for SfP has been explored as well [3, 24]. This multi-view approach also lends itself to self-supervised methods, like the one proposed by Verdie et al. [161]. Some studies have investigated the integration of complementary techniques like photometric stereo [2] and hybrid RGB+Polarization (RGB+P) methods [196]. These hybrid approaches can provide metrically accurate depth estimates, especially if the light direction is known. Additionally, existing depth maps can be refined using polarimetric cues [81]. In scenarios where the scene is assumed to be fully diffuse, the polarimetric sensing model can also aid in estimating the relative transformation of a moving polarization sensor [25]. Data-driven approaches can mitigate assumptions regarding surface properties, light direction, and object shapes. For instance, Ba et al. [5] developed a method for estimating surface normals by presenting a neural network with a set of plausible cues, enabling SfP even with ambiguous data. Our research draws inspiration from these various approaches to enhance our object pose estimation pipeline with physical priors. Unlike previous studies, we focus on object poses in an unconstrained setup without making assumptions about reflection properties or lighting conditions. The insights gained from past research allow us to design a pipeline capable of addressing pose prediction challenges for photometrically complex objects for the first time.

### 4.2.2 6D Pose Prediction

**Monocular RGB** Methods predicting 6D pose from a single image fall into different categories based on their approach: direct pose optimization, learning a pose embedding, or establishing correspondences between 3D models and 2D images.

Direct Pose Optimization: Some methods directly regress the 6D pose [90, 96, 106, 174], or discretize the task into a classification problem [15, 84]. These networks predict pose parameters as $SE(3)$ elements, reflecting the training parameterization. Implicit learning of pose parameterization is also explored [195].

Learning Pose Embeddings: This approach involves learning an implicit space that encodes the pose, from which predictions are decoded [157, 158, 173].

Establishing 2D-3D Correspondences: Contemporary and high-performing methods in this category typically adopt a two-stage approach. Initially, a network predicts 2D-3D correspondences between the image and the 3D model. These correspondences are then processed with algorithms like RANSAC/P$n$P [36, 94], the Umeyama algorithm [159], or direct regression to determine the 6D object pose. Approaches vary between sparse [67, 123, 126, 149] and dense correspondences [61, 98, 119, 144, 186], with some aiming for end-to-end learning [29, 66, 165]. ZebraPose [154] introduces hierarchical feature representations, and there is growing interest in zero-shot methods for 6D pose estimation [143]. A common challenge in correspondence-based methods [61, 98, 119, 144, 186] is the computationally intensive post-processing, typically involving RANSAC-based pose solvers. To address this, GDR-Net [165] and SO-Pose [29] employ learning-based MLP networks to directly predict the target pose from dense correspondences, enhancing computational efficiency.

**RGB-D and Refinement** Monocular pose estimation from RGB images is inherently challenging due to its ill-posed nature, where crucial depth information is missing. In this context, depth maps emerge as a valuable asset, providing essential geometric insights that aid in identifying point correspondences critical for accurate pose estimation [32]. Integrating RGB data can further enrich this geometric information [10]. While it is possible to deduce poses from depth information alone or combined RGB-D datasets, many RGB-focused methods [90, 98, 119, 158] greatly benefit from refining with depth based ICP [8] or leveraging indirect multi-view cues [90]. The synergistic use of RGB and depth data is particularly evident in pioneering works like DenseFusion [162], where encoded features from both modalities are fused. Furthering this approach, FFB6D [57] tightly couples cross-modal information across multiple feature layers, enhanced with a keypoint extraction process [58] that utilizes both geometric and texture-based cues. Other methods, including Uni6D [74], ESA6D [112], FS6D [183], and DGECN [18], also incorporate depth data into their prediction models. However, a critical limitation of these techniques is their reliance on the quality of the input. Depth sensing can be unreliable in scenarios with photometric challenges, such as reflective surfaces. In such cases, more robust polarimetric shape cues can offer a substantial advantage.

**Photometric Challenges** The field of 6D pose estimation predominantly relies on well-established datasets that provide RGB-D input [11, 60, 83, 174]. These datasets have been instrumental in testing and validating various pose estimation methodologies. Additionally, datasets featuring photometrically challenging objects, such as texture-less and reflective industrial parts, have been made publicly available [31, 62]. These datasets typically do not include polarization data, which could provide valuable additional information for pose estimation tasks.

The challenge posed by transparent objects in the context of pose estimation has already been addressed in earlier works [138], where robotic grasp points on objects using RGB stereo images are determined without relying on a 3D model. [124] showed how transparent objects with rotational symmetry could be reconstructed from two views using edge detection and contour fitting. Recent developments in this field include the work on KeyPose [102], which explores instance and category-level pose prediction from RGB stereo. Due to the limitations of their depth sensor with transparent objects, they used an opaque-transparent object pair to establish ground truth depth. ClearGrasp [137] is an RGB-D method for transparent objects. The StereOBJ-1M dataset represents a significant advancement in the field, featuring transparent, reflective, and translucent objects under varying illumination conditions and symmetries. It utilizes a binocular stereo RGB camera for pose estimation, addressing many of the challenges posed by these objects. However, despite these advancements, none of the existing datasets include RGBP (RGB + Polarization) data, which could provide richer and more nuanced information for 6D pose estimation, especially for challenging object surfaces.

## 4.3 Polarimetric Physical Conditions

Standard sensors emit or receive light to measure parameters such as wavelength and energy within a specific spectrum. In addition to these fundamental characteristics, the relative oscillation amplitude of the electromagnetic wave defines its polarization attributes. Natural light, initially unpolarized, undergoes polarization upon reflection from a surface, thereby encoding valuable information regarding the surface properties of objects. The use of RGB-D sensors in pose estimation has gained popularity due to their low cost and adaptability to a wide range of devices. However, they are susceptible to photometric challenges, including translucency and reflections, which can lead to inaccuracies in depth estimation.

### 4.3.1 Photometric Challenges for RGB-D

Commercial depth sensors typically employ photometric measurements for depth estimation utilizing active illumination techniques such as pattern projection (e.g., Intel RealSense D

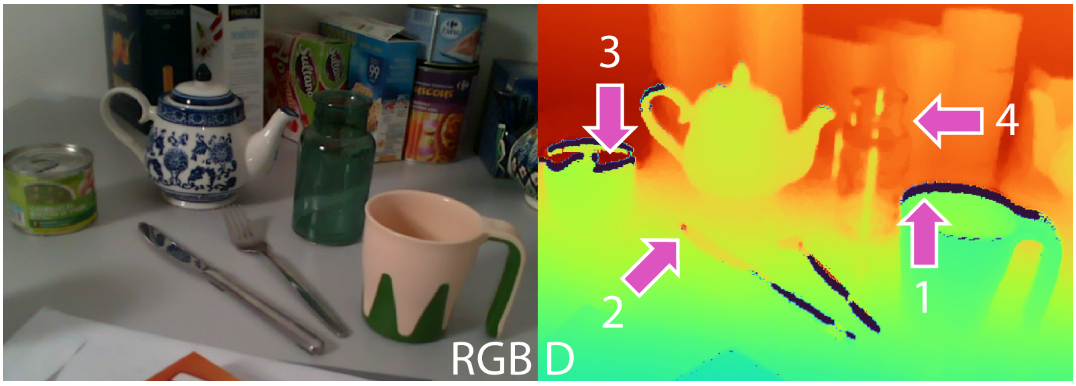


Fig. 4.1 **Depth Artifacts:** The RealSense L515 depth sensor demonstrates inaccuracies in calculating depth values for common household objects. Specifically, it encounters issues related to boundaries (1,3) that lead to the invalidation of pixels and strong reflections (2,3), resulting in depth estimates significantly different from the actual values. Additionally, the depth sensor struggles to detect semi-transparent objects such as the vase (4), leading to partial invisibility and inaccurate measurements of the distance to objects located behind them.

series), or time-of-flight (ToF) measurements (e.g., Kinect v2 / Azure Kinect, Intel RealSense L series). However, these methods face challenges when reflections artificially extend the roundtrip time of photons or translucent objects degrade the projected pattern. Fig. 4.1 provides an illustration using common household objects. In the experiment, the ToF sensor (RealSense L515) struggles to detect the semi-transparent vase, rendering it almost invisible to the sensor. Furthermore, reflections on the cutlery can result in depth estimates significantly deviating from the actual values, and strong reflections at boundaries can invalidate pixel distances.

Recognizing the potential of polarization in addressing the challenges of object geometry recovery in complex environments, the next logical step in our research is to delve into integrating shape cues from polarization for enhanced 6D object pose estimation. To embark on this path, it is essential first to understand the fundamentals of polarimetric image formation.

### 4.3.2 Polarization Model

Most artificial and natural light is unpolarized, signifying that the electromagnetic wave oscillates along all planes perpendicular to the direction of light propagation [37]. When unpolarized light interacts with a linear polarizer or is reflected at Brewster's angle from a surface, it transforms polarized. The refractive index of a material plays a role in determining the speed of light propagation through it, the extent of reflection, and the medium's Brewster's angle. When light is reflected at the same angle as the incident ray relative to the surface normal, it is referred to as specular reflection. The remaining portion penetrates the object as refracted light, which becomes partially polarized during its passage through the medium. Eventually, this partially polarized light wave exits the object, leading to what is known as diffuse reflection [37]. To illustrate these concepts, please refer to Fig. 4.2. In the case of real physical objects, the resulting reflection is typically a combination of both specular and diffuse reflection, with the proportion of each being influenced by factors such as the refractive index

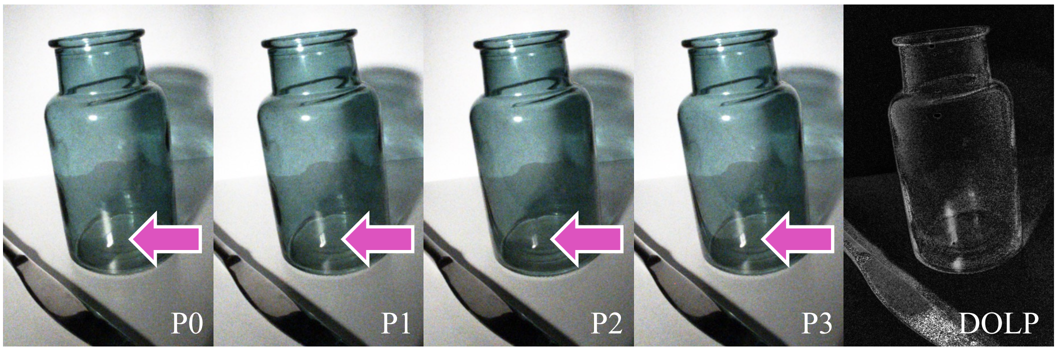


Fig. 4.2 **Degree of Polarization:** The polarization state of light undergoes alteration upon reflection from a translucent surface, resulting in discernible distinctions within the polarimetric image quadruplet with different polarization filter angles. These distinctions are directly linked to the orientation of the surface normal. More precisely, the degree of polarization (DoP) for both translucent and reflective surfaces markedly exceeds that observed in other image regions, as evident from the highlighted areas.

and the incident light angle. We propose using surface normals derived from polarization data to address RGB-D sensors' photometric challenges. Our proposed method holds applicability across various domains, including pose estimation, where the precision of 3D information is paramount.

### 4.3.3 Image Formation Model

We introduce the foundational polarization image formation model. When incident light with a specific intensity $I$ and wavelength $\lambda$ reaches the sensor, it traverses the color filter array (CFA), which separates the light into RGB wavebands. The incoming light also possesses a degree of polarization (DoP) denoted as $\rho$ and an angle of polarization (AoP) represented by $\phi$. As the light proceeds through a polarizer array positioned above a pixel unit equipped with four distinct polarization angles, $\varphi_{pol} \in \{0°, 45°, 90°, 135°\}$, the oscillation state of the light is recorded alongside its wavelength and energy [82]. The polarization image formation model, as expressed in Equation 4.1, defines the underlying parameters that contribute to the recorded polarized intensities as follows:

$$I_{\varphi_{pol}} = I_{un} \cdot \; (1 + \rho \; \cos(2(\phi - \varphi_{pol}))), \tag{4.1}$$

where the unpolarized intensity $I_{un}$ can be computed by averaging over polarized intensities $I_{\varphi_{pol}}$ measured at various polarization filter angles $\varphi_{pol} \in \{0°, 45°, 90°, 135°\}$.

We determine the values of $\phi$ and $\rho$ by reformulating the image formation model as follows:

$$
\begin{aligned}
I_{\varphi_{pol}} &= \frac{I_{max}+I_{min}}{2} + \frac{I_{max}-I_{min}}{2}\cos(2(\phi-\varphi_{pol})) && (4.2)\\
&= \begin{bmatrix} 1 \\ \cos 2\varphi_{pol} \\ \sin 2\varphi_{pol} \end{bmatrix}^T \begin{bmatrix} \frac{I_{max}+I_{min}}{2} \\ \frac{I_{max}-I_{min}}{2}\cos 2\phi \\ \frac{I_{max}-I_{min}}{2}\sin 2\phi \end{bmatrix}. && (4.3)
\end{aligned}
$$

In this formulation, we define the degree of polarization (DoP) $\rho$ as:

$$\rho = \frac{I_{max}-I_{min}}{I_{max}+I_{min}}. \tag{4.4}$$

We express the unpolarized intensity as the average of the maximum and minimum values, as follows:

$$I_{un} = \frac{I_{max}+I_{min}}{2}. \tag{4.5}$$

The degree of polarization (DoP) $\rho$ and angle of polarization (AoP) $\phi$ can be determined through an over-determined linear least squares system [69] applied to a collection of polarization images captured under various polarization filter angles, as:

$$\begin{bmatrix} I_{\varphi_{pol,1}} \\ \vdots \\ I_{\varphi_{pol,4}} \end{bmatrix} = \begin{bmatrix} 1 & \cos 2\varphi_{pol,1} & \sin 2\varphi_{pol,1} \\ & \vdots & \\ 1 & \cos 2\varphi_{pol,4} & \sin 2\varphi_{pol,4} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}. \tag{4.6}$$

The unknowns $x_i$ are defined as follows: $x_1 = I_{un}$, $x_2 = I_{un}\rho\cos 2\phi$, and $x_3 = I_{un}\rho\sin 2\phi$.

From Equation 4.6, we can solve for $[x_1, x_2, x_3]^T$, and subsequently retrieve the values of $\rho$ and $\phi$ as follows:

$$\begin{cases} I_{max} &= x_1 + \sqrt{x_2^2 + x_3^2} \\ I_{min} &= x_1 - \sqrt{x_2^2 + x_3^2} \\ \rho &= \frac{\sqrt{x_2^2+x_3^2}}{x_1} \\ \phi &= \frac{1}{2}\arctan\frac{x_3}{x_2} \end{cases} \tag{4.7}$$

It is important to note that both $\phi$ and $\phi+\pi$ can satisfy the polarization image formation model presented in Equation 4.1. This phenomenon is commonly referred to as the $\pi$-ambiguity.

### 4.3.4 Shape Priors from Polarization

With the available polarization information, we can extract object shape details by calculating the azimuth and zenith angles of the object surface normal for both the diffuse and specular reflection scenarios. The surface normal, by definition, is a vector perpendicular to the tangent plane at a given point. We conventionally choose the outward-pointing normals, which can be characterized by the azimuth angle $\alpha$ and the zenith angle $\theta$, as defined below:

$$\vec{n} = \begin{bmatrix} n_x \\ n_y \\ n_z \end{bmatrix} = \begin{bmatrix} \cos\alpha\sin\theta \\ \cos\alpha\cos\theta \\ \cos\theta \end{bmatrix}, \tag{4.8}$$

where:

- azimuth angle $\alpha \in [0, 2\pi]$
- zenith angle $\theta \in [0, \pi]$

We determine the values of $\varphi$ and $\rho$ by solving the over-determined system of linear equations using the linear least squares method. The calculation of AoP depends on the surface properties and is performed as follows:

$$\begin{cases} \phi_d[\pi] & = \alpha & \text{for diffuse reflection} \\ \phi_s[\pi] & = \alpha - \frac{\pi}{2} & \text{for specular reflection.} \end{cases} \tag{4.9}$$

Here, $[\pi]$ indicates the $\pi$-ambiguity, and $\alpha$ represents the azimuth angle of the surface normal $\mathbf{n}$. We can establish a connection between the viewing angle $\theta \in [0, \pi/2]$ and the degree of polarization (DoP) by considering Fresnel coefficients. Consequently, DoP is similarly expressed as follows [4]:

$$\begin{cases} \rho_d = \frac{(\eta - 1/\eta)^2 \sin^2(\theta)}{2 + 2\eta^2 - (\eta + 1/\eta)^2 \sin^2(\theta) + 4\cos(\theta)\sqrt{\eta^2 - \sin^2(\theta)}} \\ \\ \rho_s = \frac{2\sin^2(\theta)\cos(\theta)\sqrt{\eta^2 - \sin^2(\theta)}}{\eta^2 - \sin^2(\theta) - \eta^2\sin^2(\theta) + 2\sin^4(\theta)} \end{cases} \tag{4.10}$$

with the refractive index of the observed object material denoted as $\eta$. The values used for our objects can be seen in Tab. 4.1.

By solving Equation 4.10 for $\theta$, we obtain three solutions: $\theta_d, \theta_{s1}, \theta_{s2}$, one corresponding to the diffuse case and two to the specular cases. For each of these cases, we can subsequently determine the 3D orientation of the surface by calculating the surface normals as:

$$n = (\cos\alpha\sin\theta, \sin\alpha\sin\theta, \cos\theta)^{\mathrm{T}}. \tag{4.11}$$

Tab. 4.1 **Refractive Indices:** Refractive indices per object with specific material used for the physical model.

| Object | Material | Refractive Index |
|---|---|---|
| Teapot | ceramic | 1.54 |
| Can | aluminium composite | 1.35 |
| Fork | stainless steel | 2.75 |
| Knife | stainless steel | 2.75 |
| Bottle | glass | 1.52 |
| Cup | plastics | 1.50 |

We utilize these plausible normals $n_d, n_{s1}, n_{s2}$ as physical priors per pixel, which are later provided as inputs to a neural network for 6D object pose estimation of photometrically challenging objects.

With the assistance of the physical model defined above, we can now deduce physical polarimetric characteristics that encapsulate shape information in the form of geometric normals. More precisely, when light is reflected from the surface of an object, the shape details are encoded in the captured polarization intensities accordingly.

## 4.4 Polarimetric Physical Conditions for 6D Object Pose Prediction

This section introduces our **P**olarimetric **P**ose **P**rediction **Net**work, abbreviated as **PPP-Net**. Given a set of polarimetric images captured at four different angles, namely $I_0, I_{45}, I_{90}, I_{135}$, along with the calculated values of AoP $\phi$, DoP $\rho$, and the normal maps $N_d, N_{s1}, N_{s2}$ used as physical priors, our objective is to employ a neural network to learn a transformation $P = [R|t]$ that maps a target object from the object frame to the camera frame, leveraging a 3D CAD model of the object.

Our pipeline's physical model unveils the implicitly encoded shape information, offering object-centric priors orthogonal to intensity information. We derive a set of explicit object shape priors denoted as $N_i$ based on polarimetric intensities $I_{\varphi_{pol}}$ and properties $\rho, \phi$ [5, 197]. The inherent ambiguities within this process can result in non-unique solutions, as discussed in [5]. Nevertheless, we encode these ambiguities on a per-pixel basis to guide the network in distinguishing between different priors and extracting meaningful geometric features.

### 4.4.1 Network Architecture

Our network architecture is illustrated in Fig. 4.3. The network is divided into two distinct encoders in its initial part, each with specific responsibilities. The first encoder processes the joint polarization information derived from the raw polarimetric images concatenated with the

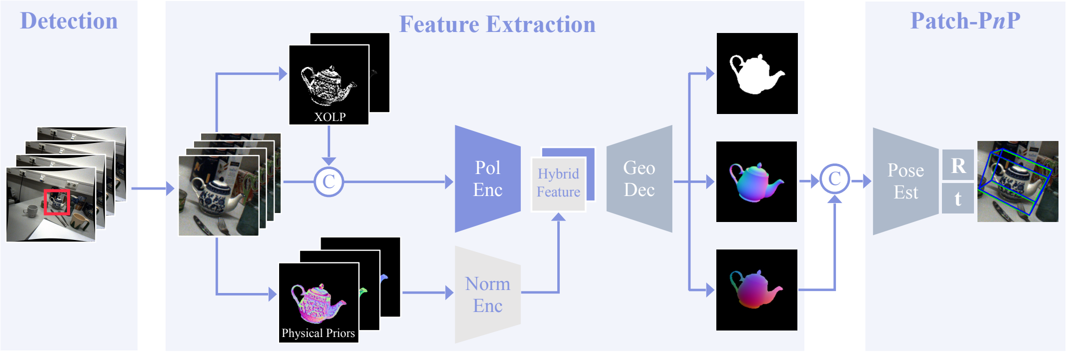


Fig. 4.3 **PPP-Net:** Our **P**olarimetric **P**ose **P**rediction **Net**work harnesses the potential of RGBP images, which constitute a set of four distinct polarized RGB images. These images facilitate the computation of AoP/DoP and normal maps via our physical model. Within our hybrid framework, the polarized data and inherent physical signals are encoded distinctly and subsequently integrated. The decoder is then tasked with predicting the object mask, the normal map, and the NOCS. The culmination of this process results in the accurate prediction of the 6D object pose, achieved through Patch-P*n*P [165].

computed AoP/DoP maps. Meanwhile, the second encoder handles physical priors, specifically the physical normals calculated from polarimetric images using the physical model. In both cases, the encoding is focused on a zoomed-in region of interest (ROI) with dimensions of 256 × 256 pixels.

Subsequently, the encoded information from both encoders is combined and passed to a decoder. The decoder takes in the jointly encoded information and further enhances it with data from skip connections originating from various hierarchical levels within the encoders. The decoder produces an object mask, a normal map, and a 3-channel dense correspondence map (NOCS). The NOCS map establishes a correspondence between each pixel and its normalized 3D coordinate.

The predicted normal map and NOCS, concatenated with corresponding 2D-pixel coordinates, are then fed sequentially into a pose estimator, following the approach described in [165]. The pose estimator comprises convolutional and fully connected layers, ultimately generating the final estimated 3D rotation and translation.

#### 4.4.1.1 Pose Parametrization

In our approach, inspired by recent advancements in the field [98, 165, 195], we adopt a continuous 6D representation for rotation, specifically an allocentric representation. For translation, we utilize a scale-invariant representation [29, 98, 165].

Given a set of polarized images $I_{\varphi_{pol}}$ and a collection of detected objects of interest $O = \{O_i | i = 1, \ldots, N\}$, together with their bounding box information $B = \{B_i | B_i = \{b_{wi}, b_{hi}, b_{xi}, b_{yi}\}, i = 1, \ldots, N\}$, our goal is to predict the 6D pose $P = [R|t]$ for each object relative to the camera. These estimates are derived from the cropped regions corresponding to each object, considering their respective 3D CAD models denoted as $M = \{M_i | i = 1, \ldots, N\}$.

The continuous 6D rotation representation, $R_{6\mathrm{d}}$, is derived from the first two columns of a standard rotation matrix $R$ [195]. By zooming in on the target object's ROI, our network focuses on the most relevant information for pose estimation. The scale-invariant translation is estimated based on the relative differences between the projected centroids of the objects and the center locations of their detected bounding boxes with respect to their size. This method offers a robust way to estimate the object's position, considering the variations in the object's size and placement within the scene.

**Rotation Parametrization** The choice of parametrization for tho 6D pose is a critical factor with significant implications for the learning process of a neural network. In the context of parametrizing the 3D rotation matrix $R$, as advocated by [195], it is well-recognized that many existing methodologies face a challenge in handling a discontinuity within Euclidean space, especially when constrained to four or fewer dimensions. This challenge arises in commonly employed parametrization schemes, including unit quaternions [174], log quaternions [117], and Euler angles. To mitigate the issues associated with these discontinuous training signals, we represent rotations using a continuous 6D formulation, denoted as $R_{6d}$, following the works by Wang et al. [165] and Di et al. [29]. This representation is defined by the first two columns of the original 3 × 3 rotation matrix.

The transformation from the 6D representation $R_{6d} = [r_1^{6d} | r_2^{6d}]$ to its original matrix representation $R = [r1|r2|r3]$ is as follows:

$$\begin{cases} r_1 = N(r_1^{6d}) \\ r_1 = N(r1 \times r_2^{6d}) \\ r_2 = r3 \times r1 \end{cases} , \tag{4.12}$$

where $r_i^{6d}$ and $r_i$ denote column vectors in $R_{6d}$ and $R$, respectively, and $N(\cdot)$ signifies the normalization operation.

Given that our neural network exclusively observes each object's cropped and zoomed-in ROI, the representation's notable viewpoint-independence characteristic becomes particularly advantageous. As a result, we undertake a further transformation of the continuous 6D representations, $R_{6d}$, shifting them from an egocentric to an allocentric perspective. This transformation is made feasible by considering factors for translation and camera intrinsics [89].

**Translation Parametrization** In recognition of the constrained global information available within the zoomed-in ROI, we choose to parameterize object translation regarding relative differences. Specifically, we express the translation as the disparity between projected object centroids and the bounding box center location, with respect to the bounding box size, as described in prior works [29, 98, 165]. The resulting translation vector is thus denoted as $t = [\delta_x, \delta_y, \delta_z]^T$, where:

$$\begin{cases} \delta_x = (o_x - b_x)/b_w \\ \delta_y = (o_y - b_y)/b_h \\ \delta_z = t_z/r \end{cases} . \tag{4.13}$$

Here, $(o_x, o_y)$ and $(b_x, b_y)$ represent the coordinates of the projected object centroids and bounding box center, respectively. Additionally, the bounding box size, denoted as $(b_w, b_h)$, plays a crucial role in computing the zoomed-in ratio, denoted as $r = s_{out}/s_{in}$, where $s_{in} = \max(b_w, b_h)$, and $s_{out}$ represents the size of the output. Note that we can recover both the rotation matrix and translation vector, provided we possess knowledge of the camera intrinsics denoted as $K$ [89, 98].

The retrieval of the conventional translation vector $T = [t_x, t_y, t_z]^T$ is achieved through the utilization of known camera intrinsics, as follows:

$$\begin{cases} t_x = \frac{(\delta_x b_w + b_x - c_x) t_z}{f_x} \\ t_y = \frac{(\delta_y b_h + b_y - c_y) t_z}{f_y} \\ t_z = r \delta_z \end{cases}, \tag{4.14}$$

where $(c_x, c_y)$ and $(f_x, f_y)$ denote the principal point and the focal length of the camera, respectively.

#### 4.4.1.2 Object Normal Map

The surface normal map, which contains the surface orientation at each discrete pixel coordinate, serves as the encoding of an object's shape. Drawing inspiration from prior works in SfP [5], we adopt a data-driven approach to recover the surface normal map. In contrast to concatenating the input physical normals with the polarized images, as proposed by Ba et al. [5], we choose to encode them separately using two ResNet encoders. Subsequently, the decoder is trained to generate the object's shape, represented by the surface normal map. Note that the estimated normals are L2-normalized to unit length. As demonstrated later in Tab. 4.3, leveraging the provided physical normals as shape priors yields high-quality normal map predictions, resulting in a notable performance enhancement for the pose estimator.

#### 4.4.1.3 Dense Correspondence Map

The Normalized Object Coordinate Space (NOCS) represents normalized 3D object coordinates while considering their associated poses. This representation explicitly establishes correspondences between 3D object coordinates and their respective projections onto 2D pixel locations. As exemplified in the work of Wang et al. [165], this representation has been demonstrated to enhance the accuracy of consecutive differentiable pose estimators when compared to traditional methods such as RANSAC/P*n*P.

### 4.4.2 Learning Objectives

The comprehensive objective encompasses both the learning of geometrical features and pose optimization, and it is formulated as $\mathscr{L} = \mathscr{L}_{\text{pose}} + \mathscr{L}_{\text{geo}}$, with:

$$\mathscr{L}_{pose} = \mathscr{L}_R + \mathscr{L}_{center} + \mathscr{L}_z \tag{4.15}$$

$$\mathscr{L}_{geo} = \mathscr{L}_{mask} + \mathscr{L}_{normals} + \mathscr{L}_{xyz}. \tag{4.16}$$

Regarding the pose loss $\mathscr{L}_{\text{pose}}$, we separate the optimization procedure into two components: one for the relative displacement with respect to the bounding box, denoted as $(\delta_x, \delta_y)$, and another for the relative depth awareness, represented as $\delta_z$. The optimization of rotation is facilitated by employing a Point-Matching loss [97]. For symmetric objects, the loss term is determined by selecting the smallest loss value from among all possible ground truth rotations.

To be specific, we utilize distinct loss terms for the given ground truth rotation $R$ and the translational components $(\delta_x, \delta_y)$ and $\delta_z$, which can be expressed as follows:

$$\begin{cases} \mathscr{L}_R & = \underset{x \in \mathscr{M}}{\operatorname{avg}} \|Rx - \hat{R}x\|_1 \\ \mathscr{L}_{center} & = \|(\delta_x - \hat{\delta}_x, \delta_y - \hat{\delta}_y)\|_1 \\ \mathscr{L}_z & = \|\delta_z - \hat{\delta}_z\|_1 \end{cases}, \tag{4.17}$$

where the notation $\hat{\bullet}$ signifies predictions. In the case of symmetric objects, the rotation loss is determined by selecting the smallest loss value from the set of all feasible ground-truth rotations considering the symmetry.

For the learning of intermediate geometrical features, we leverage $L1$ losses in the context of the object mask $\hat{M}$ and dense correspondences map NOCS $\hat{M}_{xyz}$ learning. Additionally, we employ a cosine similarity loss for the estimation of surface normals $\hat{n}$:

$$\begin{cases} \mathscr{L}_{mask} & = \|M - \hat{M}\|_1 \\ \mathscr{L}_{xyz} & = M \odot \|M_{xyz} - \hat{M}_{xyz}\|_1 \\ \mathscr{L}_{normal} & = 1 - \langle n, \hat{n} \rangle \end{cases} \tag{4.18}$$

where $\odot$ indicates the Hadamard product of element-wise multiplication, and $\langle \bullet, \bullet \rangle$ denotes the dot product.

## 4.5 Experimental Results

The primary motivation behind our proposed pipeline is to demonstrate the advantages of incorporating pixelwise physical priors derived from polarized light (RGBP) in achieving accurate and robust 6D pose estimation, particularly for objects that present photometric challenges, where traditional RGB-only and RGB-D methods often fall short. To achieve this objective, we train and test **PPP-Net** using different modalities on objects characterized by different levels of photometric complexity. Specifically, we consider a simple plastic *cup* as well as photometrically demanding objects, such as reflective and textureless stainless steel cutlery (*fork* and *knife*). As we elaborate in subsequent sections, our investigations reveal that integrating polarimetric information results in a substantial performance improvement, particularly for objects presenting significant photometric challenges.

### 4.5.1 Polarimetric Data Acquisition

This work introduces an instance-level benchmark dataset for 6D pose estimation, leveraging physical cues derived from polarimetric images tailored explicitly for objects with significant photometric challenges. The selection of objects within the dataset includes a wide range of photometric complexities, ranging from matte to highly reflective and even transparent surfaces, similar to those in PhoCal [167]. This instance-level benchmark dataset is a subset of the acquired data, as presented in the previous chapter. For this benchmark, we specifically select the following objects: *cup*, *teapot*, *can*, *fork*, *knife*, and *bottle*. These objects are chosen due to their progressively increasing photometric complexity, as visually depicted in Fig. 4.4. Note that the latter three models lack texture on their surfaces, necessitating the application of a temporary, vanishing 3D scanning spray to create a temporary opaque surface for scanning. Tab. 4.2 provides an overview of various dataset characteristics and ours for comparison.

Fig. 4.5 provides an overview of our scene settings and showcases the quality of our pose annotations. The superimposed 3D meshes of the objects illustrate the high level of accuracy achieved in our annotations. Our dataset encompasses various backgrounds, lighting conditions, and object settings, making it suitable for comprehensive evaluation. Notably, our

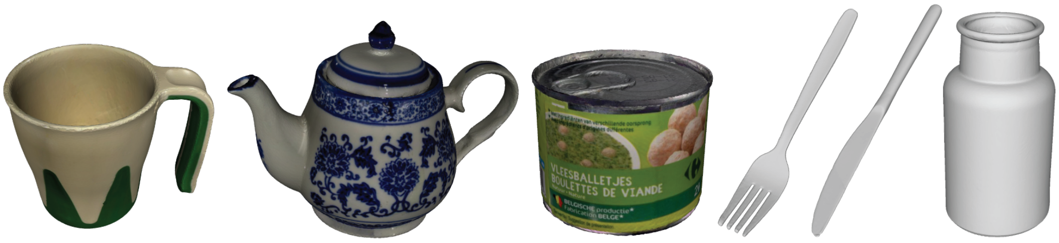


Fig. 4.4 **3D Models:** Objects with varying degrees of photometric complexity, arranged from left to right. Three of these objects lack texture due to either reflection (cutlery) or transparency (bottle).

Tab. 4.2 **Dataset Comparison.**

| Dataset | RGB | Depth | Polarisation | Robotic GT | Occlusion | Symmetry | Transparent | Reflective | Sequences |
|---|---|---|---|---|---|---|---|---|---|
| YCB-V [55] | ✓ | ✓ | | | ✓ | ✓ | | | 92 |
| T-LESS [23] | ✓ | ✓ | | | ✓ | ✓ | | | 20 |
| Linemod [21] | ✓ | ✓ | | | ✓ | ✓ | | | 15 |
| Ours | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 20 |

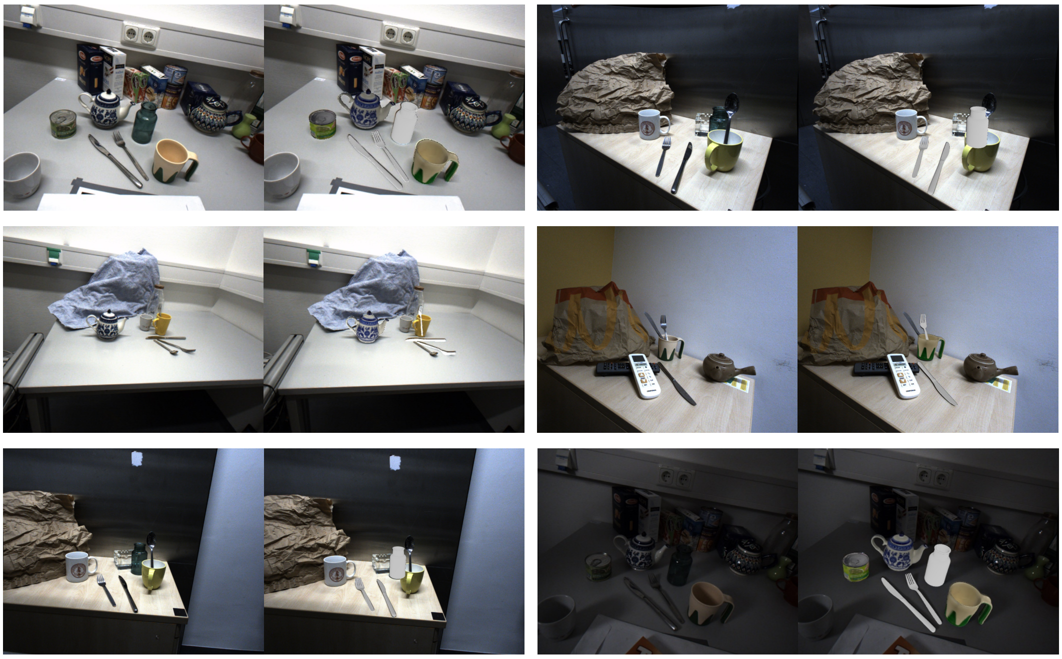


Fig. 4.5 **Dataset and Annotation Example:** The figure displays a single polarimetric image alongside the rendered 3D models.

pose annotations maintain accuracy even for challenging objects with reflective or transparent properties. The precision of these annotations is ensured through a dedicated process [167] detailed in the previous chapter, involving multiple controlled interactions with the object's surface using a calibrated tool tip attached to a robotic arm, followed by fine alignment using ICP with the pre-scanned 3D mesh of the object. While 6D pose annotations are available for all objects in the scene, we focus here on the subset introduced in Fig. 4.4, representing a broad spectrum of photometric complexities.

### 4.5.2 Experiments Setup

Initially, we fine-tune an off-the-shelf object detector, Mask R-CNN [55], directly on the polarized images $I_0$. This fine-tuning is essential to generate relevant object crops required for our approach and the RGB-only benchmark. We adopt a training/testing split strategy similar to what is commonly used for public datasets [12], where $\approx 10\%$ of the images are allocated for training, and the remaining 90% are reserved for testing. Our network is trained end-to-end using the Adam optimizer [85] for 200 epochs. We initiate training with an initial learning rate of $1 \times 10^{-4}$, which is then halved every 50 epochs. Note that due to differences in the field of view and the camera setup (with the depth sensor located beneath the polarization camera on our custom rig), the split between training and testing data for the RGB-D benchmark varies from the RGB-only training/testing split.

To evaluate our novel 6D pose estimation approach, we assess the pose estimation accuracy for each object using commonly adopted metrics, including the average distance (ADD) and its counterpart for symmetrical objects (ADD-S) [60], against different benchmarks. For surface normal estimation, we compute both mean and median errors (in degrees), along with the percentage of pixels where the estimated normals deviate by less than 11.25°, 22.5°, and 30° from the ground truth.

### 4.5.3 Experiments Evaluation

We conduct experiments to analyse the impact of different input modalities on the accuracy and robustness of 6D object pose estimation. The quantitative results are presented in Tab. 4.3, while Fig. 4.6 illustrates the qualitative improvements in the Normalized Object Coordinate Space (NOCS) representation. To assess the direct impact of polarimetric imaging on accurate and robust object pose estimation, focusing on photometrically challenging objects, we begin by establishing an RGB-only baseline. This baseline is set up by omitting the contributions of our **PPP-Net** and using unpolarized RGB images as the input. These RGB images are obtained by averaging over polarimetric images at complementary angles. The comparative analysis, as detailed in the first two rows of Table 4.3 for each object (comparing RGB with Polar RGB), reveal that the addition of polarimetric data significantly enhances the pose estimation accuracy and robustness, particularly for objects with photometrically challenging characteristics. For instance, the polarisation modality exhibits more substantial accuracy gains for the object *fork*, which presents more photometric challenges compared to the object *cup*. These findings underscore the valuable contribution of polarimetric imaging in improving the precision and robustness of pose estimation, especially for objects that pose difficulties to traditional RGB-based methods due to their surface properties. Incorporating polarimetric data is crucial for more robust and accurate pose estimation in a broader range of real-world scenarios.

The robustness and accuracy of the pose estimation can be further enhanced when the network is guided to extract additional shape information implicitly encoded in the polarization images, as demonstrated in Tab. 4.3 (rows 2nd to 3rd for each object, respectively). However, it is important to acknowledge that the quality of predicted normals in this setting remains somewhat limited. When the network is furnished with physically-induced normals obtained from polarization images as input, it gains access to a plausible priors for directly encoding shape information. Consequently, it yields significantly improved normals predictions, resulting in a substantial enhancement of pose performance, as evident from Tab. 4.3 (rows 3rd to 4th for each object, respectively). The comparison of NOCS predictions presented in Fig. 4.6 underscores that when provided with polarization and direct shape cues, the network establishes a more precise and intricate geometrical representation, aligning with the observed quantitative improvements.

In Figures 4.7 and 4.8, we illustrate the 6D pose by superimposing the image with the corresponding transformed 3D bounding box. To enhance visibility, we cropped the images and focused on the region of interest.

Tab. 4.3 **PPP-Net Input Modalities Evaluation:** Various combinations of input and output modalities are employed during training to investigate their impact on the accuracy of pose estimation (ADD(-S)) across objects with differing photometric complexity. When applicable, we also present metrics for estimated normals.

| Object | Photo. Chall. | Input Modalities | | | Output Variants | | Normal Metrics | | | | | Pose Metric |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | RGB | Polar RGB | Physical N | Normals | NOCS | mean↓ | med.↓ | 11.25° ↑ | 22.5°↑ | 30°↑ | ADD(-S) |
| Cup | | ✓ | | | | ✓ | - | - | - | - | - | 91.1 |
| | | | ✓ | | | ✓ | - | - | - | - | - | 91.3 |
| | | | ✓ | | ✓ | ✓ | 7.3 | 5.5 | 86.2 | 96.1 | 97.9 | 91.3 |
| | | | ✓ | ✓ | ✓ | ✓ | **4.5** | **3.5** | **94.7** | **99.1** | **99.6** | **97.2** |
| Teapot | † | ✓ | | | | ✓ | - | - | - | - | - | 97.8 |
| | | | ✓ | | | ✓ | - | - | - | - | - | 99.5 |
| | | | ✓ | | ✓ | ✓ | 7.9 | 5.4 | 82.5 | 94.5 | 97.1 | 99.2 |
| | | | ✓ | ✓ | ✓ | ✓ | **5.3** | **4.0** | **91.6** | **98.7** | **99.5** | **99.9** |
| Can | † | ✓ | | | | ✓ | - | - | - | - | - | 91.8 |
| | | | ✓ | | | ✓ | - | - | - | - | - | 93.2 |
| | | | ✓ | | ✓ | ✓ | **5.7** | **3.9** | **90.0** | 97.0 | 98.6 | 96.7 |
| | | | ✓ | ✓ | ✓ | ✓ | 6.0 | 4.5 | 89.0 | **97.3** | **98.9** | **98.4** |
| Fork | †† | ✓ | | | | ✓ | - | - | - | - | - | 85.4 |
| | | | ✓ | | | ✓ | - | - | - | - | - | 86.1 |
| | | | ✓ | | ✓ | ✓ | 11.0 | 7.3 | 72.6 | 90.7 | 93.9 | 92.9 |
| | | | ✓ | ✓ | ✓ | ✓ | **6.5** | **4.3** | **87.6** | **95.9** | **97.6** | **95.9** |
| Knife | †† | ✓ | | | | ✓ | - | - | - | - | - | 84.1 |
| | | | ✓ | | | ✓ | - | - | - | - | - | 88.0 |
| | | | ✓ | | ✓ | ✓ | 12.2 | 8.0 | 68.7 | 88.5 | 92.4 | 89.4 |
| | | | ✓ | ✓ | ✓ | ✓ | **6.8** | **5.4** | **88.2** | **97.3** | **98.6** | **96.4** |
| Bottle | ††† | ✓ | | | | ✓ | - | - | - | - | - | 90.5 |
| | | | ✓ | | | ✓ | - | - | - | - | - | 93.5 |
| | | | ✓ | | ✓ | ✓ | 5.6 | 4.7 | **92.9** | **99.0** | **99.6** | 94.7 |
| | | | ✓ | ✓ | ✓ | ✓ | **5.4** | **4.5** | 92.1 | **99.0** | **99.6** | **97.5** |

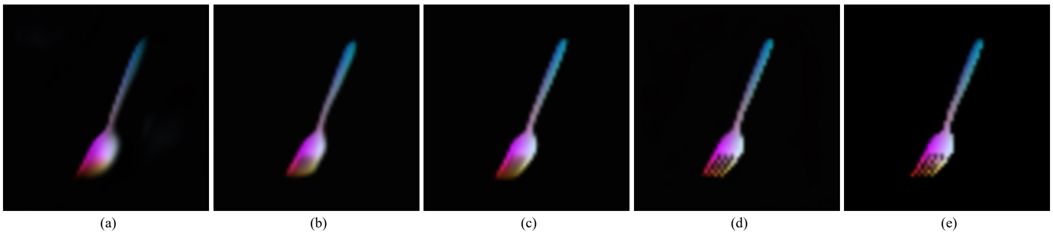


Fig. 4.6 **Visualization of Ablations on NOCS:** The quality of the geometric representations improves when incorporating physical priors. The NOCS prediction follows the same order as the ablation experiments in Tab. 4.3: (a) unpolarized RGB input with NOCS output; (b) polarization input with NOCS output; (c) polarization input with NOCS and normals output; (d) **ours:** full model with polarization and physical priors input, NOCS, and normals output; (e) ground truth NOCS.

## 4.5.4 Comparison with Established Benchmarks

The experiments involving input modalities have already demonstrated the robust capabilities of polarimetric imaging inputs for **PPP-Net**. It has proven to successfully learn reliable 6D pose prediction with high accuracy, especially for photometrically challenging objects. While the depth map from an RGB-D sensor also offers geometric information valuable for 6D object pose estimation, we compare our method against FFB6D [57]. FFB6D employs a unique design

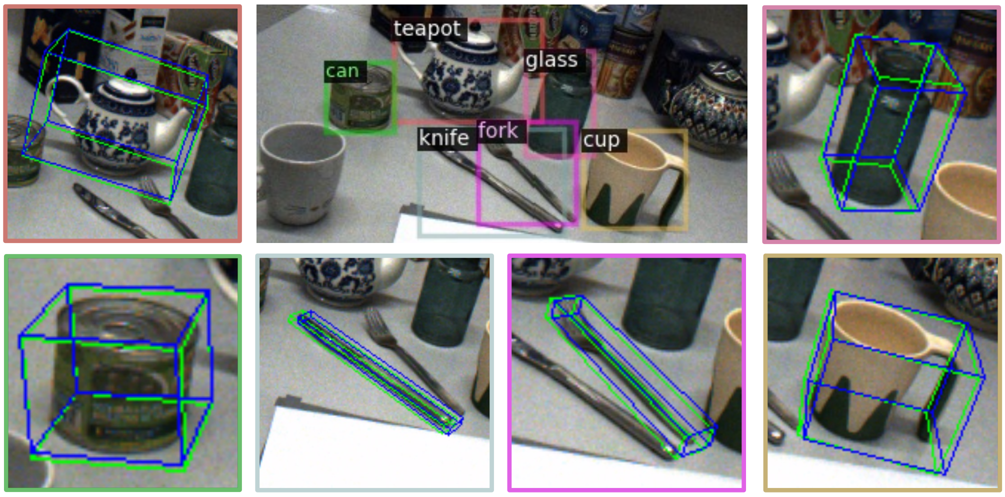


**Fig. 4.7** **Qualitative Results for a Scene:** Input images with 2D detections are displayed. The predicted 6D poses are represented by *blue* bounding boxes, while the ground truth (GT) poses are indicated by *green* bounding boxes.

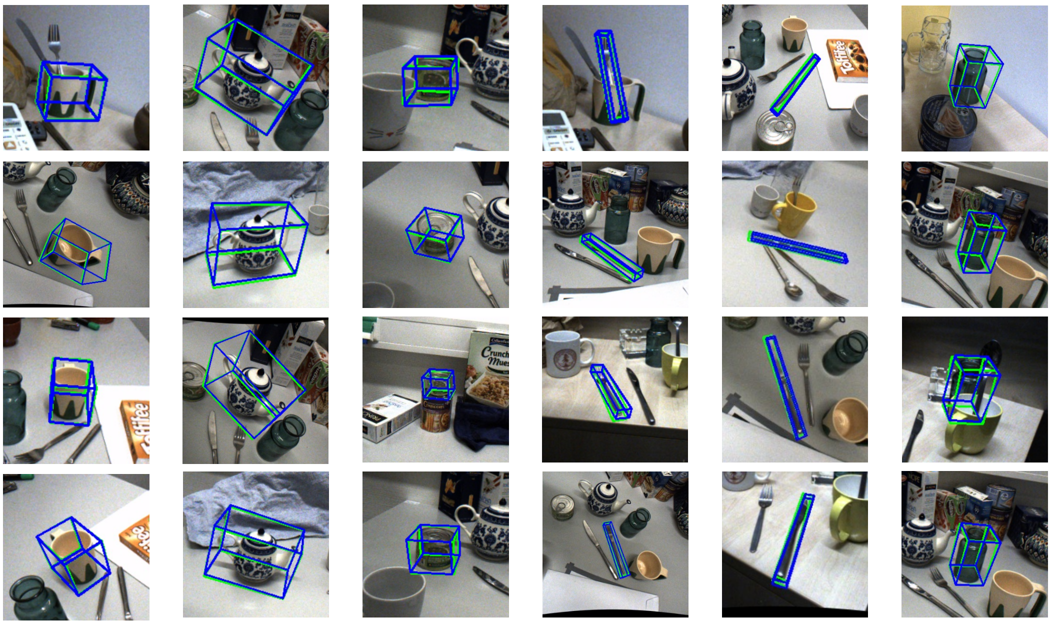


**Fig. 4.8** **Qualitative Results for Different Scenes:** Predicted and ground truth 6D poses are depicted using *blue* and *green* bounding boxes, respectively.

that learns to integrate appearance and depth information, considering both local and global aspects from the two distinct modalities.

We train FFB6D on our dataset for each object individually and report the best ADD(-S) metric for all objects in Tab. 4.4. The degree of the photometric challenge posed by each object is summarized in Tab. 4.4 and further elaborated based on their properties (refer to Fig. 4.4

Tab. 4.4 **Benchmark Comparisons:** We conduct a comparative analysis of our method against recent RGB-D (FFB6D [57]) and RGB-only (GDR-Net [165]) approaches using a diverse set of objects. These objects exhibit varying levels of photometric challenges (†) and variations in depth map quality (ranging from good: + to low: −), which serve as input for FFB6D. RGB-D and RGB-only comparisons are trained and tested on different splits due to the distinct field of view of the depth camera. We evaluate the performance using the Average Recall of ADD(-S).

| Object | Photo. Chall. | Properties | | | | | Depth Quality | RGB-D Split | | RGB Split | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Reflective | Metallic | Textureless | Transparent | Symmetric | | FFB6D | **Ours** | GDR | **Ours** |
| Cup | | | | | | | (+) | **99.4** | 98.1 | 96.7 | **97.2** |
| Teapot | † | (*) | | | | | ++ | 86.8 | **94.2** | 99.0 | **99.9** |
| Can | † | * | * | | | | - | 80.4 | **99.7** | 96.5 | **98.4** |
| Fork | †† | * | * | * | | | – | 37.0 | **72.4** | 86.6 | **95.9** |
| Knife | †† | * | * | * | | | – | 36.7 | **87.2** | 92.6 | **96.4** |
| Bottle | ††† | * | | * | * | * | None | 61.5 | **93.6** | 94.4 | **97.5** |
| Mean | | | | | | | | 67.0 | **90.9** | 94.3 | **97.6** |

for comparison). The objects are categorized into three classes based on the quality of depth maps captured by the depth sensor (also refer to Fig. 4.1). Our observations reveal that objects with good depth maps and minimal photometric challenges tend to achieve high ADD values when processed by FFB6D [57]. However, as the level of photometric complexity increases (accompanied by worse depth map quality), we notice a corresponding decrease in ADD for challenging objects.

Interestingly, the transparent *bottle* object presents an exception to this trend. Despite having an entirely invalid depth map (compare Fig. 4.1), FFB6D still achieves high ADD. We speculate that the network learns to disregard the depth map input early in training. **PPP-Net** demonstrates comparable performance to FFB6D for objects with low complexity and surpasses this strong benchmark for objects with significant photometric complexity. Our method remains robust in scenarios with noisy or inaccurate depth maps, harnessing orthogonal surface information extracted from RGBP data.

Given that **PPP-Net** significantly benefits from including physical priors obtained through polarization imaging, we undertake a comprehensive investigation to assess the extent of improvement in estimated poses, particularly for objects with significant photometric challenges. We conduct this analysis by comparing the results against the monocular RGB-only method, GDR-Net [165]. Our observations reveal that utilizing polarimetric information slightly improves pose estimation accuracy for objects that are not particularly photometrically challenging. We can achieve superior performance for items exhibiting inconsistent photometric properties due to factors like reflection or transparency. The degree of accuracy improvement offered by **PPP-Net** over GDR-Net increases proportional to the photometric complexity of the objects, highlighting the invaluable role played by our physical priors in enhancing the understanding of an object's geometry (as presented in Tab. 4.4).

#### 4.5.4.1 Additional Experiments and Ablation Studies

**Ablations on Network Architecture** Tab. 4.5 reveals that the simple concatenation of geometric priors and RGBP images as direct input to the network (as presented for SfP in [5]) results

Tab. 4.5 **Fusion Ablation:** Naive concatenation versus our proposed fusion strategy of RGB and physical priors in **PPP-Net**.

| Object | Fusion | Input Modalities | | Output Variants | | Normal Metrics | | | | | Pose Metric |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Polar RGB | Physical N | Normals | NOCS | mean↓ | med.↓ | 11.25° ↑ | 22.5°↑ | 30°↑ | ADD |
| Cup | `concat` | ✓ | ✓ | ✓ | ✓ | 6.0 | 4.9 | 91.1 | 98.1 | 99.1 | 93.6 |
| Cup | `ours` | ✓ | ✓ | ✓ | ✓ | **4.5** | **3.5** | **94.7** | **99.1** | **99.6** | **97.2** |
| Teapot | `concat` | ✓ | ✓ | ✓ | ✓ | 7.4 | 5.7 | 83.4 | 96.3 | 98.4 | 97.3 |
| Teapot | `ours` | ✓ | ✓ | ✓ | ✓ | **5.3** | **4.0** | **91.6** | **98.7** | **99.5** | **99.9** |
| Can | `concat` | ✓ | ✓ | ✓ | ✓ | 8.5 | 6.4 | 81.8 | 95.1 | 97.5 | 92.2 |
| Can | `ours` | ✓ | ✓ | ✓ | ✓ | **6.0** | **4.5** | **89.0** | **97.3** | **98.9** | **98.4** |
| Fork | `concat` | ✓ | ✓ | ✓ | ✓ | 10.7 | 7.8 | 70.0 | 91.8 | 95.0 | 87.6 |
| Fork | `ours` | ✓ | ✓ | ✓ | ✓ | **6.5** | **4.3** | **87.6** | **95.9** | **97.6** | **95.9** |
| Knife | `concat` | ✓ | ✓ | ✓ | ✓ | 10.8 | 8.5 | 67.1 | 92.8 | 96.2 | 86.1 |
| Knife | `ours` | ✓ | ✓ | ✓ | ✓ | **6.8** | **5.4** | **88.2** | **97.3** | **98.6** | **96.4** |
| Bottle | `concat` | ✓ | ✓ | ✓ | ✓ | 7.6 | 6.0 | 86.5 | 94.8 | 96.4 | 93.1 |
| Bottle | `ours` | ✓ | ✓ | ✓ | ✓ | **5.4** | **4.5** | **92.1** | **99.0** | **99.6** | **97.5** |

in inferior quality of normal predictions and offers limited improvement in pose estimation performance when compared to our approach (`concat` vs. `ours` in Tab. 4.5). These observations are consistent across all objects, with more photometrically challenging objects demonstrating a relatively more considerable improvement. These findings underscore the significance of our design choices in **PPP-Net**, particularly incorporating a dedicated encoder for physics-based geometric priors. This architectural decision positively impacts the accuracy of 6D object pose estimation.

We advocate for a careful integration strategy for incorporating physical priors into the established principles of 6D object pose estimation, as demonstrated within our novel hybrid encoder. We intentionally adopt a straightforward and general architecture for **PPP-Net** to facilitate meaningful comparisons with state-of-the-art methods. The results highlight that even such simplified encoders can achieve notable accuracy in 6D pose prediction when using physical priors derived from polarization as inputs.

**Ablations on Output Modality** 6D pose estimation heavily relies on accurate correspondence prediction through NOCS regression, as evidenced in the ablation analysis presented in Tab. 4.6. The ADD metric experiences a significant decrease when the model lacks NOCS output (w/o NOCS) prior to Patch-PnP. In this case, only shape information would be utilized for pose prediction. Still, the previous experiments demonstrate that the explicit prediction of object-centric shape information, such as the normals map, benefits 6D pose estimation. This auxiliary prediction guides the network more effectively in extracting physical shape priors from the input data, ultimately improving pose estimation accuracy.

Tab. 4.6 **PPP-Net Output Ablation:** Comparison of PPP-Net outputs with and without NOCS output.

| Object | Pose Metric (ADD) | |
|---|---|---|
| Teapot | w/ **99.9** | w/o 72.7 |
| Fork | w/ **95.9** | w/o 79.3 |

Tab. 4.7 **Bounding Box Ablations: We evaluate the performance using the Average Recall of ADD(-S)**

| Configuration | Cup | Teapot | Can | Fork | Knife | Bottle |
|---|---|---|---|---|---|---|
| Train with GT BBox/Test with pred BBox | 97.2 | 99.9 | 98.4 | 95.9 | 96.4 | 97.5 |
| Train/Test with GT BBox | 99.0 | 99.9 | 99.0 | 96.1 | 97.6 | 97.5 |

Tab. 4.8 **Refractive Index Ablation: We evaluate the performance using the Average Recall of ADD(-S)**

| Object | Cup | Teapot | Can | Fork | Knife | Bottle |
|---|---|---|---|---|---|---|
| Actual correct Refractive Index | 1.50 | 1.54 | 1.35 | 2.75 | 2.75 | 1.52 |
| Train/Test with correct index | 97.2 | 99.9 | 98.4 | 95.9 | 96.4 | 97.5 |
| Train with correct index,<br>test with incorrect (1.5) | 97.2 | 99.9 | 98.3 | 95.8 | 96.2 | 97.5 |
| Train/Test with incorrect index (1.5) | 97.2 | 99.9 | 98.0 | 93.5 | 90.1 | 97.5 |

**Ablation on Detector** We train an object detector using Faster R-CNN without making any additional modifications to the polarimetric inputs. The performance of this object detector is not significantly affected by the photometric challenges posed by the objects. This is evident from the comparable results obtained in Tab. 4.7 when we train or test **PPP-Net** using the ground truth bounding boxes or the predicted bounding boxes.

**Ablation on Refractive Index** One limitation of our model is its reliance on prior knowledge of the refractive index of materials present in the scene. To analyze the impact of incorrect refractive indices, we present pose accuracy results when the model is trained and tested with minor deviations (1.54 vs. 1.5) and large deviations (2.75 vs. 1.5) from the correct index values, as shown in Tab. 4.8. The results in the second row emphasize that our model maintains good performance even when provided with incorrect refractive indices during inference. This suggests that the model is robust enough to extract relevant features. However, when we train and test the model with significantly different indices, a slight decrease in Average Distance of Discrepancy (ADD) is observed, particularly for the *fork* and *knife* objects.

**Runtime Analysis** On a desktop PC equipped with an Intel i7 4.20GHz CPU and an NVIDIA 2080 GPU, our network processes a single object from a 512 × 612 pixel image in approximately 64 milliseconds. This time breakdown includes roughly 40 milliseconds for object detection and 13 milliseconds for calculating the physical priors. Please note that these measurements are based on our non-optimized implementation.

## 4.6 Conclusion

We introduced **PPP-Net**, a pioneering 6D object pose estimation framework that harnesses geometric insights from polarization images via physical cues. Our approach surpasses the current state-of-the-art RGB-D and RGB-only methods, excelling in scenarios involving photometrically challenging objects, while delivering competitive performance for regular objects. Extensive ablation studies underscore the crucial role of complementary polarization information in achieving precise and robust pose estimations, particularly for objects with reflective or transparent surfaces.

Our results underscore the substantial enhancements that physical priors can bring to 6D pose estimation for photometrically challenging objects. RGB-only methods, which lack geometric information, falter in scenarios featuring objects with minimal texture. Methods that attempt to leverage geometric priors from RGB-D sources [57] often struggle to reliably recover the 6D pose for such objects, given the typically degraded and corrupted nature of depth maps. In contrast, our **PPP-Net**, as RGBP 6D object pose estimation approach, effectively achieves precise and robust pose estimations, even for exceptionally challenging objects, by extracting geometric insights from physical priors. Qualitative results demonstrating this capability can be found in Figs. 4.3 and 4.7. Moreover, the advantages of using RGBP extend to the sensor technology itself. RGB-D cameras often necessitate energy-intensive active illumination and extrinsic calibration, complicating integration and introducing additional uncertainty to the final RGB-D image. As the polarization filter is seamlessly integrated into the same sensor as the Bayer filter, both modalities are intrinsically calibrated, allowing for passive image acquisition. This opens the door to sensor integration on energy-efficient and mobile devices.

# 5 Polarimetric Self-Supervision

> *“fiat lux*
> *et facta est lux”*
>
> – "**let there be light, and there was light**"
> 1. Book of Moses

## 5.1 Introduction

Building upon the insights from the previous chapter, which established the integration of polarimetric information for 6D object pose estimation, particularly with photometrically challenging objects, we now address a significant limitation: the requirement for extensive annotated real data in our supervised approach. While self-supervised RGB-D methods exist, they also suffer from noisy and corrupt depth measurements due to sensor noise, as discussed in previous chapters. On the other hand, as evidenced by the experiments for the proposed supervised method, polarimetric images provide robust surface and shape information.

We, therefore, delve deeper into the physics of polarization and propose an invertible physical model, which, together with novel network and loss components, enables self-supervision for 6D object pose estimation. Hence, we avoid the need for annotated real data and address the issue in current SOTA self-supervised RGB-D methods of corrupt depth information due to sensor artifacts, especially for photometrically challenging objects, by utilizing more robust shape priors encoded in polarimetric images.

### 5.1.1 Motivation

Recent methods in 6D object pose estimation integrate geometric information either directly as input [57] or use it for self-supervision [163]. While reliable geometric cues from depth sensors can enhance pose estimation performance, noisy or unreliable depth information can negatively affect the learning process. As discussed in the previous chapter, integrating the geometric information of polarized light enhances and robustifies 6D object pose estimation. The presented method learns features from both the estimated normal from polarization and their polarization characteristics in a supervised manner. Accuracy and robustness are remarkable, particularly concerning objects lacking distinct textures, exhibiting reflectivity, or possessing translucency. The results surpass the performance of state-of-the-art RGB-only [165] and RGB-D [57] methods. However, it is essential to note that an extensive training dataset with accurately annotated ground-truth data is requisite for these advancements. Obtaining such a dataset, particularly with high accuracy, may present practical challenges [167]. Leveraging the robust polarization information for self-supervised 6D object pose estimation would circumvent the necessity for annotated real data and avoid the issues of depth sensors.

### 5.1.2 Contributions

In our **Self-Supervised Polarimetric Pose Prediction** framework, short **S**$^2$**P**$^3$, we delve into how neural networks can effectively utilize geometric shape priors derived from polarized light for the task of 6D object pose estimation and how our approach can eliminate the necessity for annotated real data. We employ our supervised polarimetric 6D object pose estimation method from before as a teacher network, initially pre-training it on synthetically rendered polarimetric images only. Subsequently, its predictions on real data, although noisy, are utilized as weak labels to guide a student network.

For self-supervision, we employ a differentiable renderer, enabling dense geometric cue integration. We also introduce an invertible formulation of the physical polarization model, allowing for analytical computation of pixel-wise polarization characteristics from geometric normal representation. This inversion, post differentiable rendering of normals with the student network's predicted 6D pose, closes the self-supervision loop, facilitating direct comparison with input polarization data as shown in Fig. 5.1.

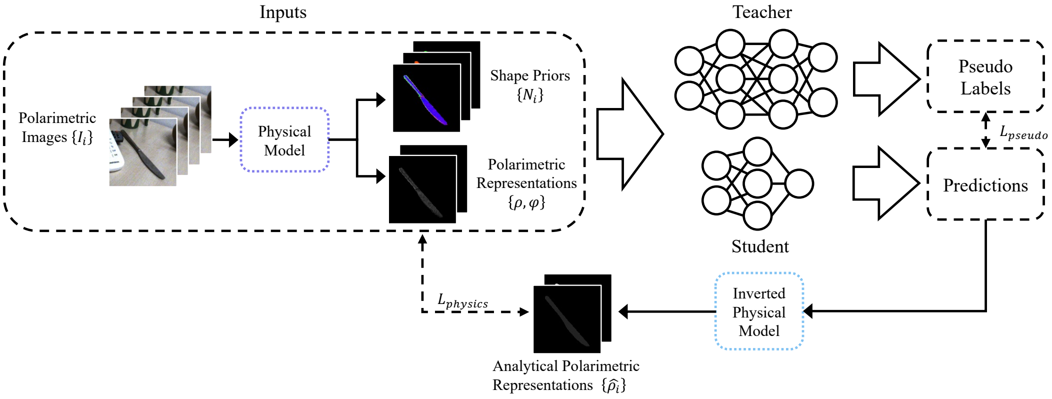

Inputs
Teacher
Polarimetric Images $\{I_i\}$
Physical Model
Shape Priors $\{N_i\}$
Polarimetric Representations $\{\rho, \varphi\}$
Pseudo Labels
$L_{pseudo}$
Predictions
Student
$L_{physics}$
Inverted Physical Model
Analytical Polarimetric Representations $\{\hat{\rho}_i\}$

Fig. 5.1 **S**$^2$**P**$^3$ **Pipeline Overview:** In our proposed teacher-student training paradigm, we employ four polarization images acquired at different polarization filter angles alongside polarimetric and geometrical representations derived from the analytical physical model. These are inputs to the teacher and student networks in a multi-modal fashion. The student network's optimization objective extends beyond the pseudo labels generated by the teacher network, denoted as $L_{pseudo}$. It also encompasses $L_{physics}$, which seeks to minimize the disparity between the polarimetric representations $\rho$ extracted from the input images post application of the analytical physical model and the corresponding $\hat{\rho}$ obtained through the inverted physical model. This derivation relies on the rendered surface normal given the estimated object pose produced by the student network.

The primary contributions in this chapter are in summary:

## Contributions

1. Analytically derived **shape priors** and **polarimetric characteristics** from a physical model are provided to the network to encode neural shape representations.
2. A **teacher network** (pre-trained on rendered synthetic data only) provides **weak pseudo labels** on real images (6D pose and geometric representations) for a student network in a **knowledge distillation scheme**.
3. A **novel self-supervised loss** formulation through differentiable rendering and an **invertible physical constraint** enables training **without annotated real data** by coupling the input polarimetric information with the predicted 6D object pose.

## 5.2 Related Work

Building on the related work in 6D object pose estimation discussed in the previous chapter, we now focus on the most relevant advancements in self-supervised 6D pose estimation. This approach is beneficial for circumventing the challenges of acquiring accurately labeled data. An essential technique in this domain is differentiable rendering, which generates synthetic images based on predicted poses for comparison against actual input images [147]. The Self6D method [164] exemplifies this approach, training a neural network on synthetic RGB data and then fine-tuning it on real RGB-D data without pose annotations in a self-supervised manner. Central to Self6D's success is the use of depth data to synchronize visual and geometric cues. Enhancing upon Self6D, the Self6D++ framework [163] introduces significant improvements. It shifts from a one-stage pose regression model to a more advanced two-stage GDR-net backbone [165] and employs a teacher-student scheme and a pose refinement layer on top of the teacher network. This adaptation notably boosts accuracy, especially in scenarios with occlusions.

Despite these advancements, challenges persist, particularly in handling photometrically challenging objects due to depth artifacts from the active depth sensors. As discussed earlier, incorporating polarimetric information significantly boosts robustness and accuracy in supervised scenarios. Building on the advancements in self-supervised learning and differentiable renderers in end-to-end learning pipelines, as demonstrated in Self6D++ [163], we apply these concepts to the multi-modal imaging domain of polarization. Unlike Self6D++, which uses a renderer to produce a depth map for comparison against a potentially noisy depth map from an active sensor, we delve into the physical properties of light. We integrate encoded shape priors into a self-supervised scheme through a differentiable analytical derivation of physical properties from surface normal information. This approach allows us to leverage the unique attributes of polarized light to improve the accuracy and robustness of 6D object pose estimation, particularly for photometrically complex objects that pose challenges to depth sensors.

## 5.3 Invertible Polarimetric Physical Model

We introduce the inversion of the physical model of polarimetric imaging. Assuming we have a normal map of an object, which can be generated by a differentiable renderer using the 3D model and an estimated 6D pose, we formulate an invertible solution to analytically derive the polarimetric representation from this normal map. It translates the information from the object's pose, parameterized as a 6D transformation, through the differentiable renderer into a geometric representation. This geometric information is then converted into encoded physical properties of light reflections. These properties can subsequently be compared against the original input data within our self-supervised learning framework. This process effectively 'closes the loop' for our network and enables end-to-end training with self-supervision from passively observed properties of light, namely polarization.

Given:

$$\overrightarrow{n} = \begin{bmatrix} n_x \\ n_y \\ n_z \end{bmatrix} = \begin{bmatrix} \cos\alpha\sin\theta \\ \cos\alpha\cos\theta \\ \cos\theta \end{bmatrix}, \tag{5.1}$$

we can derive the polarization parameters analytically, specifically the Angle of Polarization (AoP) denoted as $\phi$, and the Degree of Polarization (DoP) represented as $\rho$.

For the AoP $\phi$, the first step involves solving for the azimuth angle $\alpha$ using the equation:

$$\alpha = \arctan\frac{n_y}{n_x}. \tag{5.2}$$

The AoP is subsequently correlated with a set of potential solutions under the orthographic assumption:

$$\phi \in \left\{\alpha, \alpha - \pi, \alpha + \frac{\pi}{2}, \alpha - \frac{\pi}{2}\right\} \tag{5.3}$$

The DoP is impacted by the viewing angle $\theta_v$. Under orthographic projection, the viewing angle $\theta_v$ equals the zenith angle $\theta$. However, under perspective projection, we first compute the viewing vector $v$ and then determine the viewing angle $\theta_v$. The viewing vector $v$ is defined as follows:

$$v = -\pi^{-1}(u, v, K) = -\frac{1}{\left\|\left[\frac{u-c_x}{f_x}, \frac{v-c_y}{f_y}, 1\right]\right\|}\begin{bmatrix} \frac{u-c_x}{f_x} \\ \frac{v-c_y}{f_y} \\ 1 \end{bmatrix}, \tag{5.4}$$

where $n$ represents the rendered object surface normal map, and the viewing vector $v$ is defined as: $v = -\pi^{-1}(u, v, K)$ Here, $\pi^{-1}$ serves as the backprojection operation for the pixel $(u, v)$, utilizing the camera intrinsics $K$.

The viewing angle $\theta_v$ represents the angle between the surface normal $n$ and the viewing direction $v$:

$$\cos\theta_v = n \cdot v. \tag{5.5}$$

The analytical DoP, denoted as $\hat{\rho}$, is subsequently derived through formulations for both diffuse and specular reflection scenarios, considering a set of potential $\hat{\rho}_i$:

$$\begin{cases} \hat{\rho}_d = \frac{(\eta - 1/\eta)^2 \sin^2(\theta_v)}{2 + 2\eta^2 - (\eta + 1/\eta)^2 \sin^2(\theta_v) + 4\cos(\theta_v)\sqrt{\eta^2 - \sin^2(\theta_v)}} \\ \\ \hat{\rho}_s = \frac{2\sin^2(\theta_v)\cos(\theta_v)\sqrt{\eta^2 - \sin^2(\theta_v)}}{\eta^2 - \sin^2(\theta_v) - \eta^2\sin^2(\theta_v) + 2\sin^4(\theta_v)} \end{cases} \tag{5.6}$$

where $\eta$ represents a constant defined by the refractive index of the object materials.

The inverted physical model not only enables the optimization of the model but also does so through the utilization of object shape cues. This approach is more robust in photometrically challenging scenarios than conventional active depth sensors.

## 5.4 Physical Conditions for Self-Supervised Polarimetric Pose Prediction

The primary objective of $\mathbf{S}^2\mathbf{P}^3$ is to achieve instance-level 6D object pose prediction without the dependency on annotated real data. To realize this goal, we propose a teacher-student training paradigm that leverages pre-training on synthetic data and incorporates pseudo-labels generated by the teacher network during self-supervision, as illustrated in Fig. 5.1. Through the integration of the novel invertible physical model, $\mathbf{S}^2\mathbf{P}^3$ harnesses the complete spectrum of geometric data encoded within polarimetric images.

### 5.4.1 Network Architecture

$\mathbf{S}^2\mathbf{P}^3$, comprising a teacher network (cf. Fig. 5.2) characterized by a larger capacity and a lightweight student network (cf. Fig. 5.3), is presented in Fig. 5.4 as a schematic overview. Both of these networks undergo pre-training on synthetic data. The teacher network plays a crucial role in providing pseudo labels for real data, thereby guiding the self-supervised learning process of the student network.

The detailed architecture of $\mathbf{S}^2\mathbf{P}^3$ shows essential extensions, modifications, and significant design considerations in contrast to established teacher-student training paradigms within the area of 6D object pose estimation [163]. These aspects are detailed in the subsequent sections and are substantiated by experimental ablations discussed in the experiments section.

As depicted in Fig. 5.4, the input polarization images undergo the same forward physical model employed in PPP-Net in the previous chapter. This process is utilized to derive polarimetric properties, specifically the AoP $\phi$, the DoP $\rho$, and a collection of geometrical priors under varying assumptions regarding the reflection type, which may be diffuse or specular.

#### 5.4.1.1 Teacher Network

Inspired by the architectural framework of PPP-Net from the previous chapter, we introduce our polarimetric network, which features an extended differentiable renderer, designed to act as the teacher network within $\mathbf{S}^2\mathbf{P}^3$ (as illustrated in Fig. 5.2). Within this network, polarimetric intensity inputs and geometrical shape priors are channeled through distinct input heads, subsequently processed by an explicit decoder to yield predictions for various components:

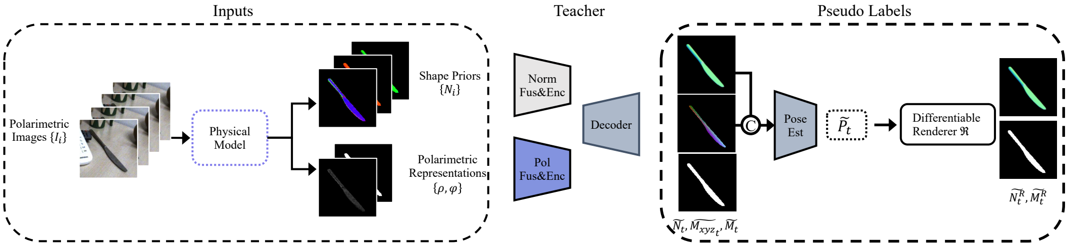

Inputs
Teacher
Pseudo Labels
Polarimetric Images $\{I_i\}$
Physical Model
Shape Priors $\{N_i\}$
Polarimetric Representations $\{\rho, \varphi\}$
Norm Fus&Enc
Pol Fus&Enc
Decoder
C
Pose Est
$\tilde{P}_t$
Differentiable Renderer $\mathfrak{R}$
$\tilde{N}_t^R, \tilde{M}_t^R$
$\tilde{N}_t, \widetilde{M_{xyz_t}}, \tilde{M}_t$

**Fig. 5.2** **S$^2$P$^3$ Teacher Network:** The network takes as input the shape priors and polarimetric representations, which are derived from the analytical physical model applied to four polarized images. Before retrieving the 6D object pose, the network makes predictions for intermediate geometrical representations. Subsequently, a differentiable renderer is employed, utilizing the predicted pose to generate a rendered normal map and object mask.

an object mask $\tilde{M}_t$, an object normal map $\tilde{N}_t$, and dense correspondences represented as a normalized object coordinate map $\tilde{M}_{xyz_t}$.

The spatial and shape correlations between $\tilde{M}_{xyz_t}$ and $\tilde{N}_t$ are then utilized as inputs to an object pose estimation module [165]. In this module, the predicted rotation vector is parameterized in the form of allocentric continuous 6D representation [195], while the predicted translation is encoded as a scale-invariant vector [98]. These parameters are subsequently transformed into standard forms: a rotation matrix $\tilde{R}_t \in \mathbb{R}^{3\times3}$ and a translation vector $\tilde{t}_t \in \mathbb{R}^3$. The collective outcome is denoted as the final pose, $\tilde{P}_t = [\tilde{R}_t \mid \tilde{t}_t]$.

We employ a differentiable renderer to derive pixel-wise geometrical pseudo labels from the predicted pose. This renderer takes as input the object's CAD model and $\tilde{P}_t$ to generate an object mask $\tilde{M}_t^R$ and an object normal map $\tilde{N}_t^R$. All the predicted and rendered quantities serve as weak pseudo labels for the student network.

### 5.4.1.2 Student Network

We introduce a lightweight student network, different to Self6D++ [163], which does not feature an explicit geometric decoder. In our approach, the network directly regresses the predicted pose for the student, denoted as $\hat{P}_s$ (as illustrated in Fig. 5.3). This design choice

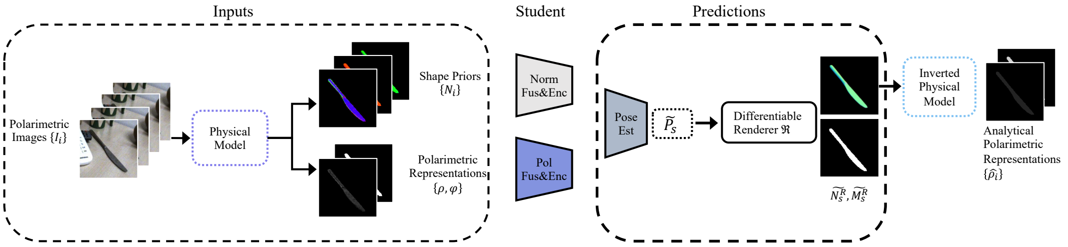

Inputs
Student
Predictions
Polarimetric Images $\{I_i\}$
Physical Model
Shape Priors $\{N_i\}$
Polarimetric Representations $\{\rho, \varphi\}$
Norm Fus&Enc
Pol Fus&Enc
Pose Est
$\hat{P}_s$
Differentiable Renderer $\mathfrak{R}$
$\tilde{N}_S^R, \tilde{M}_S^R$
Inverted Physical Model
Analytical Polarimetric Representations $\{\hat{\rho}_i\}$

**Fig. 5.3** **S$^2$P$^3$ Student Network:** In contrast to the teacher network depicted in Fig. 5.2, the student network, adopts a more lightweight architecture by omitting the explicit decoding of predicted intermediate geometric representations.

not only facilitates faster inference but also maintains a high level of accuracy. Our ablations, discussed in Table 5.4 later, affirm the superiority of our student network design.

The teacher network encompasses approximately 5.5 million weights, whereas our lightweight student does not require an explicit decoder, reducing the network to around 5 million weights. Although the reduction in the number of parameters may not be significant, the gain in inference speed and pose prediction accuracy is substantial. We compare this design later to the approach in Self6D++ [163], where the student network mirrors the teacher but omits a subsequent pose refiner.

Our student network converges towards superior predictions without the redundant explicit prediction of intermediate geometric representations, due to our proposed self-supervision mechanism. Consequently, the output exclusively contains the predicted pose of the student, $\hat{P}_s$. To establish the connection between these predictions and the geometric and polarimetric properties, we employ a differentiable renderer to generate an object normal map $\hat{N}_s$ and an object mask $\hat{M}_s$ based on $\hat{P}_s$, analogous to the teacher network. Polarimetric properties are analytically derived from the normal map. We elaborate on how this polarimetric representation of geometric information is utilized in a self-supervised loss term.

## 5.4.2 Physics-Induced Self-Supervised Training Scheme

As previously elaborated, polarimetric images encompass valuable information we provide as explicit representations to the network, allowing it to learn neural geometric encodings. This section outlines how these representations are leveraged and integrated into our physically induced self-supervised framework. The self-supervision is integrated through two aspects: firstly, through the provision of implicit and explicit weak pseudo-labels generated by the teacher network, and secondly, by establishing a direct coupling, ultimately returning to the input polarization information of the pipeline.

### 5.4.2.1 Loss Formulations

Our proposed optimization scheme encompasses two complementary paradigms (cf. Fig. 5.4). The first paradigm involves the transfer of knowledge from the pre-trained teacher network to the student network in the form of weak labels for the pose $\tilde{P}_t$ and associated object shape information $\{\tilde{M}_t, \tilde{N}_t, \tilde{M}_t^R, \tilde{N}_t^R\}$. We refer to this aspect as the pseudo label loss, denoted as $\mathcal{L}_{pseudo}$. The second paradigm leverages the inverted physical model to optimize the student's pose prediction $\hat{P}_s$ using raw polarization data within our physical loss term, denoted as $\mathcal{L}_{physics}$, which will be detailed below.

Given our objective of training the student network on real data without access to ground truth labels, the teacher network assumes the role of a pseudo ground truth provider. The loss terms provided by the teacher network are collectively denoted as $\mathcal{L}_{pseudo}$, comprising a direct pose loss denoted as $\mathcal{L}_{pose}$ and a geometrical loss denoted as $\mathcal{L}_{geo}$.

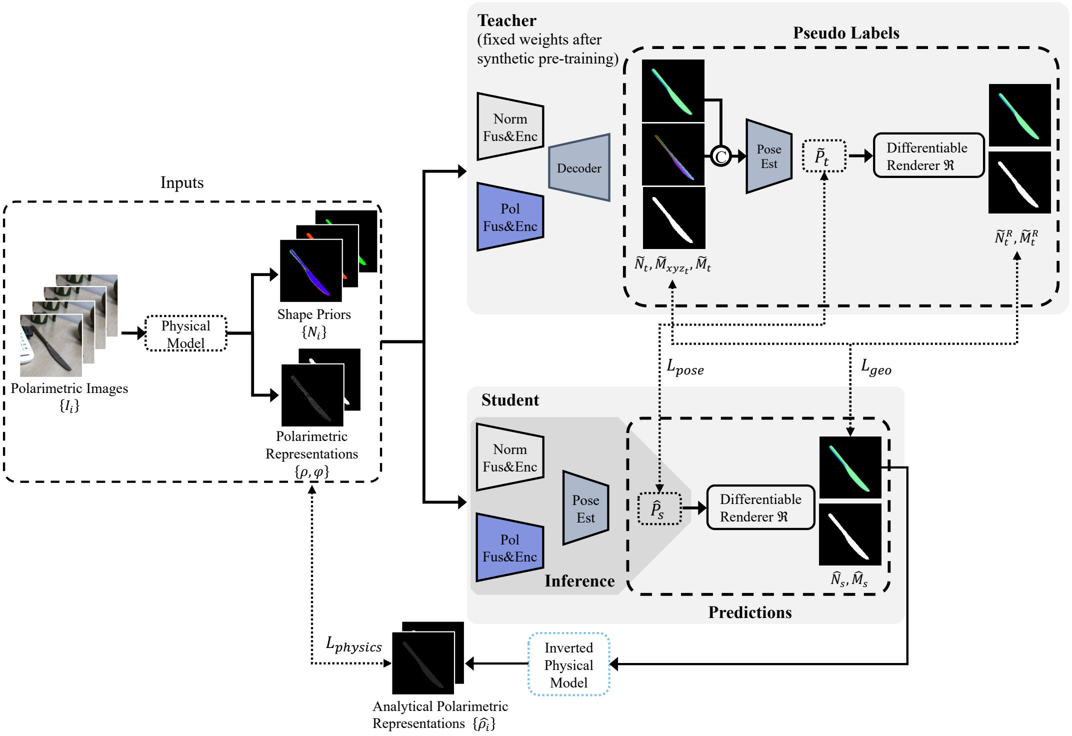


Fig. 5.4 **S$^2$P$^3$ Pipeline:** Our proposed teacher-student training scheme utilizes four polarization images captured under varying polarization filter angles and polarimetric and geometrical representations derived from the physical model as inputs for both the teacher and student networks. During training on real data, the student network is optimized not only for the pseudo labels generated by the teacher network, denoted as $L_{pseudo}$ (consisting of $L_{pose}$ and $L_{geo}$, but also by including a physics-based loss term, denoted as $L_{physics}$. This additional loss term minimizes the discrepancy between the DoP $\rho$ calculated by the physical model and the predicted DoP $\hat{\rho}$ derived from the inverted physical model. During the inference stage, the lightweight student network exclusively predicts direct pose estimates, as indicated by the gray background color.

For the pseudo pose loss $\mathscr{L}_{pose}$, we employ the pose predicted by the teacher network $\tilde{P}_t$ as the pseudo ground truth. We apply a Point-Matching loss [97] to quantify the alignment between:

$$\mathscr{L}_{pose} = \underset{x \in \mathscr{M}}{\operatorname{avg}} \|(\tilde{R}_t x + \tilde{t}_t) - (\hat{R}_s x + \hat{t}_s)\|_1. \tag{5.7}$$

For the geometrical loss term, we choose pseudo labels from two sets of geometrical information, namely the predicted geometrical representations $\{\tilde{M}_t, \tilde{N}_t\}$ and the rendered geometrical representations $\{\tilde{M}_t^R, \tilde{N}_t^R\}$. We make this distinction because the predicted geometrical representations are more likely to encode the correct pose information, leading to accurate alignment on the image plane. However, these predictions may not be pixel-perfect in terms of geometrical meanings. Conversely, the rendered representations exhibit flawless geometrical meanings but may introduce incorrect underlying pose information if the predicted pose $\tilde{P}_t$ deviates from the ground truth pose.

To leverage the accurate underlying pose information from the predictions and the precise geometrical meanings from the renderings, we introduce a misalignment coefficient $\delta$. This coefficient is calculated and normalized based on the discrepancy between the predicted mask $\tilde{M}_t$ and the rendered mask $\tilde{M}_t^R$. Suppose $\delta$ falls within a predefined threshold $r$, indicating that the teacher's predicted pose is reliable. In that case, we select the rendered representations $\{\tilde{M}_t^R, \tilde{N}_t^R\}$ as the geometrical pseudo ground truth to guide the student network. Otherwise, in cases where the large misalignment suggests a significant deviation of $\tilde{P}_t$ from the ground truth pose, we down-weight the pseudo pose loss $\mathscr{L}_{pose}$ by a factor of $\lambda = (1-\delta)$.

Finally, the pseudo label loss is defined as follows:

$$\mathscr{L}_{pseudo} = \lambda_1 \mathscr{L}_{pose} + \mathscr{L}_{geo}, \tag{5.8}$$

where the geometrical loss is combined of a mask loss and a normal loss as:

$$\mathscr{L}_{geo} = \mathscr{L}_{mask} + \mathscr{L}_{normals}, \tag{5.9}$$

where we take the L2-Loss for $\mathscr{L}_{mask}$, and the cosine-similarity loss for $\mathscr{L}_{normal}$.

**Physical Constraints** In order to facilitate self-supervision through the invertible physical model, we utilize the rendered geometric normal map $\hat{N}_s$ of the student as an input to compute analytical diffuse and specular DoP, denoted as $\{\hat{\rho_d}, \hat{\rho_s}\}$, in accordance with Equation 5.6. This approach allows us to derive training signals directly from the raw DoP $\rho$ extracted from real polarization images.

To leverage the inherent physical processes of polarimetric imaging, our physical loss $\mathscr{L}_{physics}$ incorporates a pixel-wise minimum selection mechanism for diffuse and specular solutions, inspired by the work of Verdie et al. [161]:

$$\mathscr{L}_{physics} = \min_{x \in \{\hat{\rho_d}, \hat{\rho_s}\}} \|\rho - x\|_1. \tag{5.10}$$

To mitigate the potential domain gap between the analytically derived intensity map and real polarimetric images, as discussed in [161], we formulate the loss function directly based on polarimetric properties rather than polarimetric intensities. Consequently, the student's output is fine-tuned to align with the raw Degree of Polarization ($\rho$) obtained from real polarization images.

The overall loss inlcudes information from both the teacher network and the raw data, defined as follows:

$$\mathscr{L} = \mathscr{L}_{pseudo} + \mathscr{L}_{physics}. \tag{5.11}$$

### 5.4.3 Training Procedure

Our training approach for $\mathbf{S}^2\mathbf{P}^3$ is structured into two distinct phases: "Synthetic Pre-Training" using rendered data and "Self-Supervised Training on Real Data."

#### 5.4.3.1 Synthetic Data Generation

Given a CAD model of an object, we employ random sampling of camera locations situated on its upper hemisphere for rendering. To enhance the realism of the rendered images and minimize domain discrepancies, we incorporate various backgrounds with diverse textures and lighting positions using the Mitsuba2 renderer [116]. This results in the generation of 200 to 800 sets of polarization images for each object.

We provide visual representations of our synthetic dataset from various viewpoints, as depicted in Fig. 5.5. These illustrations showcase the diversity of sampled poses, encompassing objects with varying photometric complexity, and offer insights into their appearance within the images. The synthetic dataset is used for pretraining teacher and student networks. In this dataset, we generate four polarimetric images with distinct polarization filter angles, mirroring the camera configuration utilized in the real setup.

#### 5.4.3.2 Synthetic Pre-Training

In this phase, the teacher and student models undergo pre-training with supervision based on 6D pose information from synthetic data. This stage employs a dual-component loss function: an L1 loss for translation and a point matching loss for rotation. Notably, the differentiable renderer is not used in this phase. Pre-training typically spans 4 to 5 hours per object. Following this, the self-supervised phase, which is more computationally demanding, requires about 10 hours per object.

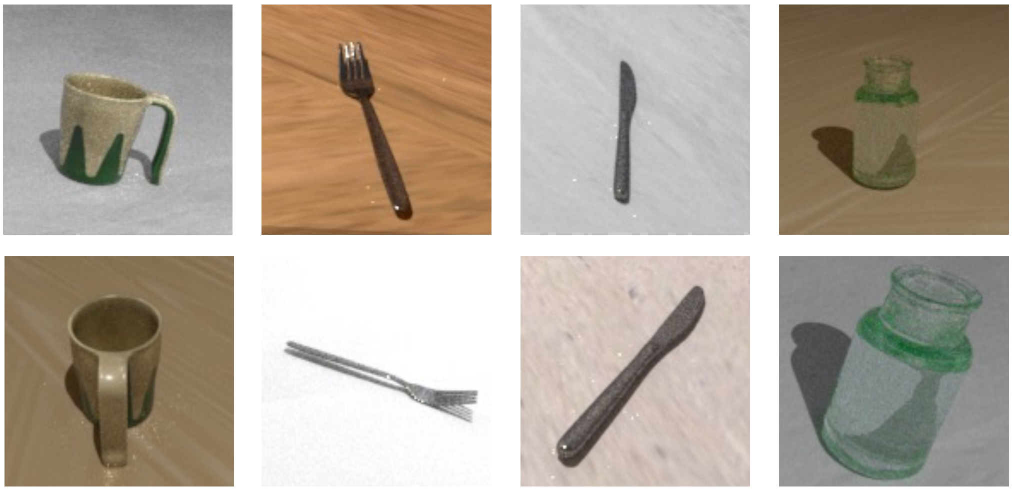


Fig. 5.5 **Synthetic Dataset:** We present samples of objects with diverse photometric complexities, depicted from various viewpoints.

#### 5.4.3.3 Self-Supervised Training on Real Data

We apply our method to a specific data split from the instance-level 6D pose estimation dataset described in the previous chapter, featuring objects with diverse photometric characteristics. For training, we sample approximately 15%—20% of the total data per object (200-300 sets of real polarimetric images) and allocate the remainder for testing (1000-2000 image sets). To minimize domain shift and enhance the effectiveness of self-supervision, we ensure that the pose distribution in the rendered synthetic data closely mirrors that in the real data. During processing, the predicted bounding box isolates the object of interest and scales it to 256 × 256 for input into the networks. The predicted object mask is then used in the physical model to generate object-specific polarimetric parameters and shape priors, which are crucial for accurate and robust pose estimation.

#### 5.4.3.4 Implementation Details

Our model is implemented using PyTorch [121] and trained on an NVIDIA 2080 GPU. The training process is conducted with the ADAM optimizer [85] on a standard desktop PC with an Intel i7 CPU processor and 32GB of RAM. We employ a customized object segmentation network to avoid reliance on ground truth pose-related information in real data. This network generates pixel-wise object labels based on polarimetric inputs, allowing us to derive object bounding boxes without knowledge of the ground truth pose. The segmentation network is initially trained on synthetic data and subsequently used to predict object masks on real data, which are further converted into bounding box information. These predicted bounding boxes enable a dynamic zoom-in strategy for both the teacher and student networks, similar to the approach used in PPP-Net [42]. Specifically, we crop the image region containing the target object and resize it to 256 × 256 for network input. The predicted mask helps filter out irrelevant information when computing physical priors.

Training of the teacher and student networks follows a two-phase process. Initially, both networks are trained on synthetic data with full supervision. Subsequently, the teacher-student training scheme is executed on real data. We employ the differentiable renderer from PyTorch3D [129] and customized shading functions to generate the required object geometrical representations. To expedite the render-and-compare training approach, we render objects on a cropped and zoomed-in image plane of dimensions 256 × 256. The camera intrinsics for each image are consistent with those used during the dynamic zoom-in operation.

For both the pre-training and self-supervised training stages of the teacher and student networks, we set the number of epochs to 100 for each object, individually, for both synthetic and real data. The initial learning rate is established at $1 \times 10^{-4}$ and is halved every 25 epochs. A batch size of 8 is utilized during pre-training, while a batch size of 4 is employed for self-supervised training. Regarding the network architecture, we employ ResNet-34 [56] as the encoder backbone for polarimetric and geometric feature extraction in both networks.

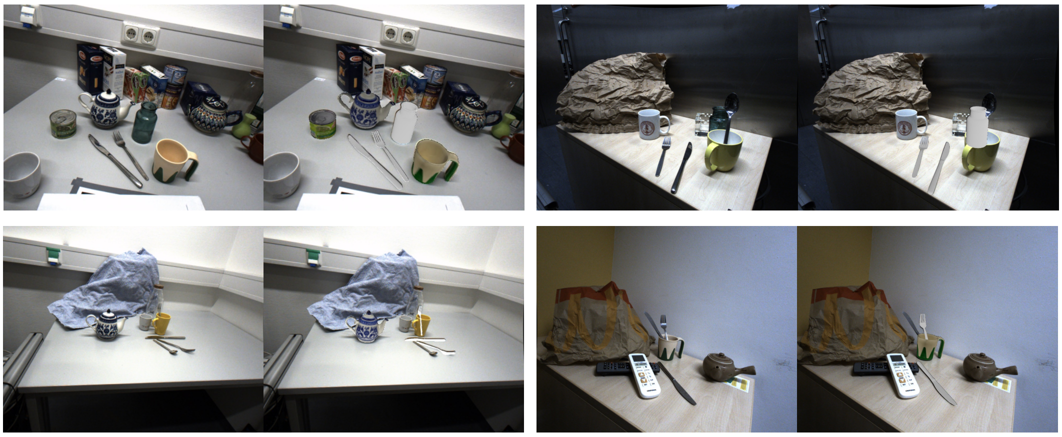


Fig. 5.6 **Real Dataset:** Samples of objects are depicted from various viewpoints. The rendered objects utilizing ground truth pose illustrate the texture rendered in white color.

## 5.5 Experimental Results

We extensively evaluate and conduct ablations using the instance-level polarimetric 6D pose dataset introduced in the previous chapters, where our supervised PPP-Net serves as a strong baseline for comparison outperforming other state-of-the-art supervised methods, including RGB-only [165] and RGB-D [57] approaches. We provide detailed quantitative results on real data and perform extensive ablations on different loss terms and modalities. Our experiments specifically investigate the influence of polarimetric physical cues in a self-supervised scheme on objects of varying photometric complexity for instance-level 6D pose prediction. Both polarimetric images and self-supervised schemes are relatively unexplored in 6D pose estimation. Therefore, we consider the supervised PPP-Net and the self-supervised Self6D++ [163], trained on RGB and RGB-D data respectively, as solid baselines for comparison.

Self6D++ [163] represents the state-of-the-art method in self-supervised 6D object pose estimation with RGB-D information, consistently outperforming other baselines by a significant margin [147, 164]. As such, it provides a valid benchmark for evaluating and justifying the improvements introduced by our method. Similarly, PPP-Net already outperforms state-of-the-art RGB-only methods, especially on photometrically challenging objects, making it a suitable representative of RGB-only methods as a strong baseline for our experiments.

In Fig. 5.6, we showcase examples from our real polarimetric dataset with annotated object poses. These samples illustrate the high quality of our data annotation, as evidenced by the objects rendered using ground truth pose labels. Additionally, the objects depicted in white color rendering, indicate their textureless nature. This characteristic of the dataset aligns well with our motivation, which intentionally omits the requirement for color texture supervision in the learning process.

For non-symmetrical objects, the results are assessed using the widely adopted Average Distance of Distinguishable Model Points (ADD) metric [60]. In this metric, a threshold equivalent to 10% of the object's diameter is employed to determine the average deviation of the transformed model points. For symmetric objects, the evaluation employs the Distance of Indistinguishable Model Points (ADD-S) metric [63], which measures the average deviation to the closest model points.

On a desktop computer equipped with an Intel i7 CPU at 4.20 GHz and an NVIDIA 2080 GPU, our student network requires approximately 7.3 milliseconds to infer the 6D pose for a single object from a $512 \times 612$ image. This represents an improvement in speed of around 30% compared to the teacher model. Furthermore, the preprocessing steps for calculating the physical priors take 13.0 milliseconds, and the object detection process requires 15.4 milliseconds.

### 5.5.1 Baseline Comparisons

$\mathbf{S^2P^3}$ introduces a novel approach for self-supervised 6D object pose estimation by leveraging polarimetric information. It specifically addresses the challenges posed by photometrically complex objects where conventional self-supervised RGB-D methods may struggle due to inherent sensor data artifacts. Additionally, supervised approaches, whether RGB-only or RGB-P, such as PPP-Net, would require a substantial amount of annotated real data, making them less practical.

Our experimental design is intentionally tailored to comprehensively analyze the effectiveness of multi-modal self-supervision using physical constraints, loss functions, and various architectural and design choices within the teacher-student framework. This approach allows us to gain valuable insights into self-supervised polarimetric 6D pose estimation.

In our evaluation, we compare $\mathbf{S^2P^3}$ against PPP-Net on our dataset split, serving as an exceptionally strong supervised baseline. This analysis helps us understand the impact of self-supervision, considering that PPP-Net has already demonstrated superior performance compared to other state-of-the-art RGB-only methods. Additionally, we evaluate $\mathbf{S^2P^3}$ against Self6D++ [163], a state-of-the-art self-supervised RGB-D method widely recognized on standard benchmark datasets. For visual results and further insights, please refer to Figs. 5.7, 5.8 and Figs. 5.9, 5.10, which showcase qualitative results, including object occlusions.

We demonstrate the effectiveness of our self-supervision pipeline through quantitative results presented in Tab. 5.1. Note that PPP-Net, which is trained on annotated real data, shares the same network architecture as our teacher model. However, in our full model $\mathbf{S^2P^3}$, we refrain from supervised training of the teacher on real data. Instead, we pre-train it exclusively on synthetic data, after which the teacher's weights are frozen, and it exclusively provides weak pseudo-labels for the teacher-student scheme on real data. Our $\mathbf{S^2P^3}$ model consistently outperforms the self-supervised Self6D++ RGB-D method [163]. Self6D++ is trained and tested on our dataset, utilizing RGB-D information from the Realsense L515 sensor. It performs comparably to the fully supervised upper-bound baseline for objects with high photometric complexity.

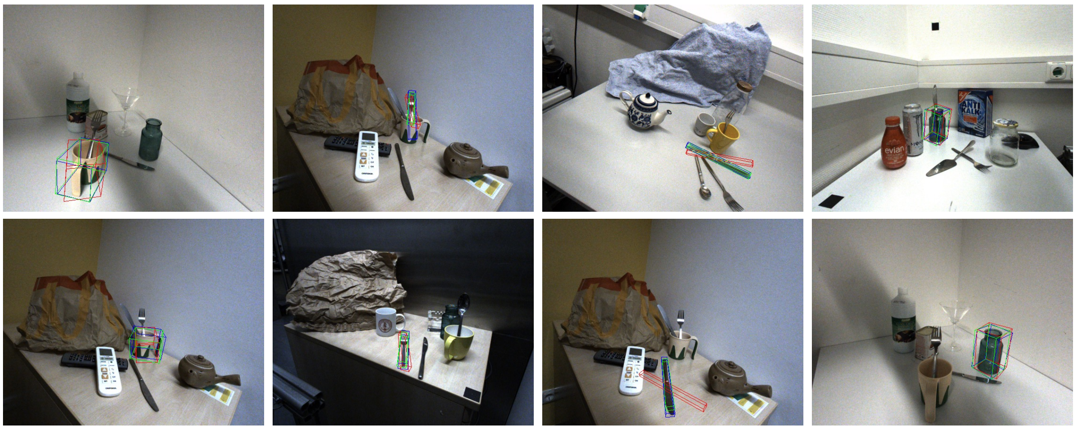


Fig. 5.7 **$S^2P^3$ Qualitative Results:** Before and after self-supervision. The projected bounding boxes in blue, red, and green represent the ground-truth 6D object poses, the results before and after applying self-supervision, respectively.

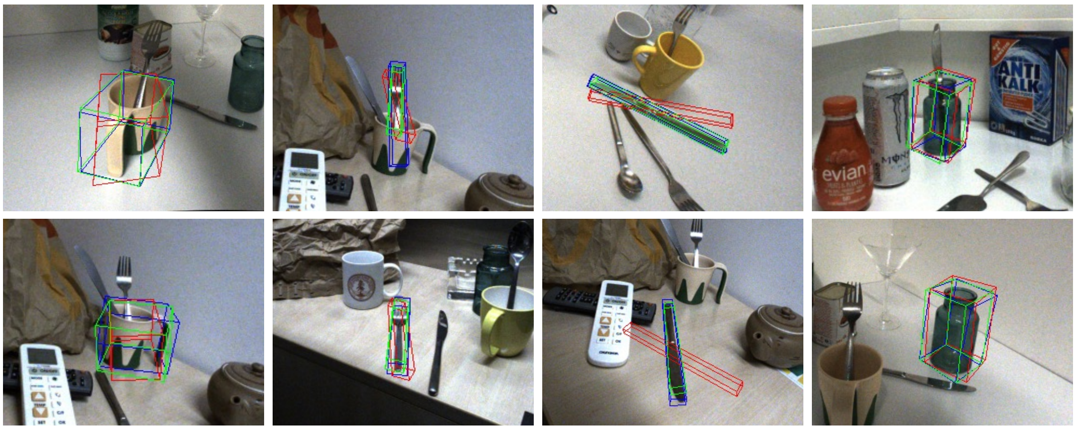


Fig. 5.8 **$S^2P^3$ Qualitative Results:** Zoomed-in from Fig. 5.7. Before and after self-supervision. The projected bounding boxes in blue, red, and green represent the ground-truth 6D object poses, the results before and after applying self-supervision, respectively.

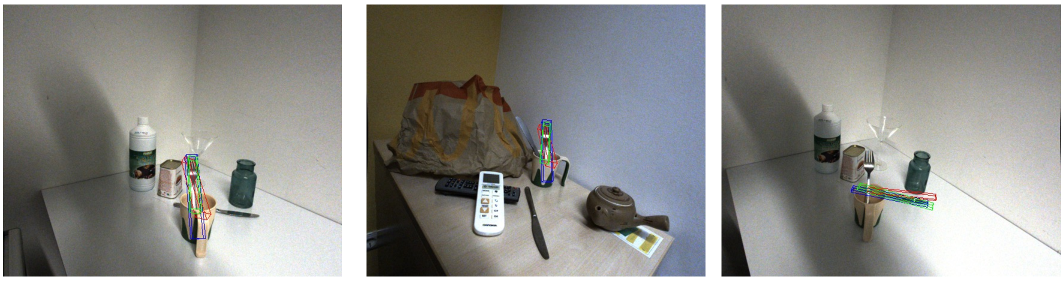


Fig. 5.9 **$S^2P^3$ Qualitative Results with Occlusions:** Before and after self-supervision. The projected bounding boxes in blue, red, and green represent the ground-truth 6D object poses, the results before and after applying self-supervision, respectively.

### 5.5.2 Ablation Studies

Our evaluation includes a series of ablation studies designed to dissect and understand the various aspects of our model. We analyze each component of our loss function, especially

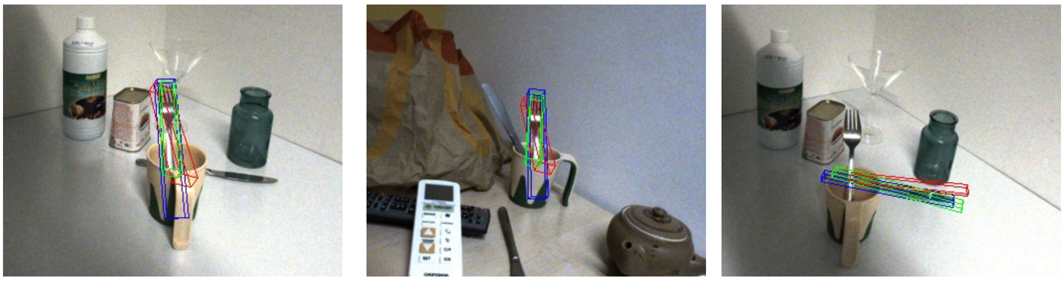


Fig. 5.10 **S$^2$P$^3$ Qualitative Results with Occlusions:** Zoomed-in from Fig. 5.9. Before and after self-supervision. The projected bounding boxes in blue, red, and green represent the ground-truth 6D object poses, the results before and after applying self-supervision, respectively.

Tab. 5.1 **S$^2$P$^3$ Quantitative Results:** The average recall of the ADD(-S) metric is reported for various objects with increasing photometric complexity. These results are compared with the performance of RGB-D self-supervised Self6D++ as presented in [163] and fully supervised PPP-Net as presented in the previous chapter.

| Methods | Training | Cup | Fork | Knife | Bottle | Mean |
|---|---|---|---|---|---|---|
| PPP-Net | Supervised | 91.4 | 91.7 | 90.0 | 89.4 | 90.6 |
| Self6D++ | Self-Supervised (RGB-D) | 68.4 | 14.3 | 17.8 | 33.5 | 34.0 |
| **S$^2$P$^3$** (Ours) | Self-Supervised (**RGB-P**) | **93.8** | **72.4** | **78.4** | **78.2** | **80.7** |

focusing on the role of our physically-induced self-supervised loss. This analysis helps highlight each loss component's contributions to the model's overall performance. We then examine how well the isolated student and teacher networks alone perform on real data when trained on synthetic data in a supervised manner. This comparison helps us understand the effectiveness of our **S$^2$P$^3$** training, particularly the benefits of combining supervised training on synthetic data with self-supervision on real data. We further explore whether a lighter-weight network model can match or surpass the performance of a larger model when employed as student network when refined with real data. This aspect assesses the necessity of a small or large student network and whether our approach of directly regressing the 6D pose for the student is advantageous. We also investigate the relative importance of depth information compared to polarimetric data to analyze which kind of data contributes more significantly to the model's accuracy and robustness in 6D pose estimation.

#### 5.5.2.1 Ablation on Loss Terms

We investigate the influence of various loss terms through a series of experiments where we exclude specific loss terms during the self-supervision stage, and we summarize the results in Tab. 5.2. Our findings underscore the critical role of direct geometrical point matching loss in the self-supervision process, denoted as $\mathcal{L}_{pose}$. Omitting this loss, which enforces alignment between the student's predictions and the weak pseudo-labels provided by the teacher, leads to a notable risk of training divergence. Furthermore, our physically-induced self-supervised loss, $\mathcal{L}_{physics}$, derived from our invertible physical model, demonstrates a substantial impact on training outcomes comparable to geometrical supervision signals from the teacher network, such as $\mathcal{L}_{normal}$ and $\mathcal{L}_{mask}$. This indicates that the real polarimetric images capture robust un-

Tab. 5.2 **Ablation on Loss Terms:** Average recall of ADD(-S) metric is reported.

| Methods | Cup | Fork | Knife | Bottle | Mean |
|---|---|---|---|---|---|
| w/o $\mathscr{L}_{pose}$ | 6.8 | 0.2 | 2.3 | 0.6 | 2.5 |
| w/o $\mathscr{L}_{physics}$ | 71.8 | 72.1 | 70.8 | 74.4 | 72.3 |
| w/o $\mathscr{L}_{normal}$ | 87.5 | 61.0 | 67.3 | 74.9 | 72.7 |
| w/o $\mathscr{L}_{mask}$ | 89.9 | 64.9 | 70.1 | 72.7 | 74.4 |
| $\mathbf{S^2P^3}$ (Ours) | **93.8** | **72.4** | **78.4** | **78.2** | **80.7** |

derlying object shape information, which seems more beneficial than the output produced by the differentiable renderer. Ultimately, the overall model performance attains the highest accuracy metrics across all objects with varying photometric complexity when all loss components are considered, as indicated in the last row of Tab. 5.2.

These findings underscore the critical role of the teacher network's weak labels in guiding the student network's pose predictions. One possible explanation for this behavior is that without the inclusion of $\mathscr{L}_{pose}$, the differentiable renderer lacks essential constraints, potentially leading to outputs with pose predictions that fall outside the field of view. Introducing dense supervision for appearance and geometric representations following differentiable rendering further enhances the network's overall performance. Substantial improvement in pose accuracy is achieved through our proposed self-supervised physically-induced loss formulation.

The impact of self-supervision is also evident in our qualitative results, as illustrated in Fig. 5.7. The projected bounding boxes in green exhibit better alignment with the ground truth (blue) after the application of self-supervision, in contrast to predictions from the pre-trained teacher (red). Additionally, Fig. 5.9 showcases results for scenarios in which parts of the object, such as the *fork* and *knife*, are occluded.

#### 5.5.2.2 Ablation on Domain Shift - S$^2$P$^3$'s Self-Supervision

Table 5.3 provides insights into the performance of individual student and teacher networks trained separately, a scenario distinct from our integrated $\mathbf{S^2P^3}$ training scheme. This training, conducted without the differentiable renderer, focuses on supervised pose estimation similar to the synthetic pre-training phase (i.e., the synthetically pre-trained networks correspond to the numbers of the lower part of Tab. 5.3 without self-supervision). The training is differentiated based on whether it occurs on real or synthetic annotated data, with all testing carried out on real data.

The results reveal that student and teacher networks exhibit decreased performance on real data when their training is limited to synthetic data. This decline results from the domain shift between synthetic and real-world environments. Notably, the teacher network, which includes a dedicated decoder and explicit intermediate geometrical representations (identical to PPP-Net [42] and marked with † in the table), consistently outperforms the smaller student network in both training scenarios.

Tab. 5.3 **Domain Shift and S$^2$P$^3$'s Self-Supervision:** This table presents the average recall of the ADD(-S) metric for various objects, showcasing a spectrum of photometric complexities. The performance of both the student and teacher networks is evaluated when trained separately under supervised conditions on either real or synthetic data, with all testing conducted on real data. Additionally, the complete **S$^2$P$^3$** pipeline, which includes synthetic pre-training followed by self-supervised training of the student on non-annotated real data, is compared for context. In this table, "Teacher †" represents the performance benchmark, being equivalent to PPP-Net [42], and "Student ⋆" reflects the **S$^2$P$^3$** performance **prior** to the implementation of our self-supervision approach.

| Configuration | Supervised | Self-Supervised | Tested on | Cup | Fork | Knife | Bottle | Mean |
|---|---|---|---|---|---|---|---|---|
| Student | Real | - | Real | 86.4 | 88.0 | 91.1 | 80.4 | 86.5 |
| Teacher † | Real | - | Real | **91.4** | **91.7** | **90.0** | **89.4** | **90.6** |
| Student ⋆ | Synthetic | - | Real | 53.7 | 64.4 | 46.1 | 47.5 | 52.9 |
| Teacher | Synthetic | - | Real | 72.3 | **75.0** | 67.3 | 76.2 | 72.7 |
| **S$^2$P$^3$** (Ours) | Syn. (Pre-trained) | Real | Real | **93.8** | 72.4 | **78.4** | **78.2** | **80.7** |

However, when implementing our full **S$^2$P$^3$** pipeline, where the student is now trained in a self-supervised manner on real data, and the teacher's weights are fixed, the results are impressive. Thanks to our self-supervision paradigm, this performance is achieved without training on real image annotations. Noticably, **S$^2$P$^3$** even surpasses the fully supervised training on real data in some cases and achieves comparable results to the fully supervised PPP-Net.

A notable observation is that the self-supervision in **S$^2$P$^3$** enhances the performance compared to the synthetically pre-trained student network. While this trend is consistent across all objects, the improvement is less pronounced for the *fork*, potentially due to significant occlusions in the majority of its data (cf. Figures 5.9 and 5.10 where the fork is inside the cup).

#### 5.5.2.3 Ablation on Network Architecture - Student Exchange

We explore the potential benefits of using a lightweight student network for faster inference. To understand the impact of a more complex network architecture, we replace the student network in **S$^2$P$^3$** with the architecture typically used for the teacher network (as depicted in Fig. 5.2) instead of the one shown in Fig. 5.3. This substitution allows us to analyze the influence of a larger network for the student, complete with a dedicated decoder and intermediate geometrical representations.

Our analysis reveals that while the larger student network with intermediate geometrical outputs shows advantages during pre-training in a supervised setting, these outputs can complicate the optimization process during self-supervised learning of the 6D object pose. Tab. 5.4 indicates that the lightweight student network (Our Student) actually outperforms the larger student network (Large Student) after fine-tuning on real data with our teacher-student training scheme and self-supervision through physical constraints $\mathscr{L}_{physics}$. The larger student network, with its additional parameters and intermediate outputs, faces more challenges in converging effectively. However, the application of physical constraints significantly enhances its performance after self-supervision (cf. Large student with None and with Self-Supervision). Overall, the ablation studies demonstrate that a lightweight student network can achieve su-

Tab. 5.4 **Ablation on Network Architecture:** A comparative analysis of various student network architectures is summarized, specifically contrasting our smaller student network design against a larger student architecture equivalent to that of the teacher. Our self-supervised student network configuration consistently delivers the best performance across all objects. The comparison is based on the average recall of the ADD(-S) metric.

| Config | Self-Sup. | Cup | Fork | Knife | Bottle | Mean |
|---|---|---|---|---|---|---|
| Our Student | None | 53.7 | 64.4 | 46.1 | 47.5 | 52.9 |
| Large Student | None | 72.3 | 75.0 | 67.3 | 76.2 | 72.7 |
| Our Student | ✓($\mathbf{S}^2\mathbf{P}^3$) | **93.8** | **72.4** | **78.4** | **78.2** | **80.7** |
| Large Student | ✓ | 88.6 | 55.9 | 69.4 | 77.8 | 73.0 |

perior performance compared to a larger network after being fine-tuned on real data within our self-supervised scheme.

#### 5.5.2.4 Ablation on Modalities

**RGB-Texture Supervision** The rendered object's texture remains white for objects that lack texture or are transparent, as these objects do not possess color information. Consequently, this simplifies the RGB-texture loss in our pipeline, essentially reducing it to the mask loss. As a result, we eliminate the necessity for texture rendering and instead leverage the inherent physical properties of polarized light.

**Depth Supervision** To assess the significance of precise and dependable geometric representations in the context of 6D object pose estimation, we train our pipeline using depth maps obtained from a direct time-of-flight (D-ToF) sensor. We then compare this approach with our polarimetric $\mathbf{S}^2\mathbf{P}^3$ method, which employs our physically-induced self-supervised loss.

For this purpose, we introduce an additional loss term into our network while keeping almost all other components unchanged. Specifically, we extend the capabilities of the student network's differentiable renderer to generate depth maps $D^R$ based on the predicted pose $\hat{P}_s$. We then employ a chamfer distance loss $\mathcal{L}_{chamfer}$, which measures the dissimilarity between the point cloud $P^R$ obtained from back-projecting the rendered depth $D^R$ and the point cloud $P$ obtained from back-projecting the depth map in the polarization camera coordinate system. This additional loss term helps optimize the alignment between the two point clouds without requiring explicit 3D-3D correspondence registrations. The formulation is as follows:

$$\mathcal{L}_{chamfer} = \underset{p \in P}{\operatorname{avg}} \min_{p^r \in P^R} \|p - p^r\|_2 + \underset{p^r \in P^R}{\operatorname{avg}} \min_{p \in P} \|p - p^r\|_2. \tag{5.12}$$

In addition to incorporating $\mathcal{L}_{chamfer}$ into the pipeline, we have excluded $\mathcal{L}_{physics}$ to ensure a fair comparison of the effectiveness of direct spatial cues from depth and object shape cues derived from polarimetric physical properties. The results, as summarized in Tab. 5.5, demonstrate that depth cues can be advantageous when the quality is reliable, e.g., the perfor-

Tab. 5.5 **$S^2P^3$ Ablations on Depth Modality:** Average recall of ADD(-S) metric is reported.

| Methods | Cup | Fork | Knife | Bottle | Mean |
|---|---|---|---|---|---|
| OURS$_{(\text{RGB-D})}$ Chamfer | **100.0** | 11.6 | 59.1 | 40.7 | 52.9 |
| OURS$_{(\text{RGB-D})}$ Pixel-wise | 86.8 | 32.3 | 62.5 | 50.3 | 58.0 |
| OURS$_{(\text{RGB-P})}$ ($\mathbf{S^2P^3}$) | 93.8 | **72.4** | **78.4** | **78.2** | **80.7** |

mance for the photometrically simple cup object significantly improves with the introduction of $\mathscr{L}_{chamfer}$.

We perform additional ablation experiments employing a pixel-wise depth loss instead of the chamfer distance loss, as presented in Tab. 5.5. These experiments illustrate that, even with the pixel-wise depth loss, inaccurate depth information can introduce erroneous geometric guidance into the pipeline, resulting in reduced performance, particularly on photometrically challenging objects.

The intrinsic limitations of the depth sensor give rise to a significant degradation in depth quality [79]. Reflective and semi-transparent objects are particularly susceptible to inaccuracies owing to their materials' reflective and translucent characteristics. This issue is visually demonstrated in comprehensive large-scale examples shown in Fig. 5.11. In these instances, the pronounced signal derived from the depth alignment loss imparts erroneous spatial awareness, resulting in poor pose prediction performance. Conversely, the shape of the object, encoded within the polarimetric image modality, offers consistent geometric information for objects with diverse material characteristics. This stability extends across a wide range of photometric complexities, encompassing matte plastic cups, reflective stainless steel cutlery, and translucent or transparent colored glass objects. The analytically obtained diffuse and specular solutions following the differentiable renderer exhibit stability across all objects under consideration. These polarization properties are calculated using our invertible model and subsequently employed in the physically-induced self-supervision framework, as depicted in the top-left image showing the raw DoP. It is essential to highlight that $\mathscr{L}_{physics}$ represents a pixelwise minimum loss, considering both diffuse and specular reflections.

## 5.6 Conclusion

The conducted experiments underscore the significance of reliable geometric priors in 6D object pose estimation. In cases where the depth map's quality is high and dependable, the spatial loss term introduced from the source depth map may yield superior performance compared to a purely object-shape-based optimization relying on polarization cues. Notably, the current model is tailored for instance-level pose estimation and does not extend to generalization for unseen objects during training. An intriguing avenue for future exploration involves incorporating this concept into a category-level pipeline.

This chapter bridges two domains, uniting a hybrid model for polarimetric pose estimation that seamlessly combines an invertible physical model with neural shape extraction through a

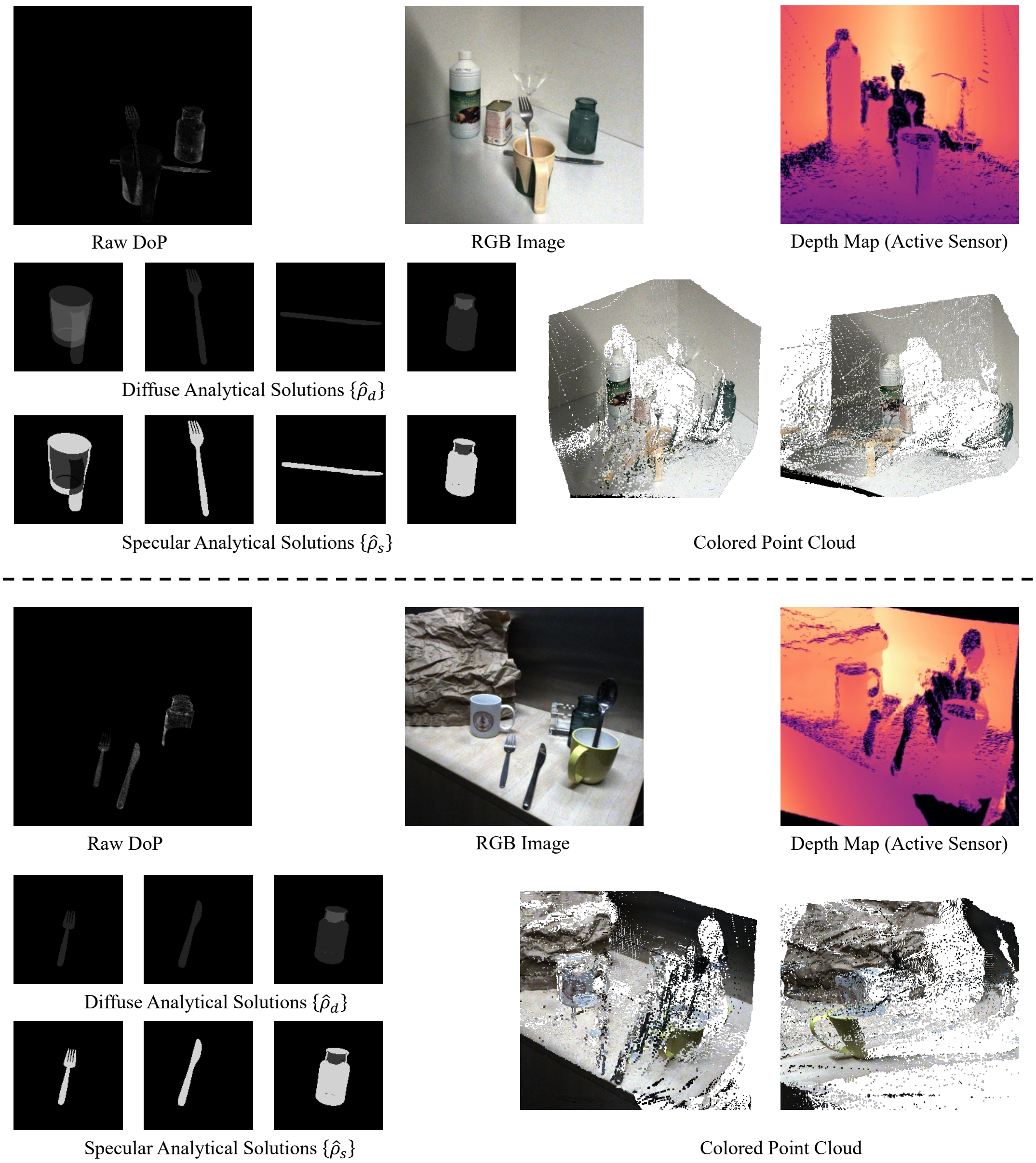

Raw DoP
RGB Image
Depth Map (Active Sensor)
Diffuse Analytical Solutions $\{\hat{\rho}_d\}$
Specular Analytical Solutions $\{\hat{\rho}_s\}$
Colored Point Cloud
Raw DoP
RGB Image
Depth Map (Active Sensor)
Diffuse Analytical Solutions $\{\hat{\rho}_d\}$
Specular Analytical Solutions $\{\hat{\rho}_s\}$
Colored Point Cloud

Fig. 5.11 **Examples of Polarimetric and Depth Quality.**

self-supervised framework. Our approach, denoted as $\mathbf{S}^2\mathbf{P}^3$, tackles the problem of instance-level 6D object pose estimation from polarimetric images without the need for annotated real data. In our proposed pipeline, a teacher network, pre-trained on a limited set of synthetic renderings, facilitates the convergence of a lightweight student network by providing weak pseudo-labels. Additionally, using a differentiable renderer yields appearance and geometric outputs, enabling effective self-supervision.

$\mathbf{S}^2\mathbf{P}^3$ outperforms methods that rely on depth measurements from active sensors, particularly for challenging photometric objects. We achieve this by carefully integrating various design choices within the teacher-student architecture and introducing our invertible physical model for self-supervision, which leverages XoP properties instead of raw polarimetric data as done in [161], thereby reducing the domain gap. Our contributions are rigorously validated through a series of comprehensive ablation studies.

Our experimental findings underscore the crucial role of self-supervision through geometric and physical cues in the domain of 6D pose estimation, offering valuable insights into the robustness of polarimetric images. These insights are particularly pronounced when dealing with photometrically challenging objects, such as those lacking texture or exhibiting reflective or translucent properties.

# Part V

## Conclusion

# 6 Conclusion

This dissertation proposed novel paradigms, methodologies, and evaluation procedures for enhancing robust learned 3D perception in challenging scenarios. A novel method ensuring temporal consistency in depth estimation for outdoor scenes and a new metric to quantify depth consistency was proposed. The dissertation also focused on creating accurate datasets for analyzing depth sensor limitations, whose artifacts affect different 3D vision tasks, like depth estimation, novel view synthesis, or 6D object pose estimation. The study also explored integrating polarization information, showing its effectiveness in challenging scenarios. It led to a novel 6D object pose estimation approach extended to self-supervision through an invertible physical polarimetric model.

The first chapter presented **TC-Depth**, which focused on self-supervised depth estimation from monocular camera ego-motion in a challenging environment, emphasizing temporal consistency. This method incorporated a spatial-temporal attention mechanism with robust geometric loss functions and a novel masking scheme. The approach was validated through experiments on standard datasets like KITTI and Cityscapes, where it outperformed existing methods in depth accuracy and temporal consistency. The Temporal Consistency Metric (TCM) was proposed as a new evaluation benchmark for depth consistency. Ablation studies highlighted the significance of each component, especially the geometric consistency loss and spatial-temporal attention mechanism, establishing **TC-Depth** as a significant advancement in self-supervised monocular depth estimation.

The subsequent chapter delved into creating a comprehensive multi-modal dataset including different depth sensors and polarization, crucial for analyzing sensor-specific artifacts in depth estimation, especially in scenarios involving reflective, textureless, or transparent materials. The first part of this chapter outlined the proposed dataset acquisition process, focusing on ensuring high accuracy for the depth ground truth data and object pose annotations. The second part provided a detailed analysis of various depth sensors, comparing their performances against dense depth ground truth. This exploration was extended to multi-view scenarios to study dense 3D vision tasks, including depth estimation and novel scene synthesis, using RGB-only information or integrating depth sensor data. The chapter aimed to develop more accurate and robust 3D perception methods for photometrically challenging environments.

The final chapters concentrated on integrating polarimetric data into 3D perception models. This integration was particularly effective for objects that conventional RGB and depth sensors struggle with due to their photometric properties. The research demonstrated that polarimetric data could offer robust shape and surface information, enhancing the ability to learn robust 6D object pose estimation in the proposed Polarimetric Pose Prediction network, short **PPP-Net**. A significant contribution was the development of a hybrid pipeline that combined polarimetric images with physical shape cues in a data-driven learning model.

The dissertation then explored leveraging polarimetric information for self-supervision, especially for 6D object pose estimation. An invertible physical model for polarimetric imaging was developed, allowing the derivation of polarimetric properties from geometric normal maps. This led to the creation of the Self-Supervised Polarimetric Pose Prediction pipeline ($\mathbf{S}^2\mathbf{P}^3$), which utilized polarimetric data to enhance pose estimation accuracy and robustness, particularly for objects with complex photometric characteristics. This approach avoids the need for annotated real data.

This dissertation marks a significant advancement in robust learned 3D perception, addressing critical challenges in depth estimation and object pose prediction. It introduces a temporally consistent monocular depth estimation pipeline, an in-depth analysis of depth sensors, and the novel integration of polarimetric data for 6D object pose estimation, setting new standards in accuracy and robustness for learned 3D perception. These breakthroughs open promising avenues for future applications in autonomous systems, robotics, and augmented reality, and enables innovative uses of sensor fusion and self-supervision in complex environments.

# List of Tables

# List of Figures

# Literature

[1] Henrik Aanæs, Rasmus Jensen, George Vogiatzis, Engin Tola, and Anders Dahl. Large-Scale Data for Multiple-View Stereopsis. *International Journal of Computer Vision*, **120**: 2016. (see pp. 45, 49)

[2] Gary A Atkinson. Polarisation Photometric Stereo. *Computer Vision and Image Understanding*, **160**: 158–167, 2017. (see p. 80)

[3] Gary A Atkinson and Edwin R Hancock. “Multi-view Surface Reconstruction using Polarization” in: *IEEE International Conference on Computer Vision (ICCV)*. vol. 1 IEEE 2005. 309–316 (see p. 80)

[4] Gary A Atkinson and Edwin R Hancock. Recovery of Surface Orientation from Diffuse Polarization. *Transactions on Image Processing*, **15**: 1653–1664, 2006. (see pp. 46, 80, 86)

[5] Yunhao Ba, Alex Gilbert, Franklin Wang, Jinfa Yang, Rui Chen, Yiqin Wang, Lei Yan, Boxin Shi, and Achuta Kadambi. “Deep Shape from Polarization” in: *European Conference on Computer Vision (ECCV)*. Springer 2020. 554–571 (see pp. 80, 87, 90, 97)

[6] V Madhu Babu, Kaushik Das, Anima Majumdar, and Swagat Kumar. “UnDEMoN: Unsupervised Deep Network for Depth and Ego-Motion Estimation” in: *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE 2018. 1082–1088 (see p. 16)

[7] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. “Mip-NeRF 360: Unbounded Anti-Aliased Neural Radiance Fields” in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022. 5470–5479 (see p. 48)

[8] Paul J Besl and Neil D McKay. “Method for Registration of 3D Shapes” in: *Sensor Fusion IV: Control Paradigms and Data Structures*. vol. 1611 International Society for Optics and Photonics 1992. 586–606 (see p. 81)

[9] Jiawang Bian, Zhichao Li, Naiyan Wang, Huangying Zhan, Chunhua Shen, Ming-Ming Cheng, and Ian Reid. “Unsupervised Scale-Consistent Depth and Ego-Motion Learning from Monocular Video” in: *Advances in Neural Information Processing Systems*. 2019. 35–45 (see pp. 4, 12–14, 16, 17, 24, 27, 30–32, 35)

[10] Tolga Birdal and Slobodan Ilic. “Point Pair Features based Object Detection and Pose Estimation Revisited” in: *IEEE International Conference on 3D Vision (3DV)*. IEEE 2015. 527–535 (see p. 81)

[11] Eric Brachmann, Alexander Krull, Frank Michel, Stefan Gumhold, Jamie Shotton, and Carsten Rother. “Learning 6D Object Pose Estimation using 3D Object Coordinates” in: *European Conference on Computer Vision (ECCV)*. Springer 2014. 536–551 (see p. 82)

[12] Eric Brachmann, Frank Michel, Alexander Krull, Michael Ying Yang, Stefan Gumhold, et al. “Uncertainty-driven 6D Pose Estimation of Objects and Scenes from a Single RGB Image” in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016. 3364–3372 (see p. 93)

[13] Benjamin Busam, Tolga Birdal, and Nassir Navab. “Camera Pose Filtering with Local Regression Geodesics on the Riemannian Manifold of Dual Quaternions” in: *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2017. 2436–2445 (see p. 13)

[14] Benjamin Busam, Matthieu Hog, Steven McDonagh, and Gregory Slabaugh. “SteReFo: Efficient Image Refocusing with Stereo Vision” in: *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. 2019. (see pp. 13, 43)

[15] Benjamin Busam, Hyun Jun Jung, and Nassir Navab. I like to move it: 6D Pose Estimation as an Action Decision Process. *arXiv preprint arXiv:2009.12678*, 2020. (see pp. 49, 79, 81)

[16] Benjamin Busam, Patrick Ruhkamp, Salvatore Virga, Beatrice Lentes, Julia Rackerseder, Nassir Navab, and Christoph Hennersperger. “Markerless Inside-Out Tracking for 3D Ultrasound Compounding” in: *Simulation, Image Processing, and Ultrasound Systems for Assisted Diagnosis and Navigation: International Workshops, POCUS 2018, BIVPCS 2018, CuRIOUS 2018, and CPM 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16–20, 2018, Proceedings*. Springer Springer, 2018. 56–64 (see p. 13)

[17] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. “A Naturalistic Open Source Movie for Optical Flow Evaluation” in: *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI 12*. Springer 2012. 611–625 (see p. 46)

[18] Tuo Cao, Fei Luo, Yanping Fu, Wenxiao Zhang, Shengjie Zheng, and Chunxia Xiao. DGECN: A Depth-Guided Edge Convolutional Network for End-to-End 6D Pose Estimation. *arXiv preprint arXiv:2204.09983*, 2022. (see p. 81)

[19] Vincent Casser, Soeren Pirk, Reza Mahjourian, and Anelia Angelova. “Depth Prediction without the Sensors: Leveraging Structure for Unsupervised Learning from Monocular Videos” in: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 33 01 2019. 8001–8008 (see p. 13)

[20] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D Data in Indoor Environments. *International Conference on 3D Vision (3DV)*, 2017. (see pp. 44, 50)

[21] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. TensoRF: Tensorial Radiance Fields. *arXiv preprint arXiv:2203.09517*, 2022. (see p. 48)

[22] Po-Yi Chen, Alexander H Liu, Yen-Cheng Liu, and Yu-Chiang Frank Wang. “Towards Scene Understanding: Unsupervised Monocular Depth Estimation with Semantic-aware Representation” in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019. 2624–2632 (see p. 47)

[23] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. (see p. 30)

[24] Zhaopeng Cui, Jinwei Gu, Boxin Shi, Ping Tan, and Jan Kautz. “Polarimetric Multi-view Stereo” in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017. 1558–1567 (see p. 80)

[25] Zhaopeng Cui, Viktor Larsson, and Marc Pollefeys. “Polarimetric Relative Pose Estimation” in: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019. 2671–2680 (see p. 80)

[26] Brian Curless and Marc Levoy. “A Volumetric Method for Building Complex Models from Range Images” in: *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*. 1996. 303–312 (see p. 48)

[27] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. “ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes” in: *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*. 2017. (see pp. 44, 50)

[28] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. “Depth-Supervised NeRF: Fewer Views and Faster Training for Free” in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022. 12882–12891 (see pp. 48, 61, 71)

[29] Yan Di, Fabian Manhardt, Gu Wang, Xiangyang Ji, Nassir Navab, and Federico Tombari. “SO-Pose: Exploiting Self-Occlusion for Direct 6D Pose Estimation” in: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021. 12396–12405 (see pp. 79, 81, 88, 89)

[30] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale” in: *International Conference on Learning Representations*. 2021. (see p. 21)

[31] Bertram Drost, Markus Ulrich, Paul Bergmann, Philipp Hartinger, and Carsten Steger. “Introducing MVTEC Itodd-a Dataset for 3D Object Recognition in Industry” in: *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*. 2017. 2200–2208 (see pp. 79, 82)

[32] Bertram Drost, Markus Ulrich, Nassir Navab, and Slobodan Ilic. “Model Globally, Match Locally: Efficient and Robust 3D Object Recognition” in: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE 2010. 998–1005 (see p. 81)

[33] Ulrich Eck, Frieder Pankratz, Christian Sandor, Gudrun Klinker, and Hamid Laga. Precise Haptic Device Co-Location for Visuo-Haptic Augmented Reality. *IEEE Transactions on Visualization and Computer Graphics*, **0**: 15, 2015. (see p. 55)

[34] David Eigen, Christian Puhrsch, and Rob Fergus. “Depth Map Prediction from a Single Image using a Multi-Scale Deep Network” in: *Advances in Neural Information Processing Systems*. 2014. 2366–2374 (see pp. 12, 16, 30, 37, 46)

[35] Hao-Shu Fang, Chenxi Wang, Minghao Gou, and Cewu Lu. “GraspNet-1Billion: A Large-Scale Benchmark for General Object Grasping” in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020. 11444–11453 (see pp. 13, 43)

[36] Martin A Fischler and Robert C Bolles. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Communications of the ACM*, **24**: 381–395, 1981. (see p. 81)

[37] Torsten Fließbach. *Elektrodynamik: Lehrbuch zur Theoretischen Physik II*. vol. 2 Springer-Verlag, 2012. (see p. 83)

[38] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. “Plenoxels: Radiance Fields Without Neural Networks” in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022. 5501–5510 (see p. 48)

[39] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. “Deep Ordinal Regression Network for Monocular Depth Estimation” in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018. 2002–2011 (see p. 16)

[40] Stefan Fuchs. “Multipath Interference Compensation in Time-of-Flight Camera Images” in: *2010 20th International Conference on Pattern Recognition*. 2010. 3583–3586 (see p. 49)

[41] Stefan Fuchs. Multipath Interference Compensation in Time-of-Flight Camera Images. *2010 20th International Conference on Pattern Recognition*, 3583–3586, 2010. (see p. 49)

[42] Daoyi Gao, Yitong Li, Patrick Ruhkamp, Iuliia Skobleva, Magdalena Wysocki, HyunJun Jung, Pengyuan Wang, Arturo Guridi, and Benjamin Busam. “Polarimetric Pose Prediction” in: *European Conference on Computer Vision*. Springer 2022. (see pp. 4, 7, 113, 118, 119, 130)

[43] Feng Gao, Jincheng Yu, Hao Shen, Yu Wang, and Huazhong Yang. Attentional Separation-and-Aggregation Network for Self-Supervised Depth-Pse Learning in Dynamic Scenes. *CoRL*, 2020. (see pp. 24, 27)

[44] N Missael Garcia, Ignacio De Erausquin, Christopher Edmiston, and Viktor Gruev. Surface Normal Reconstruction using Circularly Polarized Light. *Optics express*, **23**: 14391–14406, 2015. (see pp. 46, 80)

[45] Ravi Garg, Vijay Kumar Bg, Gustavo Carneiro, and Ian Reid. "Unsupervised CNN for Single View Depth Estimation: Geometry to the Rescue" in: *European Conference on Computer Vision*. Springer 2016. 740–756 (see pp. 16, 47)

[46] Sergio Garrido-Jurado, Rafael Munoz-Salinas, Francisco Jose Madrid-Cuevas, and Manuel Jesus Marin-Jimenez. Automatic Generation and Detection of Highly Reliable Fiducial Markers under Occlusion. *Pattern Recognition*, **47**: 2280–2292, 2014. (see p. 55)

[47] Stefano Gasperini, Patrick Koch, Vinzenz Dallabetta, Nassir Navab, Benjamin Busam, and Federico Tombari. "R4Dyn: Exploring Radar for Self-Supervised Monocular Depth Estimation of Dynamic Scenes" in: *2021 International Conference on 3D Vision (3DV)*. IEEE 2021. 751–760 (see p. 43)

[48] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets Robotics: The KITTI Dataset. *The International Journal of Robotics Research*, **32**: 1231–1237, 2013. (see pp. 13, 16, 19, 28, 30, 47)

[49] Andreas Geiger, Philip Lenz, and Raquel Urtasun. "Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite" in: *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE 2012. 3354–3361 (see p. 43)

[50] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. "Unsupervised Monocular Depth Estimation with Left-Right Consistency" in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017. 270–279 (see pp. 12, 16, 27, 47)

[51] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. "Digging into Self-Supervised Monocular Depth Estimation" in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019. 3828–3838 (see pp. 4, 12, 13, 16, 18, 20, 24, 26, 27, 30–33, 36, 47, 60, 61)

[52] Juan Luis GonzalezBello and Munchurl Kim. Forget about the LiDAR: Self-Supervised Depth Estimators with MED Probability Volumes. *Advances in Neural Information Processing Systems*, **33**: 12626–12637, 2020. (see p. 21)

[53] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. "3D Packing for Self-Supervised Monocular Depth Estimation" in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020. 2485–2494 (see pp. 13, 16, 19, 20, 30–32)

[54] Qi Guo, Iuri Frosio, Orazio Gallo, Todd Zickler, and Jan Kautz. "Tackling 3D ToF Artifacts Through Learning and the FLAT Dataset" in: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018. 368–383 (see pp. 43, 45, 47)

[55] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. "Mask R-CNN" in: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2017. 2961–2969 (see p. 93)

[56] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep Residual Learning for Image Recognition" in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016. 770–778 (see pp. 17, 113)

[57] Yisheng He, Haibin Huang, Haoqiang Fan, Qifeng Chen, and Jian Sun. "FFB6D: A Full Flow Bidirectional Fusion Network for 6D Pose Estimation" in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021. 3003–3013 (see pp. 81, 95, 97, 100, 102, 114)

[58] Yisheng He, Wei Sun, Haibin Huang, Jianran Liu, Haoqiang Fan, and Jian Sun. “PVN3D: A Deep Point-Wise 3D Keypoints Voting Network for 6DoF Pose Estimation” in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020. 11632–11641 (see p. 81)

[59] Stefan Hinterstoisser, Stefan Holzer, Cedric Cagniart, Slobodan Ilic, Kurt Konolige, Nassir Navab, and Vincent Lepetit. “Multimodal Templates for Real-time Detection of Texture-less Objects in Heavily Cluttered Scenes” in: *Proceedings of the IEEE International Conference on Computer Vision*. IEEE 2011. 858–865 (see p. 49)

[60] Stefan Hinterstoisser, Vincent Lepetit, Slobodan Ilic, Stefan Holzer, Gary Bradski, Kurt Konolige, and Nassir Navab. “Model based Training, Detection and Pose Estimation of Texture-less 3D Objects in Heavily Cluttered Scenes” in: *Asian Conference on Computer Vision (ACCV)*. Springer 2012. 548–562 (see pp. 79, 82, 94, 114)

[61] Tomas Hodan, Daniel Barath, and Jiri Matas. “EPOS: Estimating 6D Pose of Objects with Symmetries” in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020. 11703–11712 (see p. 81)

[62] Tomáš Hodan, Pavel Haluza, Štepán Obdržálek, Jiri Matas, Manolis Lourakis, and Xenophon Zabulis. “T-LESS: An RGB-D Dataset for 6D Pose Estimation of Texture-less Objects” in: *IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE 2017. 880–888 (see pp. 79, 82)

[63] Tomáš Hodaň, Jiří Matas, and Štěpán Obdržálek. “On Evaluation of 6D Object Pose Estimation” in: *European Conference on Computer Vision*. Springer 2016. 606–619 (see p. 115)

[64] Hou-Ning Hu, Qi-Zhi Cai, Dequan Wang, Ji Lin, Min Sun, Philipp Krähenbühl, Trevor Darrell, and Fisher Yu. “Joint Monocular 3D Vehicle Detection and Tracking” in: *IEEE International Conference on Computer Vision (ICCV)*. 2019. (see pp. 13, 16)

[65] Hou-Ning Hu, Yung-Hsu Yang, Tobias Fischer, Fisher Yu, Trevor Darrell, and Min Sun. Monocular Quasi-Dense 3D Object Tracking. *ArXiv:2103.07351*, 2021. (see pp. 13, 16)

[66] Yinlin Hu, Pascal Fua, Wei Wang, and Mathieu Salzmann. “Single-Stage 6D Object Pose Estimation” in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020. 2930–2939 (see p. 81)

[67] Yinlin Hu, Joachim Hugonot, Pascal Fua, and Mathieu Salzmann. “Segmentation-Driven 6D Object Pose Estimation” in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019. 3385–3394 (see p. 81)

[68] Manuel Huber, Michael Schlegel, and Gudrun Klinker. “Temporal Calibration in Multisensor Tracking Setups” in: *2009 8th IEEE International Symposium on Mixed and Augmented Reality*. IEEE 2009. 195–196 (see p. 55)

[69] Cong Phuoc Huynh, Antonio Robles-Kelly, and Edwin Hancock. “Shape and Refractive Index Recovery from Single-View Polarisation Images” in: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE 2010. 1229–1236 (see p. 85)

[70] Lam Huynh, Phong Nguyen-Ha, Jiri Matas, Esa Rahtu, and Janne Heikkilä. “Guiding Monocular Depth Estimation using Depth-Attention Volume” in: *European Conference on Computer Vision*. Springer 2020. 581–597 (see p. 17)

[71] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. “FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks” in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017. 2462–2470 (see p. 22)

[72] Md Nazrul Islam, Murat Tahtali, and Mark Pickering. Specular Reflection Detection and Inpainting in Transparent Object through MSPLFI. *Remote Sensing*, **13**: 455, 2021. (see p. 80)

[73] Joel Janai, Fatma Güney, Anurag Ranjan, Michael J. Black, and Andreas Geiger. "Unsupervised Learning of Multi-Frame Optical Flow with Occlusions" in: *European Conference on Computer Vision (ECCV)*. Springer, Cham, Sept. 2018. 713–731 (see p. 17)

[74] Xiaoke Jiang, Donghai Li, Hao Chen, Ye Zheng, Rui Zhao, and Liwei Wu. Uni6D: A Unified CNN Framework without Projection Breakdown for 6D Pose Estimation. *arXiv preprint arXiv:2203.14531*, 2022. (see p. 81)

[75] David Jiménez, Daniel Pizarro, Manuel Mazo, and Sira Palazuelos. Modeling and Correction of Multipath Interference in Time of Flight Cameras. *Image and Vision Computing*, **32**: 1–13, 2014. (see p. 49)

[76] Adrian Johnston and Gustavo Carneiro. "Self-supervised Monocular Trained Depth Estimation using Self-attention and Discrete Disparity Volume" in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020. 4756–4765 (see pp. 17, 21, 22)

[77] HyunJun Jung, Nikolas Brasch, Aleš Leonardis, Nassir Navab, and Benjamin Busam. "Wild ToFu: Improving Range and Quality of Indirect Time-of-Flight Depth with RGB Fusion in Challenging Environments" in: *2021 International Conference on 3D Vision (3DV)*. IEEE 2021. 239–248 (see pp. 43, 45, 49)

[78] HyunJun Jung, Patrick Ruhkamp, Nassir Navab, and Benjamin Busam. Multi-Modal Dataset Acquisition for Photometrically Challenging Object. 2023. (see p. 6)

[79] HyunJun Jung, Patrick Ruhkamp, Guangyao Zhai, Nikolas Brasch, Yitong Li, Yannick Verdie, Jifei Song, Yiren Zhou, Anil Armagan, Slobodan Ilic, et al. "On the Importance of Accurate Geometry Data for Dense 3D Vision Tasks" in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023. (see pp. 6, 49, 121)

[80] HyunJun Jung, Guangyao Zhai, Shun-Cheng Wu, Patrick Ruhkamp, Hannah Schieber, Giulia Rizzoli, Pengyuan Wang, Hongcheng Zhao, Lorenzo Garattoni, Sven Meier, Daniel Roth, Nassir Navab, and Benjamin Busam. "HouseCat6D – A Large-Scale Multi-Modal Category Level 6D Object Perception Dataset with Household Objects in Realistic Scenarios" in: *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*. 2024. (see p. 6)

[81] Achuta Kadambi, Vage Taamazyan, Boxin Shi, and Ramesh Raskar. Depth Sensing using Geometrically Constrained Polarization Normals. *International Journal of Computer Vision (IJCV)*, **125**: 34–51, 2017. (see pp. 46, 80)

[82] Agastya Kalra, Vage Taamazyan, Supreeth Krishna Rao, Kartik Venkataraman, Ramesh Raskar, and Achuta Kadambi. "Deep Polarization Cues for Transparent Object Segmentation" in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020. 8602–8611 (see pp. 46, 78, 80, 84)

[83] Roman Kaskman, Sergey Zakharov, Ivan Shugurov, and Slobodan Ilic. HomebrewedDB: RGB-D Dataset for 6D Pose Estimation of 3D Objects. *International Conference on Computer Vision (ICCV) Workshops*, 2019. (see p. 82)

[84] Wadim Kehl, Fabian Manhardt, Federico Tombari, Slobodan Ilic, and Nassir Navab. "SSD-6D: Making RGB-based 3D Detection and 6D Pose Estimation Great again" in: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2017. 1521–1529 (see p. 81)

[85] Diederik P Kingma and Jimmy Ba. Adam: A Method for Dtochastic Optimization. *arXiv preprint arXiv:1412.6980*, 2014. (see pp. 61, 62, 93, 113)

[86] Xin Kong, Xuemeng Yang, Guangyao Zhai, Xiangrui Zhao, Xianfang Zeng, Mengmeng Wang, Yong Liu, Wanlong Li, and Feng Wen. "Semantic Graph based Place Recognition for 3D Point Clouds" in: *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE 2020. 8216–8223 (see p. 43)

[87] Johannes Kopf, Xuejian Rong, and Jia-Bin Huang. “Robust Consistent Video Depth Estimation” in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021. 1611–1621 (see pp. 24, 43)

[88] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, **25**: 2012. (see p. 17)

[89] Abhijit Kundu, Yin Li, and James M Rehg. “3D-RCNN: Instance-level 3D Object Reconstruction via Render-and-Compare” in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018. 3559–3568 (see pp. 89, 90)

[90] Yann Labbé, Justin Carpentier, Mathieu Aubry, and Josef Sivic. “Cosypose: Consistent Multi-view Multi-object 6D Pose Estimation” in: *European Conference on Computer Vision (ECCV)*. Springer 2020. 574–591 (see p. 81)

[91] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. “Deeper Depth Prediction with Fully Convolutional Residual Networks” in: *2016 Fourth International Conference on 3D Vision (3DV)*. IEEE 2016. 239–248 (see pp. 12, 16, 47)

[92] Sihaeng Lee, Janghyeon Lee, Byungju Kim, Eojindl Yi, and Junmo Kim. “Patch-Wise Attention Network for Monocular Depth Estimation” in: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 35 3 2021. 1873–1881 (see pp. 17, 47)

[93] Chenyang Lei, Xuhua Huang, Mengdi Zhang, Qiong Yan, Wenxiu Sun, and Qifeng Chen. “Polarized Reflection Removal with Perfect Alignment in the Wild” in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020. 1750–1758 (see p. 80)

[94] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. EPnP: An Accurate O(n) Solution to the PnP Problem. *International Journal of Computer Vision*, **81**: 155, 2009. (see p. 81)

[95] Xinghui Li, Kai Han, Shuda Li, and Victor Prisacariu. “Dual-Resolution Correspondence Networks” in: *Conference on Neural Information Processing Systems (NeurIPS)*. 2020. (see p. 21)

[96] Yi Li, Gu Wang, Xiangyang Ji, Yu Xiang, and Dieter Fox. “DeepIM: Deep Iterative Matching for 6D Pose Estimation” in: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018. 683–698 (see p. 81)

[97] Yi Li, Gu Wang, Xiangyang Ji, Yu Xiang, and Dieter Fox. “DeepIM: Deep Iterative Matching for 6D Pose Estimation” in: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018. 683–698 (see pp. 91, 110)

[98] Zhigang Li, Gu Wang, and Xiangyang Ji. “CDPN: Coordinates-based Disentangled Pose Network for Real-time RGB-based 6-DoF Object Pose Estimation” in: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019. 7678–7687 (see pp. 81, 88–90, 108)

[99] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. “BARF: Bundle-Sdjusting Neural Radiance Fields” in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021. 5741–5751 (see p. 48)

[100] Chao Liu, Jinwei Gu, Kihwan Kim, Srinivasa G Narasimhan, and Jan Kautz. “Neural RGB-D Sensing: Depth and Uncertainty from a Video Camera” in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019. 10986–10995 (see pp. 18–20, 31, 32)

[101] Xingyu Liu, Shun Iwase, and Kris M Kitani. “StereOBJ-1M: Large-scale Stereo Image Dataset for 6D Object Pose Estimation” in: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021. 10870–10879 (see pp. 49, 50, 79)

[102] Xingyu Liu, Rico Jonschkowski, Anelia Angelova, and Kurt Konolige. “Keypose: Multi-view 3D Labeling and Keypoint Estimation for Transparent Objects” in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020. 11602–11610 (see pp. 45, 49, 82)

[103] Adrian Lopez-Rodriguez, Benjamin Busam, and Krystian Mikolajczyk. “Project to Adapt: Domain Adaptation for Depth Completion from Noisy and Sparse Sensor Data” in: *Proceedings of the Asian Conference on Computer Vision*. 2020. (see pp. 5, 27, 42, 43)

[104] Keyang Luo, Tao Guan, Lili Ju, Yuesong Wang, Zhuo Chen, and Yawei Luo. “Attention-Aware Multi-View Stereo” in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020. 1590–1599 (see p. 17)

[105] Xuan Luo, Jia-Bin Huang, Richard Szeliski, Kevin Matzen, and Johannes Kopf. Consistent Video Depth Estimation. *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH)*, **39**: 71–1, 2020. (see pp. 13, 17, 18, 24, 47)

[106] Fabian Manhardt, Diego Martin Arroyo, Christian Rupprecht, Benjamin Busam, Tolga Birdal, Nassir Navab, and Federico Tombari. “Explaining the Ambiguity of Object Detection and 6D Pose from Visual Data” in: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019. 6841–6850 (see pp. 79, 81)

[107] Nikolaus Mayer, Eddy Ilg, Philipp Fischer, Caner Hazirbas, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. What makes Good Synthetic Training Data for Learning Disparity and Optical Flow Estimation? *International Journal of Computer Vision*, **126**: 942–960, 2018. (see p. 47)

[108] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. “A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation” in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016. 4040–4048 (see p. 46)

[109] Robert McCraith, Lukas Neumann, Andrew Zisserman, and Andrea Vedaldi. Monocular Depth Estimation with Self-supervised Instance Adaptation. *arXiv preprint arXiv:2004.05821*, 2020. (see p. 30)

[110] S Mahdi H Miangoleh, Sebastian Dille, Long Mai, Sylvain Paris, and Yagiz Aksoy. “Boosting Monocular Depth Estimation Models to High-Resolution via Content-Adaptive Multi-Resolution Merging” in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021. 9685–9694 (see p. 47)

[111] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. *Communications of the ACM*, **65**: 99–106, 2021. (see pp. 48, 61, 62, 72)

[112] Ningkai Mo, Wanshui Gan, Naoto Yokoya, and Shifeng Chen. ES6D: A Computation Efficient and Symmetry-Aware 6D Pose Regression Framework. *arXiv preprint arXiv:2204.01080*, 2022. (see p. 81)

[113] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. ORB-SLAM: A Versatile and Accurate Monocular SLAM System. *IEEE Transactions on Robotics*, **31**: 1147–1163, 2015. (see pp. 13, 16)

[114] Raul Mur-Artal and Juan D Tardós. ORB-SLAM2: An Open-Source SALM System for Monocular, Stereo, and RGB-D Cameras. *IEEE Transactions on Robotics*, **33**: 1255–1262, 2017. (see pp. 13, 16)

[115] Richard A. Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J. Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. “KinectFusion: Real-time Dense Surface Mapping and Tracking” in: *2011 10th IEEE International Symposium on Mixed and Augmented Reality*. 2011. 127–136 (see pp. 13, 16, 48)

[116] Merlin Nimier-David, Delio Vicini, Tizian Zeltner, and Wenzel Jakob. Mitsuba 2: A retargetable forward and inverse renderer. *ACM Transactions on Graphics (TOG)*, **38**: 1–17, 2019. (see p. 112)

[117] Keunhong Park, Arsalan Mousavian, Yu Xiang, and Dieter Fox. "Latentfusion: End-to-end Differentiable Reconstruction and Rendering for Unseen Object Pose Estimation" in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020. 10710–10719 (see p. 89)

[118] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. "NeRFies: Deformable Neural Radiance Fields" in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021. 5865–5874 (see p. 48)

[119] Kiru Park, Timothy Patten, and Markus Vincze. "Pix2pose: Pixel-wise Coordinate Regression of Objects for 6D Pose Estimation" in: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019. 7668–7677 (see p. 81)

[120] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. "Automatic Differentiation in PyTorch" in: *NIPS-W*. 2017. (see p. 61)

[121] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems*, **32**: 2019. (see p. 113)

[122] Vaishakh Patil, Wouter Van Gansbeke, Dengxin Dai, and Luc Van Gool. Don't forget the Past: Recurrent Depth Estimation from Monocular Video. *IEEE Robotics and Automation Letters*, **5**: 6813–6820, 2020. (see pp. 13, 16, 18)

[123] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. "PVNet: Pixel-Wise Voting Network for 6DOF Pose Estimation" in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019. 4561–4570 (see p. 81)

[124] Cody J Phillips, Matthieu Lecce, and Kostas Daniilidis. "Seeing Glassware: From Edge Detection to Pose Estimation and Shape Recovery" in: *Robotics: Science and Systems*. vol. 3 2016. (see p. 82)

[125] Andrea Pilzer, Stephane Lathuiliere, Nicu Sebe, and Elisa Ricci. "Refine and Distill: Exploiting Cycle-Inconsistency and Knowledge Distillation for Unsupervised Monocular Depth Estimation" in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019. 9768–9777 (see p. 26)

[126] Mahdi Rad and Vincent Lepetit. "BB8: A Scalable, Accurate, Robust to Partial Occlusion Method for Predicting the 3D Poses of Challenging Objects without using Depth" in: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2017. 3828–3836 (see p. 81)

[127] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. "Vision Transformers for Dense Prediction" in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021. 12179–12188 (see p. 47)

[128] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **44**: 2022. (see p. 47)

[129] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3D Deep Learning with Pytorch3D. *arXiv preprint arXiv:2007.08501*, 2020. (see p. 113)

[130] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. "Common Objects in 3D: Large-Scale Learning and Evaluation of Real-Life 3D Category Reconstruction" in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021. 10901–10911 (see p. 48)

[131] Barbara Roessle, Jonathan T Barron, Ben Mildenhall, Pratul P Srinivasan, and Matthias Nießner. “Dense Depth Priors for Neural Radiance Fields from Sparse Input Views” in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022. 12892–12901 (see pp. 48, 61, 62, 71, 72)

[132] Patrick Ruhkamp, Daoyi Gao, Hanzhi Chen, Nassir Navab, and Beniamin Busam. “Attention meets Geometry: Geometry guided Spatial-Temporal Attention for Consistent Self-Supervised Monocular Depth Estimation” in: *2021 International Conference on 3D Vision (3DV)*. IEEE 2021. 837–847 (see pp. 6, 43, 60)

[133] Patrick Ruhkamp, Daoyi Gao, Hanzhi Chen, Nassir Navab, and Benjamin Busam. Spatial-Temporal Attention through Self-Supervised Geometric Guidance. 2021. (see p. 6)

[134] Patrick Ruhkamp, Daoyi Gao, HyunJun Jung, Nassir Navab, and Benjamin Busam. Polarimetric Information for Multi-Modal 6D Pose Estimation of Photometrically Challenging Objects with Limited Data. 2023. (see p. 7)

[135] Patrick Ruhkamp, Daoyi Gao, Nassir Navab, and Benjamin Busam. S2P3: Self-Supervised Polarimetric Pose Prediction. 2024. (see p. 7)

[136] Assem Sadek and Boris Chidlovskii. Self-Supervised Attention Learning for Depth and Ego-motion Estimation. *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 10054–10060, 2020. (see p. 17)

[137] Shreeyak Sajjan, Matthew Moore, Mike Pan, Ganesh Nagaraja, Johnny Lee, Andy Zeng, and Shuran Song. “Clear Grasp: 3D Shape Estimation of Transparent Objects for Manipulation” in: *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE 2020. 3634–3642 (see p. 82)

[138] Ashutosh Saxena, Justin Driemeyer, and Andrew Y Ng. Robotic Grasping of Novel Objects using Vision. *The International Journal of Robotics Research*, **27**: 157–173, 2008. (see p. 82)

[139] Daniel Scharstein and Richard Szeliski. A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. *International Journal of Computer Vision*, **47**: 7–42, 2002. (see pp. 4, 43, 45, 49)

[140] Johannes Lutz Schönberger and Jan-Michael Frahm. “Structure-from-Motion Revisited” in: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016. (see p. 55)

[141] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. “Pixelwise View Selection for Unstructured Multi-View Stereo” in: *European Conference on Computer Vision (ECCV)*. 2016. (see p. 55)

[142] Chang Shu, Kun Yu, Zhixiang Duan, and Kuiyuan Yang. “Feature-metric Loss for Self-supervised Learning of Depth and Egomotion” in: *ECCV*. 2020. (see pp. 18, 30)

[143] Ivan Shugurov, Fu Li, Benjamin Busam, and Slobodan Ilic. OSOP: A Multi-Stage One Shot Object Pose Estimation Framework. *arXiv preprint arXiv:2203.15533*, 2022. (see p. 81)

[144] Ivan Shugurov, Sergey Zakharov, and Slobodan Ilic. DPODv2: Dense Correspondence-Based 6 DoF Pose Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. (see p. 81)

[145] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. “Indoor Segmentation and Support Inference from RGBD Images” in: *European Conference on Computer Vision*. Springer 2012. 746–760 (see pp. 43, 45, 49)

[146] William AP Smith, Ravi Ramamoorthi, and Silvia Tozza. Height-from-Polarisation with unknown Lighting or Albedo. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, **41**: 2875–2888, 2018. (see pp. 46, 80)

[147] Juil Sock, Guillermo Garcia-Hernando, Anil Armagan, and Tae-Kyun Kim. "Introducing Pose Consistency and Warp-Alignment for Self-Supervised 6D Object Pose Estimation in Color Images" in: *2020 International Conference on 3D Vision (3DV)*. IEEE 2020. 291–300 (see pp. 105, 114)

[148] Kilho Son, Ming-Yu Liu, and Yuichi Taguchi. "Learning to Remove Multipath Distortions in Time-of-Flight Range Images for a Robotic Arm Setup" in: *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE 2016. 3390–3397 (see p. 45)

[149] Chen Song, Jiaru Song, and Qixing Huang. "Hybridpose: 6D Object Pose Estimation under Hybrid Representations" in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020. 431–440 (see p. 81)

[150] Jaime Spencer, Richard Bowden, and Simon Hadfield. "DeFeat-Net: General Monocular Depth via Simultaneous Unsupervised Representation Learning" in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020. 14402–14413 (see p. 47)

[151] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica Dataset: A Digital Replica of Indoor Spaces. *arXiv preprint arXiv:1906.05797*, 2019. arXiv: 1906.05797 [cs.CV] (see pp. 44, 49, 50)

[152] Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. "A Benchmark for the Evaluation of RGB-D SLAM Systems" in: *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE 2012. 573–580 (see pp. 43, 45, 49)

[153] Yongzhi Su, Yan Di, Guangyao Zhai, Fabian Manhardt, Jason Rambach, Benjamin Busam, Didier Stricker, and Federico Tombari. OPA-3D: Occlusion-Aware Pixel-Wise Aggregation for Monocular 3D Object Detection. *IEEE Robotics and Automation Letters*, 2023. (see p. 43)

[154] Yongzhi Su, Mahdi Saleh, Torben Fetzer, Jason Rambach, Nassir Navab, Benjamin Busam, Didier Stricker, and Federico Tombari. ZebraPose: Coarse to Fine Surface Encoding for 6DoF Object Pose Estimation. *arXiv preprint arXiv:2203.09418*, 2022. (see p. 81)

[155] Cheng Sun, Min Sun, and Hwann-Tzong Chen. "Direct Voxel Grid Optimization: Super-fast Convergence for Radiance Fields Reconstruction" in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022. 5459–5469 (see p. 48)

[156] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-Free Local Feature Matching with Transformers. *CVPR*, 2021. (see pp. 21, 22)

[157] Martin Sundermeyer, Maximilian Durner, En Yen Puang, Zoltan-Csaba Marton, Narunas Vaskevicius, Kai O Arras, and Rudolph Triebel. "Multi-Path Learning for Object Pose Estimation across Domains" in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020. 13916–13925 (see p. 81)

[158] Martin Sundermeyer, Zoltan-Csaba Marton, Maximilian Durner, Manuel Brucker, and Rudolph Triebel. "Implicit 3D Orientation Learning for 6D Object Detection from RGB Images" in: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018. 699–715 (see p. 81)

[159] Shinji Umeyama. Least-Squares Estimation of Transformation Parameters between two Point Patterns. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 376–380, 1991. (see p. 81)

[160] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *Advances in Neural Information Processing Systems*, **30**: 2017. (see pp. 17, 21)

[161] Yannick Verdie, Jifei Song, Barnabe Mas, Benjamin Busam, Ales Leonardis, and Steven McDonagh. “CroMo: Cross-Modal Learning for Monocular Depth Estimation” in: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE 2022. 3937–3947 (see pp. 43, 46, 49, 80, 111, 123)

[162] Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martín-Martín, Cewu Lu, Li Fei-Fei, and Silvio Savarese. “Densefusion: 6D Object Pose Estimation by Iterative Dense Fusion” in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019. 3343–3352 (see p. 81)

[163] Gu Wang, Fabian Manhardt, Xingyu Liu, Xiangyang Ji, and Federico Tombari. Occlusion-Aware Self-Supervised Monocular 6D Object Pose Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. (see pp. 102, 105, 107–109, 114, 115, 117)

[164] Gu Wang, Fabian Manhardt, Jianzhun Shao, Xiangyang Ji, Nassir Navab, and Federico Tombari. “Self6D: Self-supervised Monocular 6D Object Pose Estimation” in: *European Conference on Computer Vision (ECCV)*. Springer 2020. 108–125 (see pp. 105, 114)

[165] Gu Wang, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. “GDR-Net: Geometry-Guided Direct Regression Network for Monocular 6D Object Pose Estimation” in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021. 16611–16621 (see pp. 4, 81, 88–90, 97, 102, 105, 108, 114)

[166] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J Guibas. “Normalized Object Coordinate Space for Category-level 6D Object Pose and Size Estimation” in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019. 2642–2651 (see p. 49)

[167] Pengyuan Wang, HyunJun Jung, Yitong Li, Siyuan Shen, Rahul Parthasarathy Srikanth, Loranzo Garattoni, Sven Meier, Nassir Navab, and Benjamin Busam. “PhoCaL: A Multimodal Dataset for Category-Level Object Pose Estimation with Photometrically Challenging Objects” in: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE 2022. (see pp. 43, 49, 50, 52, 53, 92, 93, 102)

[168] Pengyuan Wang, Fabian Manhardt, Luca Minciullo, Lorenzo Garattoni, Sven Meier, Nassir Navab, and Benjamin Busam. “DemoGrasp: Few-shot Learning for Robotic Grasping with Human Demonstration” in: *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE 2021. 5733–5740 (see pp. 13, 43)

[169] Yang Wang, Yi Yang, Zhenheng Yang, Liang Zhao, Peng Wang, and Wei Xu. “Occlusion Aware Unsupervised Learning of Optical Flow” in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018. 4884–4893 (see p. 17)

[170] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. NeRF−: Neural Radiance Fields without known Camera Parameters. *arXiv preprint arXiv:2102.07064*, 2021. (see p. 48)

[171] Jamie Watson, Oisin Mac Aodha, Victor Prisacariu, Gabriel Brostow, and Michael Firman. “The Temporal Opportunist: Self-Supervised Multi-Frame Monocular Depth” in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021. 1164–1174 (see pp. 13, 14, 18–20, 26, 30–32, 34, 35, 37, 47, 60)

[172] Felix Wimbauer, Nan Yang, Lukas von Stumberg, Niclas Zeller, and Daniel Cremers. “MonoRec: Semi-Supervised Dense Reconstruction in Dynamic Environments from a Single Moving Camera” in: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021. (see p. 18)

[173] Paul Wohlhart and Vincent Lepetit. “Learning Descriptors for Object Recognition and 3D Pose Estimation” in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR*. 2015. 3109–3118 (see p. 81)

[174] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and. PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes. *Robotics: Science and Systems*, 2018. (see pp. 4, 49, 81, 82, 89)

[175] Jianxiong Xiao, Andrew Owens, and Antonio Torralba. "Sun3d: A database of Big Spaces Reconstructed using SfM and Object Labels" in: *Proceedings of the IEEE International Conference on Computer Vision*. 2013. 1625–1632 (see pp. 43, 45)

[176] Junyuan Xie, Ross Girshick, and Ali Farhadi. "Deep3d: Fully Automatic 2D-to-3D Video Conversion with Deep Convolutional Neural Networks" in: *European Conference on Computer Vision*. Springer 2016. 842–857 (see pp. 16, 47)

[177] Yiheng Xie, Towaki Takikawa, Shunsuke Saito, Or Litany, Shiqin Yan, Numair Khan, Federico Tombari, James Tompkin, Vincent Sitzmann, and Srinath Sridhar. Neural Fields in Visual Computing and Beyond. *Computer Graphics Forum*, 2022. (see p. 48)

[178] Guanglei Yang, Hao Tang, Mingli Ding, Nicu Sebe, and Elisa Ricci. Transformers Solve the Limited Receptive Field for Monocular Depth Prediction. *arXiv preprint arXiv:2103.12091*, 2021. (see p. 17)

[179] Nan Yang, Rui Wang, Jorg Stuckler, and Daniel Cremers. "Deep Virtual Stereo Odometry: Leveraging Deep Depth Prediction for Monocular Direct Sparse Odometry" in: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018. 817–833 (see p. 13)

[180] Zhenheng Yang, Peng Wang, Yang Wang, Wei Xu, and Ram Nevatia. "LEGO: Learning Edge with Geometry all at once by Watching Videos" in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018. 225–234 (see p. 47)

[181] Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Long Mai, Simon Chen, and Chunhua Shen. "Learning to Recover 3D Scene Shape from a Single Image" in: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (CVPR)*. 2021. (see p. 47)

[182] Zhichao Yin and Jianping Shi. "GeoNet: Unsupervised Learning of Dense Depth, Optical Flow and Camera Pose" in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018. 1983–1992 (see p. 17)

[183] He Yisheng, Wang Yao, Fan Haoqiang, Chen Qifeng, and Sun Jian. FS6D: Few-Shot 6D Pose Estimation of Novel Objects. *CVPR*, 2022. (see p. 81)

[184] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. "Dilated Residual Networks" in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017. 472–480 (see pp. 20, 21)

[185] Ye Yu, Dizhong Zhu, and William AP Smith. "Shape-from-Polarisation: A Nonlinear Least Squares Approach" in: *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*. 2017. 2969–2976 (see pp. 46, 80)

[186] Sergey Zakharov, Ivan Shugurov, and Slobodan Ilic. "DPOD: 6D Pose Object Detector and Refiner" in: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019. 1941–1950 (see p. 81)

[187] Andy Zeng, Shuran Song, Matthias Nießner, Matthew Fisher, Jianxiong Xiao, and Thomas Funkhouser. "3DMatch: Learning Local Geometric Descriptors from RGB-D Reconstructions" in: *CVPR*. 2017. (see p. 48)

[188] Guangyao Zhai, Dianye Huang, Shun-Cheng Wu, HyunJun Jung, Yan Di, Fabian Manhardt, Federico Tombari, Nassir Navab, and Benjamin Busam. "MonoGraspNet: 6-DoF Grasping with a Single RGB Image" in: *IEEE International Conference on Robotics and Automation*. IEEE 2023. (see pp. 13, 43)

[189] H. Zhan, C. S. Weerasekera, J. -W. Bian, and I. Reid. "Visual Odometry Revisited: What Should Be Learnt?" in: *2020 IEEE International Conference on Robotics and Automation (ICRA)*. 2020. 4203–4210 (see p. 24)

[190] Haokui Zhang, Chunhua Shen, Ying Li, Yuanzhouhan Cao, Yu Liu, and Youliang Yan. “Exploiting Temporal Consistency for Real-time Video Depth Estimation” in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019. 1725–1734 (see pp. 16, 17)

[191] Yinda Zhang, Sameh Khamis, Christoph Rhemann, Julien Valentin, Adarsh Kowdle, Vladimir Tankovich, Michael Schoenberg, Shahram Izadi, Thomas Funkhouser, and Sean Fanello *ActiveStereoNet: End-to-End Self-Supervised Learning for Active Stereo Systems* 2018 arXiv: 1807.06009 [cs.CV] (see p. 49)

[192] Hengshuang Zhao, Jiaya Jia, and Vladlen Koltun. “Exploring Self-Attention for Image Recognition” in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020. 10076–10085 (see p. 17)

[193] Wang Zhao, Shaohui Liu, Yezhi Shu, and Yong-Jin Liu. “Towards Better Generalization: Joint Depth-Pose Learning without PoseNet” in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020. 9151–9161 (see pp. 17, 24, 30)

[194] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. “Unsupervised Learning of Depth and Ego-Motion from Video” in: *CVPR*. 2017. (see p. 16)

[195] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. “On the Continuity of Rotation Representations in Neural Networks” in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019. 5745–5753 (see pp. 81, 88, 89, 108)

[196] Dizhong Zhu and William AP Smith. “Depth from a Polarisation + RGB Stereo Pair” in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019. 7586–7595 (see p. 80)

[197] Shihao Zou, Xinxin Zuo, Yiming Qian, Sen Wang, Chi Xu, Minglun Gong, and Li Cheng. “3D Human Shape Reconstruction from a Polarization Image” in: *European Conference on Computer Vision*. Springer 2020. 351–368 (see p. 87)