Technische Universität München
Fakultät für Luftfahrt, Raumfahrt und Geodäsie
Professur für Signalverarbeitung in der Erdbeobachtung

Deep Learning for Matching High-Resolution SAR and Optical Imagery

Lloyd Haydn Hughes

Vollständiger Abdruck der von der Fakultät für Luftfahrt,
Raumfahrt und Geodäsie der Technische Universität München
zur Erlangung des akademischen Grades eines
Doktor-Ingenieurs (Dr.-Ing.)
genehmigten Dissertation.

Vorsitzender: Prof. Dr. Martin Werner

Prüfer der Dissertation:

1. Priv.-Doz. Dr.-Ing. habil. Michael Schmitt
2. Prof. Dr.-Ing. habil. Xaioxiang Zhu
3. Prof. Dr. Florence Tupin

Die Dissertation wurde am 09.07.2020 bei der Technische Universität München eingereicht und durch die Fakultät für Luftfahrt, Raumfahrt und Geodäsie am 19.08.2020 angenommen.

"*No one can give you any answers. There aren't any. You have to discover for yourself – you must learn to navigate the mystery.*"

Bill Hicks

Abstract

The rise of the New Space era has led to a rapid increase in the availability and accessibility of high-resolution Earth observation data, and with it, the need for efficient mechanisms to extract insights. To this end, data fusion has become an indispensable tool for the large-scale exploitation of ever-growing remote sensing data archives. A particularly important case is the joint exploitation of highly complementary data captured by synthetic aperture radar (SAR) and optical sensor modalities.

However, the first step in image-based data fusion endeavours is the determination of correspondences and the subsequent alignment of the various data sources. In this context, the main objective of this thesis is to expand the current research on the application of deep learning to the problem of SAR-optical image matching. The aim is to develop a fully automatic deep learning-based SAR-optical matching pipeline capable of matching high-resolution imagery.

The objective is achieved through the investigation and development of deep learning-based solutions to the various sub-problems within the realm of SAR-optical matching. This thesis focuses on the core sub-problem of accurately determining corresponding points across these modalities, and proposes numerous deep matching architectures to address this problem in different ways. These supervised networks are found to outperform existing SAR-optical matching approaches across the board, with one such approach significantly raising the bar for high-resolution SAR-optical matching. Furthermore, approaches to matching under scarce data conditions are also investigated. However, the complexity of these formulations and the numerous additional deep learning related challenges mean that these approaches require more research to be comparable to existing approaches' matching accuracy and precision.

Additionally, the sub-problems of SAR-optical feature detection and outlier identification are addressed from a multi-modal deep learning perspective. The proposed solution to the multi-modal feature detection problem is found to significantly increase the accuracy of the correspondence network – in comparison to the commonly-used single modality feature detection approach. Furthermore, the use of a convolutional neural network for outlier identification is found to further increase the overall matching accuracy.

Finally, these various developments are combined to form a comprehensive, deep learning-based framework for matching SAR and optical imagery. This matching pipeline is evaluated on large-scale high-resolution test imagery, and is found to provide spatially diverse correspondences with an accuracy and precision suitable for their use in other data fusion endeavours.

The contributions of this thesis are described in detail in the six peer-reviewed papers, which comprise the body of this thesis. In summary, the results highlight both the progress made and the remaining challenges in the realm of SAR-optical matching, and lay the groundwork for further development of generalizable solutions to SAR-optical image matching based on deep learning methods.

Zusammenfassung

Der Anbruch eines neuen Zeitalters der Raumfahrt hat zu einem rasanten Anstieg der Verfügbarkeit und Zugänglichkeit zu hochauflösenden Erdbeobachtungsdaten und damit zu einem gestiegenen Bedarf für effizienter Mechanismen zur Gewinnung von Erkenntnissen geführt. Zu diesem Zweck ist die Datenfusion zu einem unverzichtbaren Werkzeug für die umfassende Nutzung ständig wachsender Fernerkundungsdatenarchive geworden. Ein besonders wichtiger Fall ist die gemeinsame Nutzung hoch komplementärer Daten, die mit Radar mit synthetischer Apertur (SAR) und optischen Sensormodalitäten erfasst werden.

Der erste Schritt bei bildbasierten Datenfusionsbemühungen ist jedoch die Bestimmung von Entsprechungen und die anschließende Ausrichtung der verschiedenen Datenquellen. In diesem Zusammenhang besteht das Hauptziel dieser Arbeit darin, die aktuelle Forschung zur Anwendung von Deep Learning auf das Problem der SAR-optischen Bildanpassung zu erweitern. Ziel ist es, eine vollautomatische, auf Deep Learning basierende SAR-optische Matching-Pipeline zu entwickeln, mit der hochauflösende Bilder abgeglichen werden können.

Das Ziel wird durch die Untersuchung und Entwicklung von Deep-Learning-basierten Lösungen für die verschiedenen Teilprobleme im Bereich der SAR-optischen Anpassung erreicht. Diese Arbeit konzentriert sich auf das Unterproblem der genauen Bestimmung entsprechender Punkte über diese Modalitäten hinweg und schlägt zahlreiche Deep-Matching-Architekturen vor, um dieses Problem auf unterschiedliche Weise anzugehen. Es wurde festgestellt, dass diese überwachten Netzwerke bestehende SAR-optische Anpassungsansätze auf breiter Front übertreffen, wobei ein solcher Ansatz die Messlatte für hochauflösende SAR-optische Anpassungen erheblich höher legt. Darüber hinaus werden Ansätze zum Matching unter knappen Datenbedingungen untersucht. Die Komplexität dieser Formulierungen und die zahlreichen zusätzlichen Herausforderungen im Zusammenhang mit Deep Learning führen jedoch dazu, dass diese Ansätze weitere Forschungsarbeit erfordert, um mit der Genauigkeit und Präzision bestehender Ansätze vergleichbar zu sein.

Zusätzlich werden die Unterprobleme der SAR-optischen Merkmalserkennung und der Ausreißeridentifikation aus einer multimodalen Deep-Learning-Perspektive angesprochen. Es wurde festgestellt, dass die vorgeschlagene Lösung für das Problem der Erkennung multimodaler Merkmale die Genauigkeit des Korrespondenznetzwerks erheblich erhöht – im Vergleich zu dem üblicherweise verwendeten Ansatz zur Erkennung einzelner Merkmale. Darüber hinaus wurde festgestellt, dass die Verwendung eines Faltungs-Neuronalen Netzwerks zur Identifizierung von Ausreißern die Gesamtanpassungsgenauigkeit weiter erhöht.

Schließlich werden diese verschiedenen Entwicklungen kombiniert, um ein umfassendes, auf Deep Learning basierendes Framework für die Anpassung von SAR- und optischen Bildern zu bilden. Diese Matching-Pipeline wird anhand hochauflösender Testbilder in großem Maßstab ausgewertet und liefert räumlich unterschiedliche Entsprechungen mit

einer Genauigkeit und Präzision, die für ihre Verwendung in anderen Datenfusionsbemühungen geeignet sind.

Die Beiträge dieser Arbeit werden in den sechs von Experten begutachteten Arbeiten, die den Hauptteil dieser Arbeit bilden, ausführlich beschrieben. Zusammenfassend heben die Ergebnisse sowohl die erzielten Fortschritte als auch die verbleibenden Herausforderungen im Bereich der SAR-optischen Anpassung hervor und bilden die Grundlage für die Weiterentwicklung verallgemeinerbarer Lösungen für die SAR-optische Bildanpassung auf der Grundlage von Deep-Learning-Methoden.

Contents

List of Figures

List of Tables

List of Abbreviations

AE	Auto-Encoder
BCE	Binary Cross Entropy
CNNs	Convolutional Neural Networks
DEM	Digital Elevation Model
DSM	Digital Surface Model
DoG	Difference of Gaussians
EM	Electro-Magnetic
FN	False Negative
FPR	False Positive Rate
FP	False Positive
GAN	Generative Adversarial Network
GCP	Ground Control Point
GEE	Google Earth Engine
GRD	Ground Range Detected
HOPC	Histogram Of Phase Congruency
IW	Interferometric Wide
MI	Mutual Information
MSE	Mean Squared Error
NCC	Normalized Cross Correlation
NMS	Non-Maximal Suppression
ORN	Outlier Reduction Network
PSiam	Pseudo-Siamese
RANSAC	Random Sample Consensus
RIFT	Radiation Invariant Feature Transform
ReLU	Rectified Linear Unit
SAR	Synthetic Apature Radar
SIFT	Scale Invariant Feature Transform
VGG16	Visual Geometry Group

1. Introduction

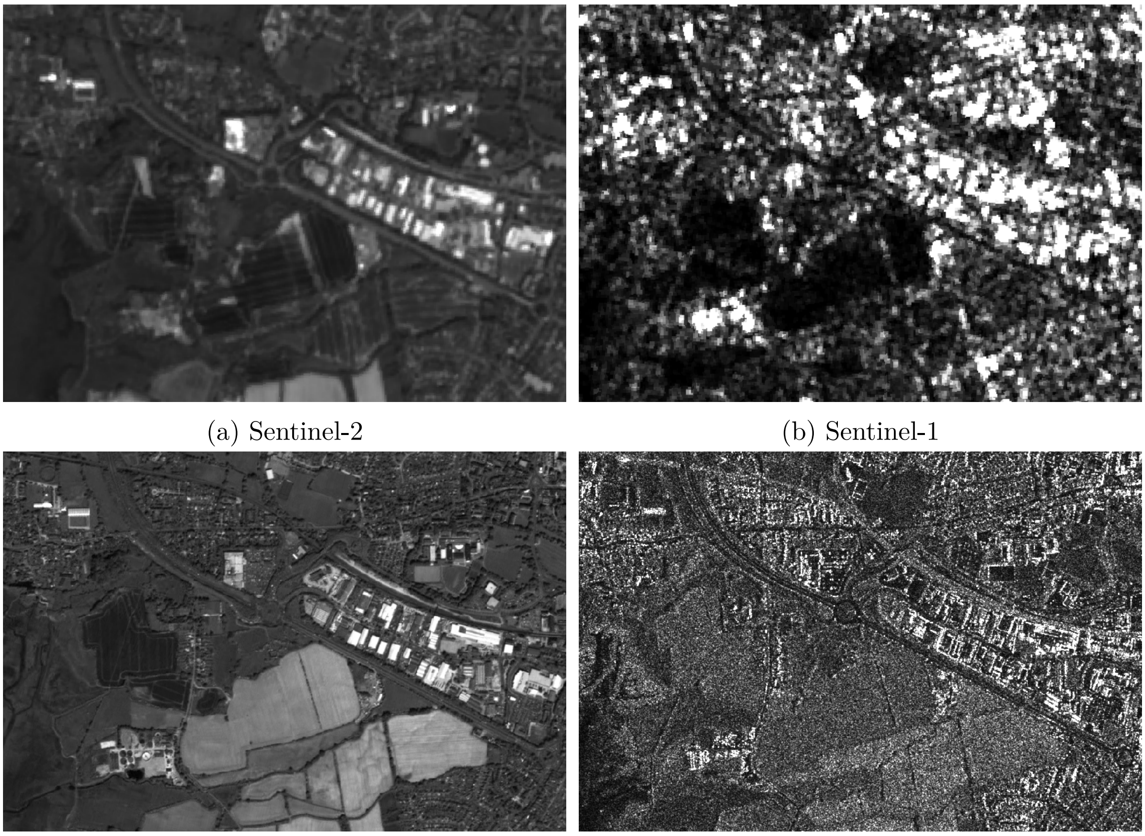
1.1 Motivation

In recent years technological advancements in sensor systems, the miniaturisation of satellites and ride-share based launch solutions have significantly reduced the barrier of entry to space and thus led us into the New Space era (Butler, 2014; Denis et al., 2017). The birth of this era has led to a rapid increase in the number of space-borne Earth observation missions being founded at both a national and international level, across multiple sectors. This in turn has fuelled an exponential growth in the amount and diversity of available Earth observation data. Furthermore, it has brought on the need for the development of advanced and efficient means to exploit this vast data in order to tackle some of humanities most pressing challenges, as outlined by the United Nations in the Sustainable Development Goals (United Nations, 2015). For this reason data fusion has become a key topic within the field of remote sensing, as it enables maximal utilisation of information generated by past, present and future Earth observation missions (Gamba, 2014; Zhang, 2010).

Despite algorithmic advancements and the importance of the data fusion within remote sensing; data fusion has largely been constrained to application across similar modalities, in the case of high-resolution fusion, or low and medium resolution, in the case of multi-modal fusion. The core reason behind this lies within the fundamental need for being able to determine corresponding points, and perform subsequent alignment of data sources prior to being able to embark on data fusion endeavours. Within the realm of image based data, as commonly dealt with in remote sensing, these correspondences are obtained via the process of image matching, whereby common points are co-located across a set of images. This process is largely considered solved, from an operational standpoint, when dealing with data from the same or similar modalities, such as optical and infrared imagery. However, this is not the case when looking to determine correspondences in high-resolution and multi-modal settings.

Synthetic aperture radar (SAR) and optical imagery constitute two of the most widely used and available data sources within remote sensing. This, combined with their vastly different yet highly complementary characteristics, makes SAR and optical data fusion, and thus matching, a highly warranted pursuit (Schmitt et al., 2017; Tupin, 2010).

The imaging concepts employed by SAR and optical sensors are vastly different, which makes finding common points and defining generic similarity metrics a challenging task. Synthetic aperture radar relies on a side-looking acquisition, and range-Doppler geometry to capture the physical properties of a scene, such as moisture content and surface roughness. This leads to geometric image distortions such as layover, foreshortening and radar shadow which have no analogous concepts in the optical domain. On



(a) Sentinel-2

(b) Sentinel-1

(c) PRISM

(d) TerraSAR-X

FIGURE 1.1: Example of optical (a,c) and SAR (b,d) imagery of the same scene at different resolutions. Images (a) and (b) have a ground sampling distance of 10m and are considered to be medium resolution, while (c) and (d) have a ground sampling distance of 2.5m and are considered to be high-resolution, as defined by (Thenkabail, 2018).

the other hand, optical imagery is formed using a projective acquisition geometry and is often captured at near-nadir angles, thus reducing the geometric distortion of the scene. However, unlike SAR, optical imagery captures the chemical characteristics of a scene, but this acquisition can only take place under cloud-free and daylight conditions and thus introduces illumination and shadow inconsistencies across the scene.

These factors combined with the different wavelengths captured by each modality lead to significantly different geometric and radiometric properties between SAR and optical imagery. Thus an object which is visible in SAR imagery, may appear with a completely different nature in the optical image, if it appears at all. Furthermore, SAR imagery suffers from a deterministic, multiplicative noise, called *speckle*, which further distorts the interpretability of features within the image. These factors, specifically those relating to geometric differences, become more even more pronounced as the resolution of the imagery is increased, this can clearly be seen in Figure 1.1.

By this very nature, the determination of corresponding points between high-resolution SAR and optical imagery, as defined by Thenkabail (2018), is a multi-facetted and deeply complex task, even for human experts. While several approaches to SAR-optical matching have been developed over the years, the majority of these approaches rely on intensity-based methods (Chen et al., 2003; Siddique et al., 2012; Suri & Reinartz, 2010) or hand-crafted features (Ma et al., 2017; Xiang et al., 2018; Ye et al., 2016). Intensity-based methods rely directly on pixel intensity values and are thus sensitive to speckle and the radiometric differences between SAR and optical imagery. Feature-based methods, on the other hand, rely on hand-crafted feature detectors to detect and describe similar points across the modalities. Although these approaches perform better than purely intensity-based methods, they still lack robustness to speckle, and more so to geometric differences between the scenes. More recently, a number of deep learning based approaches to single modality remote sensing image matching have been proposed, along with a few approaches specifically tailored towards SAR-optical matching (Bürgmann et al., 2019; Hoffmann et al., 2019; Merkle, Luo, et al., 2017; Mou et al., 2017). Due to the ability of deep neural networks to learn complex representations of data, these matching approaches have shown promising results and better robustness to speckle and radiometric differences between SAR and optical imagery. However, even in the case of deep learning these approaches have been limited to medium resolution imagery, or flat rural areas where geometric differences between the modalities are still constrained. Furthermore, these approaches still largely rely on manually selected features, or features generated from expensive, auxiliary processes.

Thus the development of a generic approach to high-resolution SAR-optical matching, which is not constrained to hand-crafted features, or specific geometric conditions, is still an open problem and forms the basis for the topic of this thesis.

1.2 Objectives

This thesis aims to expand current research on the application of deep learning to SAR-optical image matching, and the sub-problems involved within this domain. To a larger degree, this thesis aims to transfer knowledge from the domains of deep learning and image matching into the domain of remote sensing, by providing a toolbox of solutions to the various challenges and formulations of the SAR-optical matching problem.

The main objective of this thesis is the development of a novel, fully automatic deep learning-based SAR-optical matching pipeline capable of matching high-resolution SAR and optical imagery. Such a pipeline would handle all aspects of the matching process; from multi-modal feature recommendation, to the determination of SAR-optical correspondences and finally, the removal of outliers. In order to achieve this, numerous sub-objectives were formed and investigated, these can be summarized as follows:

- **Deep learning-based approaches to matching**
  The development of various deep matching architectures which enable the matching of high-resolution SAR and optical imagery.

- **Matching under scarce data**
  The development of deep learning-based techniques, and semi-supervised methods to enable matching and improve robustness of existing networks when large-scale labelled training datasets are not available.

- **Large-scale dataset creation**
  The creation of large-scale SAR-optical correspondence datasets which are suitable for training deep neural networks in a supervised manner.
- **Auxiliary tasks for end-to-end matching**
  The investigation and implementation of a multi-modal feature detection mechanism. As well as the development of an outlier identification mechanism which does not rely on auxiliary feature transfer models.

1.3 Thesis Structure

The structure of the thesis is as follows:

Chapter 2 introduces fundamental concepts related to SAR and optical imaging modalities, and image matching. It further provides a review of existing approaches and the state-of-the-art in SAR-optical matching in Section 2.3. Chapter 3 describes the key contributions of this thesis, to which the related publications are contained within Chapter 4. Chapter 5 provides a unified evaluation of the proposed approaches to SAR-optical matching. Finally, Chapter 6 provides a discussion of the findings and outlines future directions for research into SAR-optical matching.

2. Theoretical Background

This chapter introduces the reader to the foundational concepts which are required for understanding the challenges involved within SAR-optical matching, and contributions of this thesis. Firstly, the chapter addresses the basic concepts of SAR and optical sensors and the differences between their image characteristics. This is followed by a general overview of general image matching concepts which form the basis of prior work, and from which the key ideas for addressing SAR-optical matching have been built upon. Finally, this chapter provides a review of previous work in SAR-optical image matching and an analysis of the current the state-of-the-art.

2.1 Spacebourne Optical and Synthetic Aperture Radar Imagery

The taxonomy of Earth observation sensors can be split up in many ways based on the operational frequency or bandwidth of the sensor. However, for the case of this thesis, it is most useful to divide the taxonomy by the acquisition geometry; into projective, and doppler geometries. Within these classes, focus is drawn to the two main subclasses, namely optical and synthetic aperture radar (SAR) imagery. These two modalities constitute the most important modalities within space-borne remote sensing as they capture vastly different characteristics of a scene. Still, each does so with a unique set of benefits and drawbacks.

Through the use of data fusion, the drawbacks of one modality can be augmented by the benefits of the other, and thus a more complete picture of the scene can be created (Schmitt et al., 2016). However, this highly complementary nature is also what makes matching SAR and optical imagery a deeply complex problem. In this section, the fundamental principles of optical and SAR sensors are presented. These are relevant for the reader to gain an understanding of the differences between the modalities. Furthermore, the different characteristics of these modalities are explored, and finally, the key aspects which need to be addressed to match SAR and optical imagery successfully are outlined.

As this section is designed to provide the reader with an overview of the fundamental concepts of SAR and optical imagery relevant to the image matching problem, the reader is referred to the following resources for an in-depth review of SAR (Cumming et al., 2005), optical (Prasad et al., 2011) and general remote sensing (Lillesand et al., 2015) sensors.

2.1.1 Fundamentals of Optical and SAR Imaging

Both SAR and optical sensors image a scene through the use of the electromagnetic (EM) spectrum. However, their imaging mechanism and the information which can be derived from the EM-spectrum at these various frequencies are inherently different (Lillesand et al., 2015).

Active and Passive Sensors

Optical sensors are passive imaging devices, which means they take advantage of existing sources of illumination and the natural reflective properties of objects to form images. In the case of optical Earth observation imagery, the sun is used as a global illumination source. Thus optical image sensors tend to operate on a sun-synchronous orbit to ensure the scene is well illuminated during acquisition. On the other hand, SAR uses the principle of active sensing, whereby the sensor acts as both an illumination source as well as the acquisition unit. A scene is thus imaged by alternating between emitting bursts of EM-radiation and measuring the strength and time delay of the reflected signal. The proportion of the signal which is reflected towards the sensor by an object is known as *backscatter*.

The Electromagnetic Spectrum

Given that objects interact (reflect and absorb) different parts of the EM-spectrum in a unique manner, several realizations of optical and SAR sensors exist, each designed for a specific purpose and to take advantage of the properties observable within a particular part of the EM-spectrum. Optical sensors detect very high-frequency radiation within the visible to thermal infrared section of the EM-spectrum. In contrast, SAR sensors operate at much lower frequencies within the microwave section of the spectrum. It is this use of lower frequencies which allow SAR sensors to be mostly independent of atmospheric conditions, and thus they can acquire imagery through thick clouds, or smoke. The same, however, is not true for optical sensors which require clear atmospheric conditions to observe ground level reflections within the visible light and near-infrared spectrum.

In large, optical sensors can be classified, by the number of spectral bands which they image, into hyperspectral, multispectral, and panchromatic sensors. Whereby hyperspectral sensors image the full optical subsection of the EM-spectrum into hundreds of narrow spectral bands, while panchromatic sensors image the visible light spectrum using a single, wide spectral band. Similarly, SAR sensors can also be classified based on their operational bandwidth into C, L and X-band, with L-band being the lowest frequency and thus the highest level of vegetation and soil penetration. However, due to the relationship between wave-length and spatial resolution present in SAR imagery, (Cumming et al., 2005), X-band imagery exhibits the highest spatial resolution and thus provides the detailed information about surface structure. In optical imagery, a similar trade-off exists, except it is between spatial resolution and spectral resolution. Due to data storage and throughput constraints hyperspectral and multispectral images tend to have a lower resolution than panchromatic imagery. The operational frequency and bandwidth of each of these sensors within the EM-spectrum is further described by Figure 2.1.

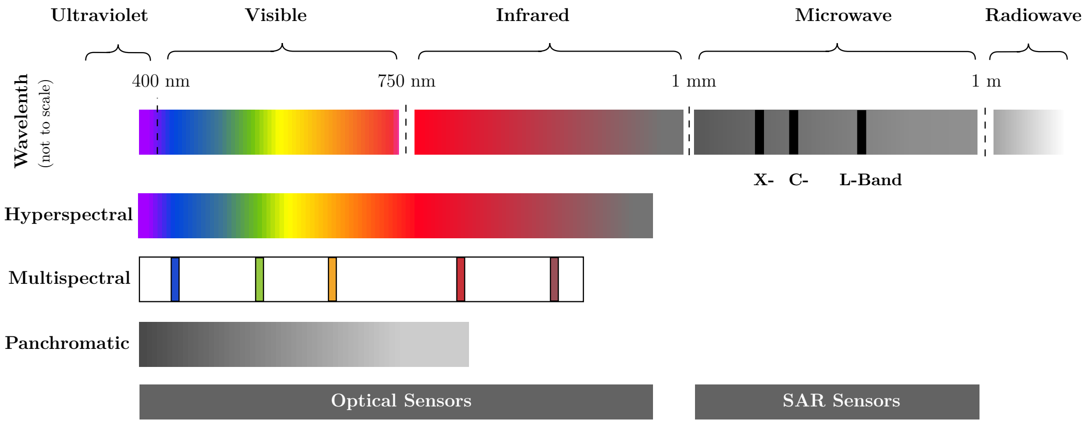


FIGURE 2.1: The electromagnetic spectrum with the position and bandwidth of various classes of optical and SAR sensors depicted. As the intensity of the panchromatic imagery is a function of multiple wavelengths, the depiction of blue as black and infrared as white is merely for effect and not a true mapping of the colour space.

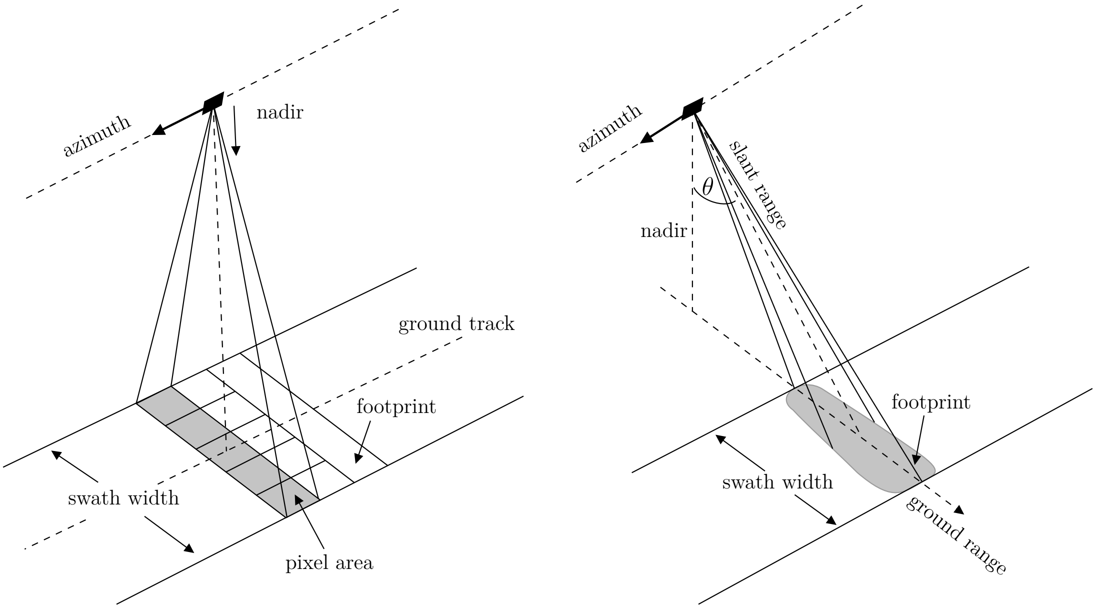
Image Acquisition Geometry

Apart from their illumination and operating spectrum differences, optical and SAR sensors rely on vastly different imaging techniques and acquisition geometries. Optical sensors make use of a linear array of photosensitive detectors which simultaneously record the reflection of light from the Earth's surface. Thus each image pixel directly represents the accumulation of reflected radiation for a specific area on the ground. In optical imaging, the sensor array is often positioned such that the acquisition occurs from a near-nadir (downward looking) perspective, with the image being formed in a line by line along the azimuth (flight direction) as the satellite orbits over a scene (Girard et al., 2003).

In contrast to this, SAR sensors utilize a single antenna to emit EM-signals and measure the corresponding magnitude, range and Doppler-shift of these signals in the backscatter. As only a single detector element exists, SAR is based around the concept of synthesizing an aperture in the azimuth direction in order to form a 2-dimensional image (Cumming et al., 2005; Orth, 2018). Due to this imaging concept, SAR sensors are required to be side looking in order to prevent ambiguities in the image formation process. The angle between the SAR antenna and an object on the ground is known as the *incidence angle*, θ, and it plays a large role in the geometric distortions which are present in SAR imagery, as will be discussed in the next section. Due to this imaging concept, the SAR image is formed along the slant range, where each pixel represents the accumulation of backscatter from points which are the same distance away from the sensor. The acquisition geometry of optical and SAR sensors is depicted in Figure 2.2.

Image Geo-localization

An important aspect of remote sensing imagery which separates it from other types of imagery is the absolute geo-localization of the data. This refers to the fact that each pixel in an optical or SAR image can be related back to a specific area on the Earth, with the area being directly related to the spatial resolution of the sensor. For



(a) Optical Sensor

(b) SAR Sensor

Figure 2.2: Illustration of (a) optical and (b) SAR image acquisition geometry. The optical sensor has a nadir geometry with a rectangular footprint, and fixed area pixel, while the SAR sensor has a side-looking geometry with a non-regular footprint which is forms the aperture as the sensor progresses along the azimuth. SAR pixels are formed along the slant range, based on the time-of-flight of the signal.

this reason satellites continuously monitor their *state* (attitude, position and velocity) relative to the Earth. Using this state, an Earth model and sensors specific image formation models it is possible to perform the geo-localization process for both optical and SAR imagery.

For optical sensors, the image formation process of a single line in the image can be mathematically modelled by a set of co-linearity equations which form a perspective projection model. This same model is used for full-frame cameras in conventional computer vision applications. The perspective projection is then applied to each line in the image to relate 3-dimensional ground coordinates, to pixel coordinates within the sensors image frame, thereby geo-localizing the image (Girard et al., 2003). A similar process can be applied to SAR imagery by modelling the projection of Earth coordinates to slant-range image coordinates using a set of range-Doppler equations (Cumming et al., 2005). The 3-dimensional ground coordinates used in the geo-localization process for both sensors usually take the form of measured ground control points (GCPs) or are derived from existing surface models. Thus the accuracy of these points directly affects the geo-localization accuracy of the imagery.

Furthermore, the geo-localization process in both modalities relies on knowledge of the sensor state. However, the estimation of this state is fraught with errors and uncertainties. Due to the nature of the optical imaging process, these state errors have a significant effect on the absolute geo-localization accuracy. This is because small angular errors in the satellite attitude propagate into large offset errors on the ground. Thus for high-resolution optical sensors, these inaccuracies can lead to geo-localization errors of tens of meters (Merkle, 2018). In contrast, the effect of state inaccuracies on SAR geo-localization is significantly lower as they are primarily compensated for in the SAR sensor model and signal-processing procedures. Thus high-resolution SAR sensors can produce geo-localized imagery with sub-meter accuracy, and in some situations centimeter accuracy (Eineder et al., 2010).

2.1.2 Geometric and Radiometric Characteristics

Due to the previously discussed differences between the SAR and optical image acquisition process, the character of the resultant imagery is vastly different, as seen in Figure 1.1. While optical imagery is easily interpretable by humans, due its imaging concept being similar to that of the human eye, SAR imagery is difficult for humans to interpret without the use of auxiliary information or expert knowledge (Schulz et al., 2009).

In this section the key characteristics of each modality are described and compared to provide the reader with an understanding of the considerations which need to be held in mind when developing and evaluating SAR-optical matching approaches.

Geometric Characteristics

Due to their distinct imaging concepts, the geometric characteristics of a scene appear differently in optical and SAR imagery. When imaging flat terrain these geometric differences are negligible. However, if the scene contains objects with a height above ground the geometric distortions become significant, especially in the case of high-resolution imagery.

In the case of optical imaging, objects perpendicular to the azimuth get projected away from the sensor in the image plane. However, due to the near-nadir acquisition geometry used in optical sensors, these distortions remain moderate throughout the scene. On the contrary, the time-of-flight based imaging principle used in SAR imagery leads to above ground objects being projected towards the sensor in the image plane. To further clarify these differences in projection, the case of imaging a building with SAR and optical sensors is presented in Figure 2.3.

In Figure 2.3 the distortion of the roof (towards and away from the sensor) can be clearly seen in the SAR and optical image planes. Furthermore, the reason why distortions are negligible for ground level objects is also visible. What is not visible from the figure are the other types of geometric distortions which are present in SAR imagery, and which have no analogous concept in optical imagery. These differences can further be seen in the real-world example imagery depicted in Figure 2.4.

Layover, foreshortening and shadow are three additional types of geometric distortion which occur when imaging above ground objects using SAR, and which have a significant impact on the appearance of the resultant imagery (Curlander, 1982). Layover occurs when the slope of the object is greater than the incidence angle, θ, thus causing mixing of the ground and object backscatter and an inversion of the object in the image. This distortion is particularly common in urban and mountainous regions, and is one of the biggest challenges for SAR-optical matching. Foreshortening refers to the absolute distance between two points being shortened upon projection to the image

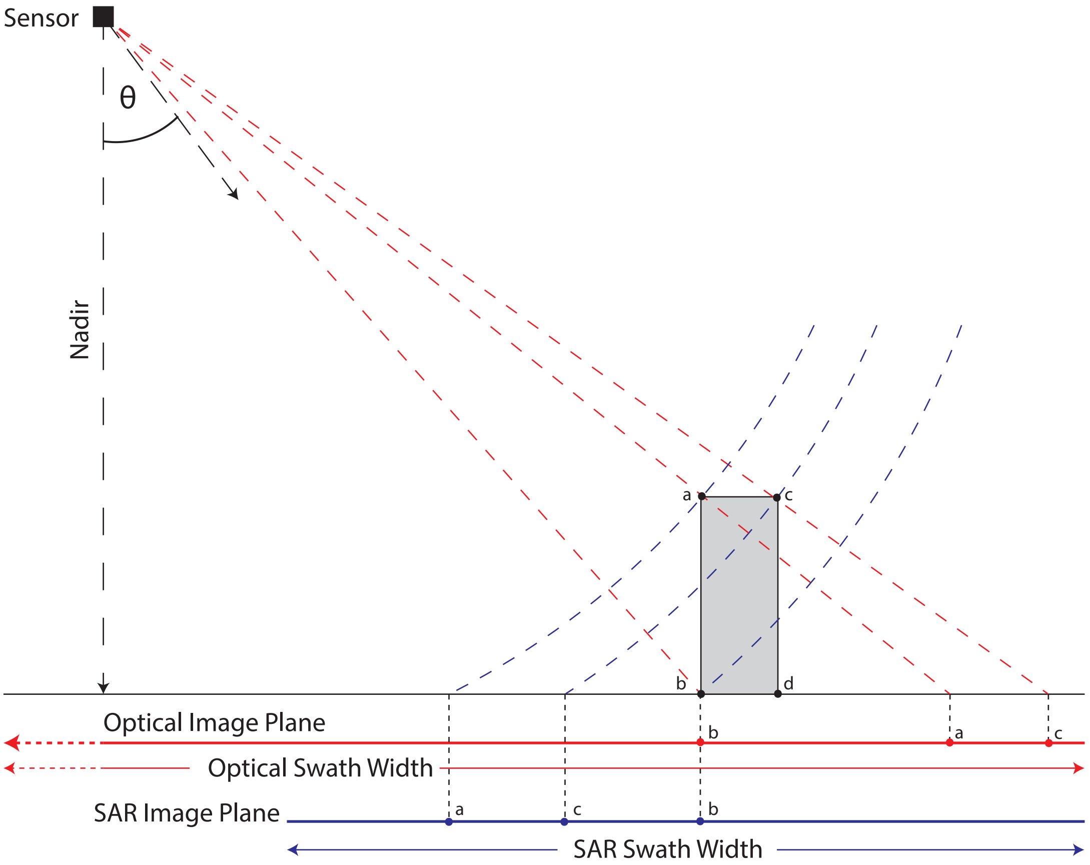


Figure 2.3: Comparison of optical (red) and SAR (blue) image formation concepts. The raised points (a) and (c) can be seen being projected towards the sensor in the SAR image plane, and away from the sensor in the optical image plane. Point (b) which is on the Earth is not geometrically distorted by the different modalities, and projects to the same locations in both images. While (d) is not visible to either sensor.

plane. This occurs when the slope of the terrain or object is less than the θ or if the slope is facing away from the sensor with an angle of less than $90° - \theta$. While the distance is shortened there is no inversion of the object.

The final distortion, shadowing, occurs when there is no direct line-of-site from the sensor to the object, thus no information can be obtained and the region appears dark in the final image. Shadowing commonly occurs in urban and mountainous areas where a raised point obscures the few of the ground or objects behind it. An example of these three SAR specific geometric effects can be seen in Figure 2.5

Although layover and foreshortening have no analogue in optical imagery, shadowing can be thought of as being similar to occlusion. The main difference being that shadowing appears as a dark region in the SAR image (as the visible face is subject to layover), while in an optical image the visible face falls away from the sensor to occlude the objects behind it.

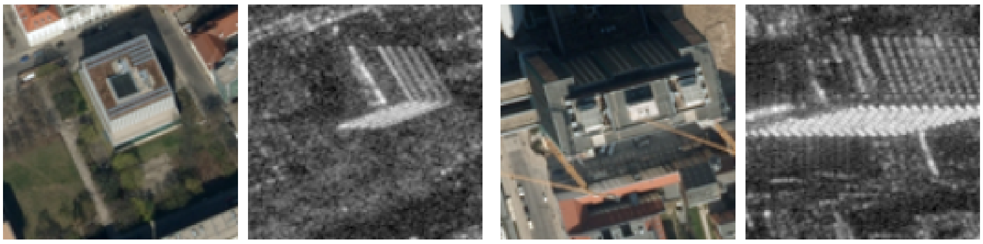


FIGURE 2.4: Two examples of optical and SAR imagery of the same scene, highlighting the distortion of tall objects in each modality.

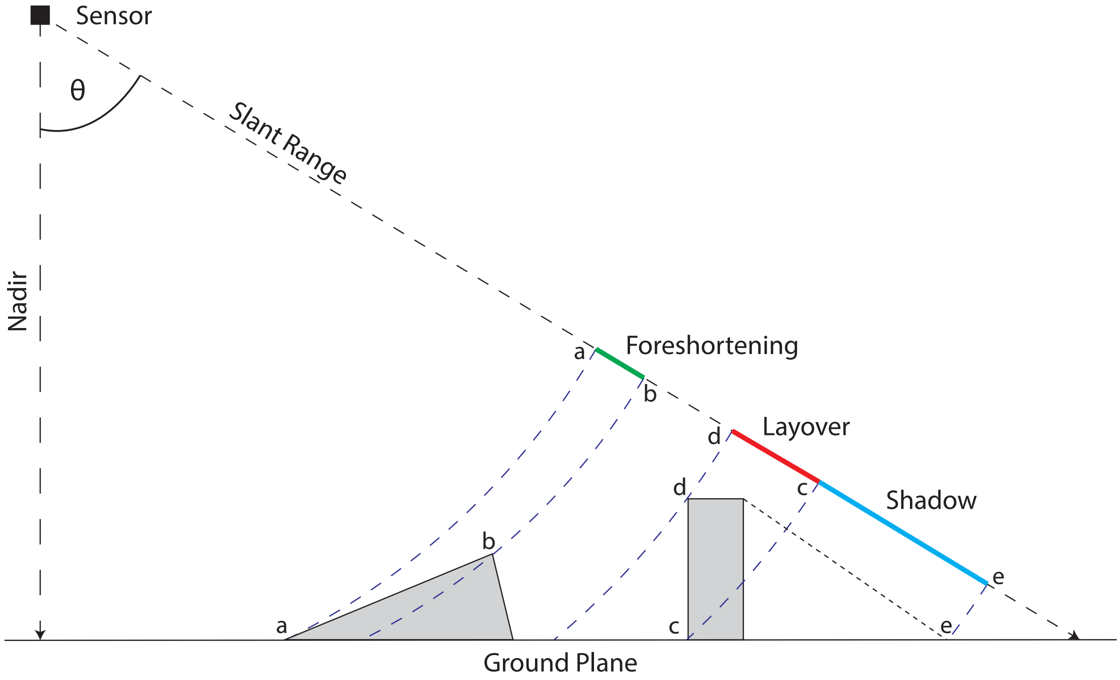


FIGURE 2.5: Geometric distortions which occur in the SAR image plane. Foreshortening (green) causes the compression of an objects dimensions, while layover (red) leads to coordinate inversion as the object 'falls' towards the sensor, finally shadow (blue) can be seen as dark/empty regions in the image.

Radiometric Characteristics

An objects response to EM-radiation is dependent on both object properties, such as roughness, conductivity, and reflectivity, as well as on the polarization and wavelength of the exciting signal. As the frequency of the signals used in SAR and optical image formation differ by several degrees of magnitude, it is not surprising that the radiometric properties of the resultant imagery are vastly different.

As optical sensors operate within the range of visible and infrared radiation, the pixel values in optical imagery can be interpreted as a characterization of the chemical composition of the scene. On the other hand, the lower frequency used in SAR imaging provides less information about the chemical composition and more about the structural and geometric properties of the scene. The intensity of pixels in a SAR image directly related to the roughness, electrical conductivity and orientation of the object relative to the sensor (Curlander, 1982).

Another effect of the lower frequency of SAR imagery is the salt-and-pepper like noise, *speckle*, which occurs across the image. This is not noise, but rather the effects of additive and destructive interference which arises due to multi-path effects and multiple scatterers existing within the same resolution cell. Speckle, as well as the other radiometric differences, can be seen in Figure 1.1.

2.1.3 Considerations for SAR-Optical Matching

From the discussion on image geo-localization in Section 2.1.1, it can be seen that finding correspondences between SAR and optical imagery can lead to improvements in optical geolocalization (Müller et al., 2012). While this is an important use case many other applications exist, such as SAR-optical stereogrammetry (Bagheri et al., 2018; Qiu et al., 2018), location of ground control points (GCPs) in optical imagery (Bürgmann et al., 2019) and many other data fusion tasks (Schmitt et al., 2017; Tupin, 2010).

Given the substantial radiometric and geometric differences which exist between SAR and optical imagery, as discussed in Section 2.1.2, it is clear as to why matching these modalities is a deeply complex problem. These differences directly affect the robustness and suitability of various matching approaches to the problem. Furthermore, they make human interpretation and manual matching of the imagery a difficult task, even for experts. This has a significant effect on the availability of labelled training data, and the suitability of supervised learning on the matching problem.

At low and medium resolutions the geometric differences are, in many cases, negligible and the matching problem is largely reduced to dealing with the radiometric differences. However, as the resolution increases the geometric distortions become more pronounced and add to the complexity of the problem. This is especially true in urban and suburban areas where layover and shadowing are common occurrences and begin to interact with each other due to the density of above ground structures.

In Chapter 3 the problem of matching high-resolution optical and SAR imagery is addressed, and various proposals are made as to how to deal with these complexities from both a deep learning, as well as, an image matching perspective.

2.2 The Fundamentals of Image Matching

Image matching has long been a topic of research within the field of computer vision and almost all fields related to image analysis and processing, including remote sensing (Gruen, 2012; Szeliski, 2010). While there have been a significant developments in image matching approaches across the board, the majority of approaches remain domain specific and are not directly transferable to other modalities (Gruen, 2012).

However, domain specific approaches to image matching often take inspiration from computer vision methodologies and thus many parallels exist between them. Most notably, the fundamental concepts and taxonomy of image matching approaches are transferable across domains.

In general the image matching process can be broken down into three steps, namely, feature detection, matching and outlier removal. While these three stages often exist as separable components of a matching pipeline, it is also possible that two or more of them are combined into a single step. However, each component fulfils an important role in the image matching process and is required in order to enable fully automatic image matching, irrespective of the imaging modality.

In this section each of these foundational components are described and a high-level overview of the most common methodologies for each are described. Due to the breadth and depth of research in image matching the discussion is limited to methodologies which have later seen adaptation within the field of SAR-optical image matching. For a complete review of image matching approaches and methodologies the reader is referred to (Steger et al., 2018; Szeliski, 2010). Furthermore, for an in-depth review of the historical development of image matching within photogrammetry and remote sensing the reader is referred to (Gruen, 2012).

2.2.1 Feature Detectors

Features can be defined as sub-regions in an image which contain a pattern that is distinctive from immediately nearby pixels. Thus features can usually be linked to a physical image or object property such as corners, edges or blobs (Leng et al., 2019; Li et al., 2015).

The role of a feature detector is to find these high saliency and descriptive local regions, such that a spatial extent can be extracted, and matching algorithms can be applied in order to determine correspondences between images.

While many feature detectors have been proposed, they are very tightly coupled to a specific modality. This combined with a broad diversity of image conditions (i.e. illumination and viewpoint changes, image quality, resolution) means that no ideal feature detector exists, and the design or selection of a feature detector is mostly dependent on the application (Salahat et al., 2017). This problem is further exacerbated when dealing with feature detection in multi-modal imagery, where the goal of the detector is not only to locate salient features but to do so in a manner which ensures that there is an overlap between the features detected in each modality.

For these reasons many feature detectors are designed around the use of secondary information, such as image gradients. This allows for the detector to be adapted to for use across a wide range of modalities by replacing the gradient operator with the applicable operator for the new modality. Furthermore, features based on image gradients are less affected by varying image conditions, and thus lead to improved robustness of the detector Leng et al. (2019). The utility of this approach can be seen in the widespread and continued use of the Harris corner detector (Harris et al., 1988), as well as the Scale Invariant Feature Transform (SIFT) (Lowe, 2004).

While being one of the oldest gradient-based feature detectors, the Harris corner detector remains in widespread use due to its simple nature. The Harris detector is based around finding 3 × 3 pixel regions which exhibit high variation (a large sum of squared difference) when subject to a small translation in any direction (Harris et al., 1988). Regions which exhibit a high variation in a single direction are labelled as edges, while regions with high variation in multiple directions are labelled as corners. On the other hand, low variation regions are deemed to be unsuitable as feature points. Due to the simplicity of this approach, it is computationally efficient and produces a large number of feature points. To reduce the overall number of features, Harris detectors usually include a non-maximal suppression (NMS) phase which limits the final set of features by suppressing the response of non-maximal features within a certain radius of a local-maxima (Szeliski, 2010).

Compared to the Harris detector, the Scale Invariant Feature Transform (SIFT), proposed by Lowe (2004), is a computationally expensive feature detection algorithm. However, it has become one of the most widely used and adapted feature detectors due to its robustness and adaptability (Burger et al., 2016; Suri, Schwind, et al., 2010; Xiang et al., 2018). To detect robust feature points, SIFT constructs a difference of Gaussian (DoG) pyramid. This is done by convolving each level of an image scale space (octave) by a set of Gaussian kernels of increasing standard deviation. The differences between adjacent filtered images in each octave are then computed to form the DoG pyramid. Local extrema are then detected by comparing each pixel to its eight spatial neighbours, as well as the nine pixels in the scale space above and below it. If the pixel value is larger or smaller than all of its neighbours, it is considered a candidate feature point. Next, the keypoint location is optimized to sub-pixel accuracy, and then two final checks are applied to remove unstable feature points, such as edges. These checks are based on the eigenvalues of the Hessian matrix for the key-points, and follow similar criteria to the Harris corner detector to differentiate corners from edges.

While both Harris and SIFT still see extensive utilization in many image matching pipelines, in recent years, there has been a move away from hand-crafted feature detectors in favour of learned detectors (DeTone et al., 2018; Laguna et al., 2019; Yi et al., 2016). This has primarily been driven by the maturation of deep learning techniques and image matching within optical computer vision applications. However, while these learned detectors have shown state-of-the-art results in conventional computer vision applications, they are not yet suitable for use in SAR-optical matching. This is due to the large amounts of training data they require, or the training mechanism is based around a set of assumptions which do not hold under multi-modal conditions.

2.2.2 Methodologies for Matching

The matching stage of the pipeline is responsible for determining which points across a set of images are likely to represent the same point in actuality. This is the fundamental stage in an image matching pipeline, and the feature detection and outlier removal stages can be seen as auxiliary tasks which simplify the problem of determining correspondence by reducing the search space and removing errors.

Historically matching approaches can largely be separated into two main classes, namely, intensity-based and feature-based matching (Shapiro et al., 1992). However, more recently the third class of matching approaches has appeared, namely, deep matching. These approaches rely on the descriptive nature of feature maps extracted by convolutional neural networks (CNNs) to enable matching in complex high-dimensional spaces (Fischer et al., 2014).

Intensity-based Matching

Intensity-based approaches rely directly on pixel intensity values to determine correspondences across a set of images. These approaches usually forego the feature detection stage of the matching pipeline in favour of a more direct, albeit, computationally intensive approach.

To determine correspondences across images, intensity-based methods make use of similarity metrics which compute the agreement between image regions based on the intensity values within the sub-region. Due to this, these methods are generally implemented in a sliding window manner whereby a small template is progressively moved within a larger search region to determine the point of best correspondence (Ghaffary, 1986), as depicted in Figure 2.6.

To ensure a diverse spatial distribution of correspondences, intensity-based methods frequently select search and template patches based on a uniform grid of points sampled across the images to be matched. However, this approach relies on the assumptions that the images are related by a local offset and that the upper bound of this offset is known. As this is the case in most SAR-optical matching tasks, where the geolocation and upper bound of the optical geo-localization error are known, intensity-based methods have seen significant use within this domain. In cases where the images are not coarsely aligned, intensity-based methods fall back to a global search which comes at a considerable computational cost (Zitová et al., 2003).

While less common for image matching, intensity-based methods can be initialised using features detected in one image and the assumption of small local offsets between images. This approach is known as optical-flow (Horn et al., 1981) and is commonly used in video-based object tracking.

Due to the reliance on raw intensity values, and the multitude of distortions which can occur, the design of a robust similarity metric is paramount to the success of intensity-based matching methods. Within the realm of remote sensing, and multimodal matching the most frequently used similarity metrics are based on normalized cross-correlation (NCC) and mutual information (MI) (Merkle, 2018; Suri & Reinartz, 2010; Wang et al., 2012).

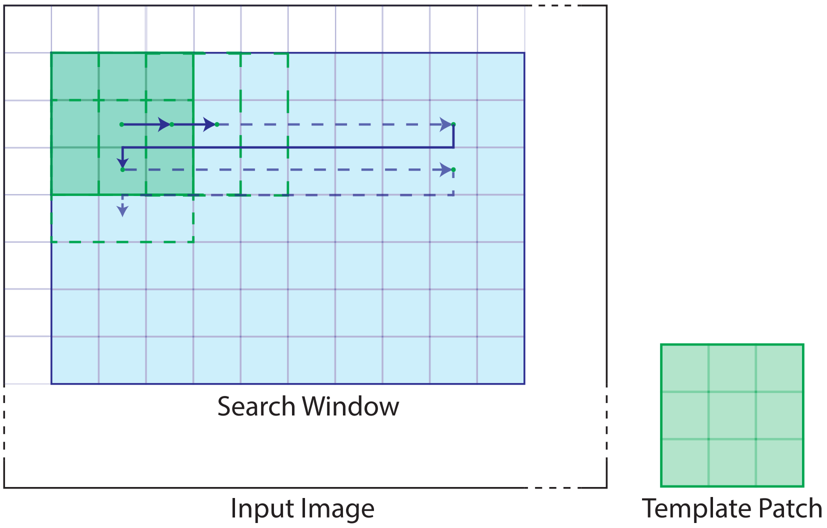


FIGURE 2.6: Illustration of the sliding window mechanism used in intensity-based matching approaches. The template patch **T** is iteratively moved across a defined search window in a left-to-right, top-to-bottom manner. For each offset, a similarity metric is computed between **T** and the overlapped region **R**, within the search window. The resultant value is an indication of the likelihood of the offset being the point of correspondence between **T** and **R**.

Cross-correlation metrics, such as NCC, are based on the idea of finding the offset which maximizes the correlation function between the search and template patch (Ghaffary, 1986). This implies calculating the NCC similarity metric between the overlapping regions for each offset within the search window, as depicted in Figure 2.6.

Considering an image patch **R** of size (N, M) and centered at the point (x, y) in the reference image **I**, the NCC similarity metric between this region and a template patch **T** of size (N, M), extracted from the input image **I**$'$, is defined as,

$$\mathrm{NCC}(x,y) = \frac{\sum\limits_{(u,v)\in\mathbf{T}} \Big(\mathbf{R}(x+u,y+v) - \overline{\mathbf{R}}_{u,v}\Big)\Big(\mathbf{T}(u,v) - \overline{\mathbf{T}}\Big)}{\sqrt{\sum\limits_{(u,v)\in\mathbf{T}} \Big(\mathbf{R}(x+u,y+v) - \overline{\mathbf{R}}\Big)^2 \sum\limits_{(u,v)\in\mathbf{T}} \Big(\mathbf{T}(u,v) - \overline{\mathbf{T}}\Big)^2}} \tag{2.1}$$

Where $(u, v) \in \mathbf{T}$ represents the set of all pixel offsets in the template patch and $\mathbf{R}(x+u, y+v)$ and $\mathbf{T}(u, v)$ are the intensity values of the region and template at the offset (u, v). Furthermore, $\overline{\mathbf{T}}$, and $\overline{\mathbf{R}}$ represent the mean intensity value of the template patch, and the image region overlapping with the template patch, respectively. To determine the point of maximum correspondence the NCC is computed at each (x, y) coordinate in the search region. The argument of the maxima is then selected to be the point of correspondence if the maxima is above a predefined threshold (Zitová et al., 2003).

Another widely used similarity metric for intensity-based matching, is mutual information (MI). It is computes the similarity between two image patches based on a comparison between their pixel intensity distributions. Thus the MI between a template patch $\mathbf{T}$ and an image region $\mathbf{R}$, cropped from the search window at (x, y), can be defined as follows,

$$\mathrm{MI}(\mathbf{T}, \mathbf{R}) = H(\mathbf{T}) + H(\mathbf{R}) - H(\mathbf{T}, \mathbf{R}), \tag{2.2}$$

where $H(\circ)$ represents the marginal Shannon entropy of the respective patch, and $H(\mathbf{T}, \mathbf{R})$ is the joint entropy. Both the marginal and joint entropy can be directly calculated from a 2-dimensional co-occurance matrix of $\mathbf{T}$ and $\mathbf{R}$ image intensities, as per the formulations presented in (Chen et al., 2003).

As with NCC, the process of determining the most likely point of correspondence involves computing the MI at every location (x, y) within the search window, and then selecting the argument of the maxima as the point of correspondence. The MI value has a range from 0 to 1 with 1 indicating that the patches are identical. Thus the point of correspondence can be further verified using a threshold on the maximum MI value.

Both NCC and MI include normalization mechanism, which allows for these metrics to be relatively robust to changes in illumination and other radiometric distortions between the template and search images. However, due to the direct approach of these methods, they are very sensitive to geometric differences and can often break down quickly even under modest geometric distortions.

Feature-based Matching

Feature-based approaches to matching rely on the description of distinctive salient features within the image space. These points are usually found using hand-crafted feature detectors, as was discussed in Section 2.2.1.

Once features have been detected, a feature description algorithm is used in order to create a unique vector description of the feature which can be used for matching. This is commonly done by extracting a small template patch around the identified feature point and performing a series of transformations on it in order to derive an N-dimensional latent vector for the feature. This vector should not only be unique, but should also have the property that similar features are mapped to a similar location within the latent space (Szeliski, 2010).

Correspondences are then determined by computing a matching score between sets of feature vectors extracted from multiple images. In many cases these matching scores are simply computed as the Euclidean distance between two vectors, however, they can be based on other metrics too. A pair of feature vectors is thus labelled as corresponding when the matching score exceeds a predefined threshold.

From this it can be seen that the success of feature-based matching is highly dependent on the design of a good feature descriptor. Over the years a wide variety of feature descriptors have been designed which trade-off speed, accuracy and robustness depending on the application. However, to date, the SIFT descriptor (Lowe, 2004) still remains the most widely used descriptor and has seen a number of modifications for application in other domains, and to improve speed and robustness (Chen et al., 2003; Wu et al., 2013).

The SIFT feature descriptor is a 128-dimensional vector which encodes information about the local image gradients around a feature point. To form this vector a 16×16 pixel window is selected around the previously detected feature point. The gradient vectors for each pixel within this window are computed, and a Gaussian weighting is then applied to prioritize gradients nearest to the feature point. The descriptor window is then sub-divided into sixteen 4×4 windows and an eight bin gradient histogram is computed for each sub-window. These histograms are then concatenated and the resultant vector is normalized to unit length to form the final feature vector.

The application of Gaussian weighting and the normalization of the feature vector make SIFT descriptors largely invariant to small translational offsets as well as to large variability in contrast and illumination. However, this invariance does not translate to robustness when considering multi-modal imagery with vastly different image properties, although modifications to SIFT have been proposed to deal with these cases (Suri, Schwind, et al., 2010; Xiang et al., 2018).

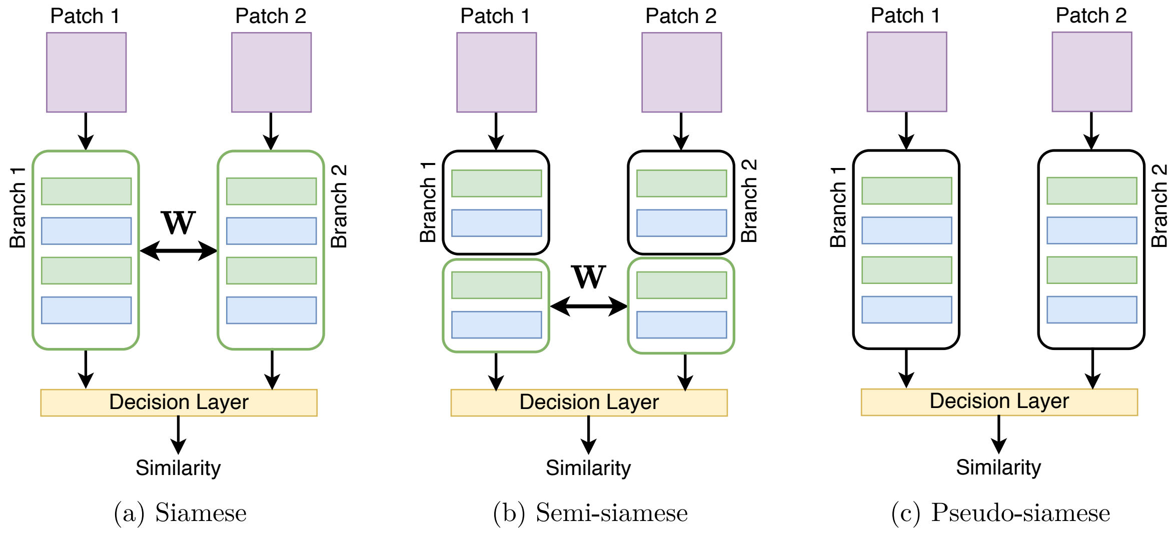
The biggest constraint of feature-based approaches lies in the fact that the feature detector and descriptor need to be carefully designed such that the same feature points can be located and accurately matched across a set of images. Thus they are often designed to with a specific application or sensor modality in mind, and do not translate well to other modalities without modification. This constraint becomes more complex to resolve when dealing with matching across different modalities, where features can have vastly different appearances, and large geometric distortions can exist between images.

Deep Matching

In recent years deep learning-based approaches to image matching have gained much popularity and are evermore becoming the go-to approach in image correspondence problems. This is primarily due to the ability of deep Convolutional Neural Networks (CNNs) to model complex features and relationships within the image domain (O'Mahony et al., 2019; Schönberger et al., 2017). Which has allowed for the development of deep matching solutions which are more robust to variation in image condition and view-point than their intensity and feature-based predecessors. These improvements have, however, come at the cost of increased computational complexity; although this is less and less becoming a deciding factor in algorithm selection.

Fischer et al. (2014) provided one of the earliest insights into applying deep learning to the image matching problem, by using features extracted from the last layer of a CNN in place of SIFT as a feature descriptor. In doing so, it was shown that the features learned by CNNs can outperform conventional feature descriptors in image matching tasks.

Unlike intensity and feature-based matching approaches which have a relatively fixed methodological form, deep matching approaches have been developed in many forms which often take inspiration from these methodologies. Initial deep matching architectures aimed to replace the similarity metrics used in intensity-based methods (Fischer et al., 2014; Zagoruyko et al., 2015), or the feature descriptor component of feature-based methodologies (Balntas et al., 2016; Han et al., 2015; Simo-Serra et al., 2015). In both these initial forms of approaches, the networks relied on pre-existing feature



(a) Siamese (b) Semi-siamese (c) Pseudo-siamese

FIGURE 2.7: Taxonomy of 2-stream matching architectures. Each stream consists of a CNN which either (a) share weights **W** across all layers, (b) share weights in high-level layers to allow for initial feature independence, or (c) are completely independent to allow for modality-specific features to be learned at all levels.

detectors or template style matching, where correspondences were determined between fixed-size image patches. However, in recent years deep matching architectures have evolved to offer a comprehensive solution to the image matching problem by handling feature detection, matching and outlier removal in a single end-to-end deep learning architecture, and thus being able to operate on full-scale imagery DeTone et al. (2018) and Yi et al. (2016).

However, out of the diverse arrangement of deep matching architectures which exist, two-stream network architectures, such as those depicted in Figure 2.7, have shown the most promise for matching applications which involve geometric distortions or multi-modal imagery (Liu et al., 2018; Simo-Serra et al., 2015; Zagoruyko et al., 2015; Zhu et al., 2019).

Two-stream networks commonly take one of two forms; either they include a decision (or metric) network which directly outputs a similarity score, as depicted in Figure 2.7. Or they output feature vectors for each stream which are then matched using conventional distance metrics such as the Euclidean distance.

While two-stream networks and other deep matching approaches have grown in popularity due to their robustness, accuracy and adaptability, they are still limited to use within a specific domain or application. However, unlike other methodologies, this limitation is not inherent in their design but instead based upon the data and loss functions which were used to train the model. Thus deep matching approaches can easily be adapted and modified for use in other domains, given sufficient labelled training data and computational resources. Furthermore, recent deep matching approaches have begun to explore the use of self-supervised methods to reduce the requirement for labelled training data (DeTone et al., 2018). However, these approaches require mathematical models of the expected distortions and image space transformations to generate artificial scenarios which accurately model reality.

The constraints introduced by the need for large-scale labelled training data, as well as accurate mathematical models of feature transfer are by far the biggest obstacles facing the development and success of deep matching within complex multi-modal domains.

2.3 SAR-Optical Matching: A Review

Optical and SAR are two of the most important modalities in spaceborne remote sensing applications. Due to the highly complementary nature of their imagery (Section 2.1) data fusion of these modalities has become a critical task in deriving insights for global-scale Earth observation applications (Schmitt et al., 2017; Zhang, 2010). For this reason, research into SAR-optical image matching has grown substantially, with a wide range of approaches having been proposed over the years. In this section, a review of the existing approaches to the SAR-optical matching problem is provided, and the open problems which this thesis aims to address are highlighted.

One of the seminal works in this domain was proposed by Li et al. (1995), who approached the problem as an image registration task. Whereby strong contours in each modality where extracted and then iteratively aligned.

Several years later, after significant developments within the field of traditional image matching, Dare et al. (2000) proposed the first work detailing the use of domain-specific features for SAR-optical matching. This lead to various feature-based approaches being developed, which were primarily based around the idea of using edge and contour segments as features within a regression-based matching framework (Cheng et al., 2004; Zhaohui et al., 2004).

The development and growing reputation of the SIFT descriptor (Lowe, 2004), within the domain of traditional computer vision applications, lead to many modifications of it being proposed for use in remote sensing image matching (Dellinger et al., 2015; Gong et al., 2014; Suri, Schwind, et al., 2010). While these approaches provided a relatively successful means of matching spacebourne SAR imagery, they still failed to provide a means for enabling SAR-optical matching. This was largely due to the features detected by these SAR-specific SIFT implementations being independent of those detected in the optical domain (Ma et al., 2017).

To address this a number of additional modifications were added to the SIFT pipeline in order to enforce consistency between the detected feature points and extracted descriptors in each modality (Fan et al., 2012; Fan et al., 2014; Xu et al., 2015). However, these approaches were largely limited by the fundamental concept of SIFT, which is related to the use of image gradients for feature detection and description. In optical imagery the concept of image gradients is largely related to edges and object boundaries, while in SAR imagery speckle and the nature of the imaging concept leads to ambiguous object boundaries and significant fluctuations in local intensity. Thus the original and modified SIFT detectors could still not provide robust features for matching in high-resolution environments.

To account for these large differences between SAR and optical imagery at higher resolutions, Ye et al. (2016) argued for the use of an auxiliary image representation as a proxy for image gradients. This lead to the development of the histogram of oriented phase congruency (HOPC) descriptor. In a similar vein, Xiang et al. (2018) coupled modality specific gradient operators with a Harris scale-space to better account for the vast radiometric differences between the modalities. Li et al. (2020) combined these insights and phase-congruency to develop the Radiation-variation Insensitive Feature Transform (RIFT).

Unlike feature-based approaches, which have seen a rich and diverse history of development, intensity-based methods have seen limited see use in SAR-optical matching. A significant reason for this is their lack of robustness to extreme differences in radiometric and geometric distortions. However, this comes with the exception of Suri and Reinartz (2010), who successfully confirmed the use of MI as a similarity metric for matching SAR and optical imagery. This approach was later extended to include a genetic search algorithm which lead to improved robustness and a lower computational cost (Fischer et al., 2018).

Although these previous approaches can detect and match features across SAR and optical modalities, their success has primarily been limited to within the bounds of specific geometric and radiometric constraints. Often these constraints restrict them to use within flat, semi-urban or rural environments where the differences between SAR and optical imagery are primarily constrained to the radiometric properties.

Driven by the success and maturation of deep learning for conventional image matching tasks, Section 2.2.2, remote sensing practitioners adopted these methodologies to address the shortcomings of feature-based approaches for SAR optical matching. In doing so, several deep matching approaches have been proposed which deal with the inherent heterogeneity between SAR and optical imagery. The seminal works on SAR-optical deep matching were proposed in short succession by Merkle, Luo, et al. (2017) and Mou et al. (2017). While both these approaches made use of two-stream networks, the approaches they took to the matching problems were fairly distinct. Merkle, Luo, et al. (2017) proposed a siamese network which operated on a search and template patch to create pixel-wise feature descriptors. These were then matched using a dot-product to create a correspondence heatmap, from which the match could be extracted. Alternatively, Mou et al. (2017) proposed the use of a pseudo-siamese architecture with modality-specific streams and a fusion (metric) network to compute a similarity score for a SAR-optical patch pair. Later this approach was adapted by Citak et al. (2019) to include SAR and optical saliency maps as an attention mechanism in the modality-specific streams. Bürgmann et al. (2019) argued for the use of a modified HardNet architecture (Mishchuk et al., 2017) to incorporate the use of hard negative mining and a triplet loss in the SAR-optical matching problem. Hoffmann et al. (2019) trained a single stream Fully Convolutions Network (FCN), using the concatenation of the optical and SAR imagery as input, to estimate the similarity between the patches. Taking a multi-stage, multi-scale approach, Ma et al. (2019) used features extracted from a fine-tuned VGG16 (Simonyan et al., 2015) model to propose a coarse-to-fine registration pipeline. In a vastly different approach, Generative Adversarial Networks (GANs) were proposed to translate SAR patches into pseudo-optical template patches which could then be matched in the optical domain using standard intensity and feature-based methods (Merkle, Auer, et al., 2017).

Although significant progress has been made in development of SAR-optical matching methodologies, the accuracy and success of these matching methods is largely reliant on the quality of the feature points which are used for the extraction of candidate patch pairs. However, this problem has received little focus over the years due to the inherent complexity of determining jointly visible, and salient features across such vastly different modalities. Thus previous matching approaches have relied on the detection of features using computationally intensive auxiliary data (Bürgmann et al., 2019; Merkle, Luo, et al., 2017), or assumed correspondence based purely on geo-localization (Citak et al., 2019; Hoffmann et al., 2019; Ma et al., 2019). While these approaches have worked for creating training datasets and evaluating matching methodologies, the assumptions they hold or data they require have limited the application and development of deep SAR-optical matching approaches to medium resolution imagery or semi-urban areas. Thus the development of a deep learning-based SAR-optical matching pipeline suitable for matching high-resolution imagery, across a wide range of scenes has not yet been achieved.

To this end, the main contribution of this thesis is the further development of existing matching methodologies into a scalable, comprehensive and fully-automatic framework for end-to-end matching of high-resolution SAR and optical imagery without the reliance on auxiliary data or strong assumptions.

3. Deep Learning for SAR-Optical Image Matching

This chapter presents the key contributions of this thesis in terms of developing deep learning-based approaches to enable automatic matching of high-resolution SAR and optical imagery. The addressed topics directly relate to the publications in Chapter 4 which form the basis of this cumulative thesis.

Firstly, the topic of deep learning and SAR-optical correspondence is addressed, and various architectures are proposed to enable matching of these modalities. Following which the topic of scarce data is presented and solutions to dealing with the lack of large-scale training data are proposed. Finally, deep learning-based architectures for feature detection and outlier removal are proposed and these various sub-tasks are chained together in a logical manner to complete the requirements for a comprehensive SAR-optical matching framework.

3.1 Deep Learning for Determining SAR-Optical Correspondences

Peer-Reviewed Publications Related to this Section

Hughes, L. H., Schmitt, M., Mou, L., Wang, Y., & Zhu, X. X. (2018). Identifying corresponding patches in SAR and optical images with a pseudo-Siamese CNN. *IEEE Geoscience and Remote Sensing Letters*, *15*(5), 784–788.

Hughes, L. H., Marcos, D., Lobry, S., Tuia, D., & Schmitt, M. (2020). A framework for sparse matching of SAR and optical imagery [Under Review]. *ISPRS Journal of Photogrammetry and Remote Sensing.*

Schmitt, M., Hughes, L. H., & Zhu, X. X. (2018). The SEN1-2 dataset for deep learning in SAR-optical data fusion. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *IV-1*, 141–146.

As introduced in Section 2.2, the concept of image matching is largely based around three sub-tasks, namely, feature detection, correspondence and outlier removal. While feature detection and outlier removal play an important role in enabling large-scale image matching, the determination of correspondences remains the most fundamental sub-task in image matching endeavours. The importance of the image correspondence problem in SAR-optical matching is founded in the need for a method which can determine similarity across the vast differences between the modalities.

As discussed in Section 2.3, past approaches to determining correspondences between SAR and optical imagery have largely been based around hand-crafted feature descriptors (Suri, Schwind, et al., 2010; Xiang et al., 2018; Ye et al., 2016; Ye et al., 2017). Although these approaches have seen successful application in SAR-optical matching, their application is largely constrained to specific scene geometry and sensor resolutions. More recently, advancements in deep learning have shown the ability for deep matching architectures to outperform conventional feature-based matching approaches (Schönberger et al., 2017). Based on these advancements, a number of deep matching approaches have been proposed towards the problem of determining correspondences between SAR and optical imagery (Merkle, 2018; Mou et al., 2017). These initial SAR-optical deep matching approaches have shown great promise in being robust to the large radiometric differences between the modalities. However, these approaches are far from being robust to geometric distortions, and have largely been proposed in the frame of initial investigations (Mou et al., 2017) or for use in rural and suburban areas (Merkle, 2018).

Taking inspiration from these initial investigations, and the advancements made within conventional deep matching, two novel deep learning-based approaches to the SAR-optical correspondence problem were developed within the scope of this thesis. Furthermore, to support the development of generalizable deep learning based solutions, two large-scale SAR-optical correspondence datasets were also created.

3.1.1 Datasets for Deep Matching

One of the largest determining factors in the success and generalizability of supervised deep learning methods, is the quality and quantity of the data used to train and evaluate the model (Sun et al., 2017).

In conventional applications, a manual labelling process is often used to create large-scale labelled datasets. However, the vast radiometric and geometric differences between SAR and optical imagery (Section 2.1.1) make this task intractable, even for domain experts. Automated approaches have been proposed for creating SAR-optical correspondence datasets (Merkle, 2018; Wang et al., 2018). However, the methods either rely on computational complex processes (Wang et al., 2018), or are based on strong assumptions (i.e. only using non-urban areas) that limit the generalization of models trained on them (Merkle, 2018).

Thus prior to the development of deep-matching approaches, and within the frame of this thesis, two novel, large-scale SAR-optical correspondence datasets are created. The first dataset, called *SEN1-2*, is based on openly available medium resolution imagery. While the second dataset is based on high-resolution imagery from the Urban Atlas project (Schneider et al., 2010).

SEN1-2: Medium Resolution

The *SEN1-2* dataset consists of 282,384 corresponding pairs of Sentinel-1 SAR and Sentinel-2 optical imagery. The dataset was created using the Google Earth Engine

(GEE) platform and data catalogue (Gorelick et al., 2017), and every patch was manually verified by removing patches which contained artefacts, clouds or other errors.

Sentinel-1 ground-range-detected (GRD) data products acquired in interferometric wide swath (IW) mode were used as the basis of the dataset. To ensure precise orthorectification the products were calibrated and terrain corrected based on precise orbit information and a 30m DEM (either SRTM-DEM or ASTER-DEM depending on the latitude). For simplicity, the σ^0 backscatter coefficient in Decibels, and VV polarization were used to create the SAR patches. Each image has a spatial resolution of 5m in azimuth and 20m in range, and uses square 10×10m pixel spacing. For the Sentinel-2 imagery only the red, green and blue channels (bands 4,3 and 2) were selected to create the resultant RGB images. As the Sentinel-2 imagery is provided as accurately geo-referenced granules, no further pre-processing was required. However, the initial image selection was filtered by cloud coverage, to ensure the images contained as few clouds as possible.

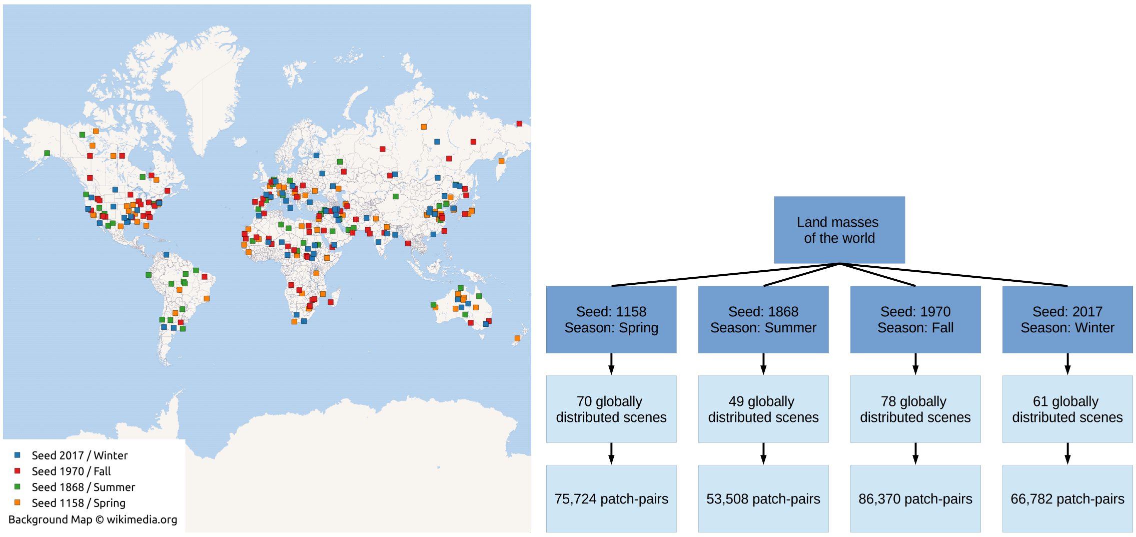
In order to generate a globally representative dataset, locations were sampled uniformly across the landmasses of the Earth. An approximately 100km^2 region was created around each point, and a season identifier was randomly assigned. These season identifiers were later used to specify the time range, according to northern hemisphere meteorological seasons, of imagery to use in creating a mosaic for each region.

The resultant mosaicked images were then exported from GEE and tiles of 256×256 pixels were created for each scene. During the tiling process a stride of 128 pixels was used to create patches with an overlap of 50%. This was deemed to be the best trade-off between patch independence and the number of patches per scene. Finally, each SAR and optical patch was visually inspected to ensure that no no-data areas, clouds, shadows or other significant artefacts were present. If an artefact was found in either the SAR or optical imagery, the patch pair was removed from the dataset. A summary of the final dataset, and the distribution of the 282,384 patches is depicted in Figure 3.1, while a collection of example patches is shown in Figure 3.2.

As the SEN1-2 dataset is based on medium resolution imagery, the geometric differences between the modalities become less apparent. This combined with the fact that the source imagery is accurately co-registered, means that correspondence between the image patches can be assumed based on the geo-location. Thus the complexity of the image matching largely related to the radiometric differences between the modalities. While this is a simpler problem, the SEN1-2 dataset has still seen widespread use in the development of many other deep learning based approaches (Bürgmann et al., 2019; Citak et al., 2019; Hoffmann et al., 2019).

Urban Atlas: High-Resolution

The original Urban Atlas dataset (Schneider et al., 2010) consists of 46 manually co-registered, high-resolution SAR and optical images acquired over 13 cities across Europe. The spatial distribution of these cities is depicted in Figure 3.3. The SAR images were acquired by the TerraSAR-X sensor operating in stripmap mode with a spatial resolution of 1.25m. These images were then processed into Enhanced Ellipsoid Corrected (EEC) data products (Breit et al., 2009). The optical images were acquired at a spatial resolution of 2.5m by the panchromatic PRISM sensor. As mentioned


■ Seed 2017 / Winter
■ Seed 1970 / Fall
■ Seed 1868 / Summer
■ Seed 1158 / Spring
Background Map © wikimedia.org
Land masses of the world
Seed: 1158 Season: Spring
Seed: 1868 Season: Summer
Seed: 1970 Season: Fall
Seed: 2017 Season: Winter
70 globally distributed scenes
49 globally distributed scenes
78 globally distributed scenes
61 globally distributed scenes
75,724 patch-pairs
53,508 patch-pairs
86,370 patch-pairs
66,782 patch-pairs

(a) (b)

FIGURE 3.1: The spatial and seasonal distribution of the final SEN1-2 dataset, after the removal of artefact-affected patches and regions. (a) shows the spatial distribution of the regions and selected seasons, (b) depicts the breakdown of patches within the dataset by season and quantity.

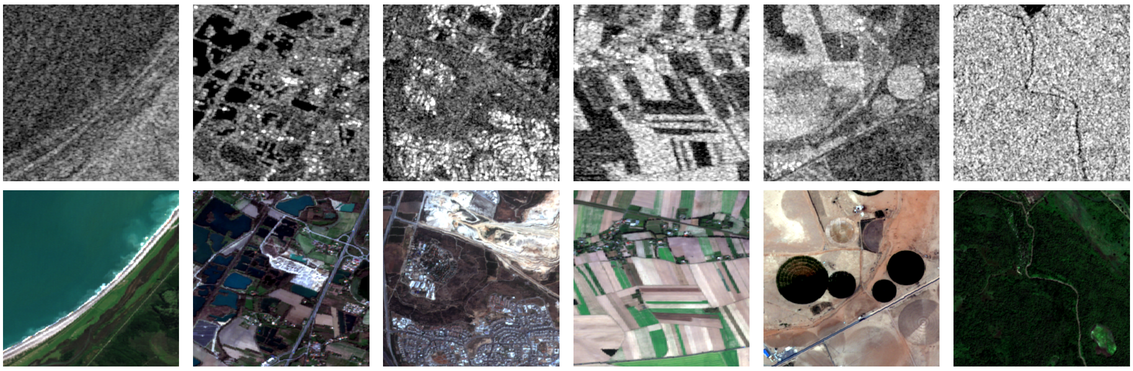


FIGURE 3.2: Some exemplary patch-pairs from the SEN1-2 dataset. Top row: Sentinel-1 SAR image patches, bottom row: Sentinel-2 RGB image patches.

in Section 2.1, the geo-coding of optical imagery suffers from inaccuracies, and thus the alignment between the TerraSAR-X and PRISM imagery contained errors of on average around 23 meters. As part of the Urban Atlas project, an intensive manual co-registration was carried out to reduce these errors to within 3 meters.

To simplify the process of creating and using the dataset, the SAR imagery is resampled to a spatial resolution of 2.5m. These accurately aligned SAR and optical images then served as the basis for the creation of a high-resolution SAR-optical correspondence dataset which is suited for deep learning applications. This accurate alignment and ortho-rectification means ground-level points can be assumed to be corresponding. With this in mind, a Harris corner detector was applied to the optical imagery to create

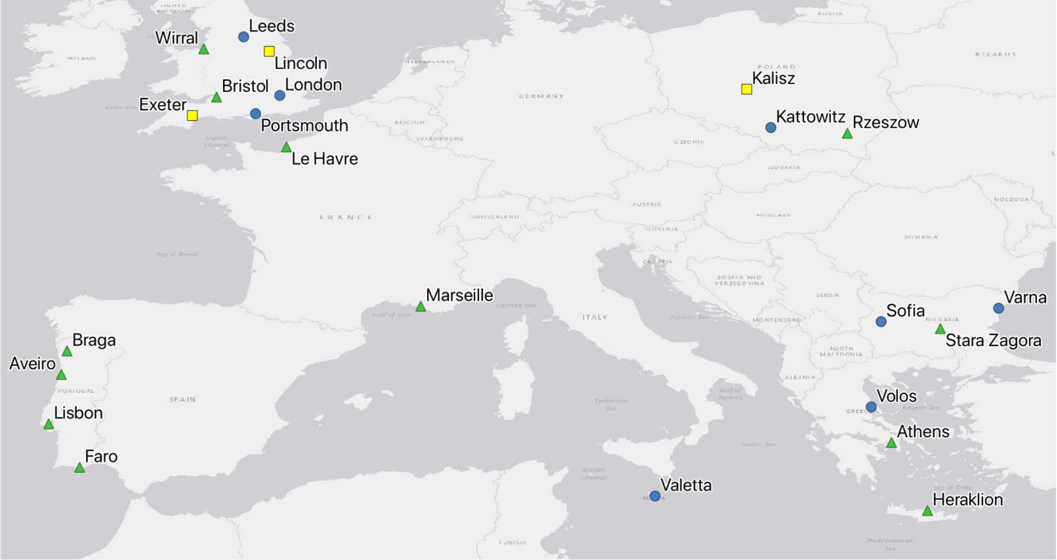


FIGURE 3.3: The spatial distribution of cities in the Urban Atlas dataset. The cities for the creation of the training, validation and testing datasets are depicted as green triangles, yellow squares and blue circles respectively.

an initial set of points which contain salient features, in at least in one modality. Non-maximal suppression was then used to reduce the overall number of points and ensure minimal overlap between the final patches. As the assumptions about correspondence only hold for ground-level points, OpenStreetMap (OpenStreetMap contributors, 2017) data was used to filter the point set by excluding points which were too near to building footprints, forests or other raised structures. After filtering, patches of size 256 × 256 pixels were extracted around each point in the SAR and optical images. Finally, the optical imagery was normalized to a range of $[0, 255]$, while the SAR backscatter was converted to Decibels and clipped to the global 3σ range, $[10, 30]$dB. This lead to a final dataset of 50,872 corresponding patch pairs, split into 40,314 training pairs, 4,205 validation pairs and 6,353 testing pairs, respectively. Some exemplary patch pairs are depicted in Figure 3.4, along with the suggested assignment of train, test and validation scenes.

While this high-resolution dataset is still based on the assumption of ground-level correspondence, this assumption only affects the center pixel in each patch. Thus the dataset captures a diverse set of scenes, including rural and dense urban areas, which makes it suitable for learning more generalizable models. However, due to the approach taken in creating the dataset, there are no guarantees that the SAR patches contain salient features. Although this can affect the learning process, the filtering based on OpenStreetMap features (i.e. roads and railways), combined with the large spatial extent of the patches significantly reduces the likelihood of the corresponding SAR patches containing no salient features.

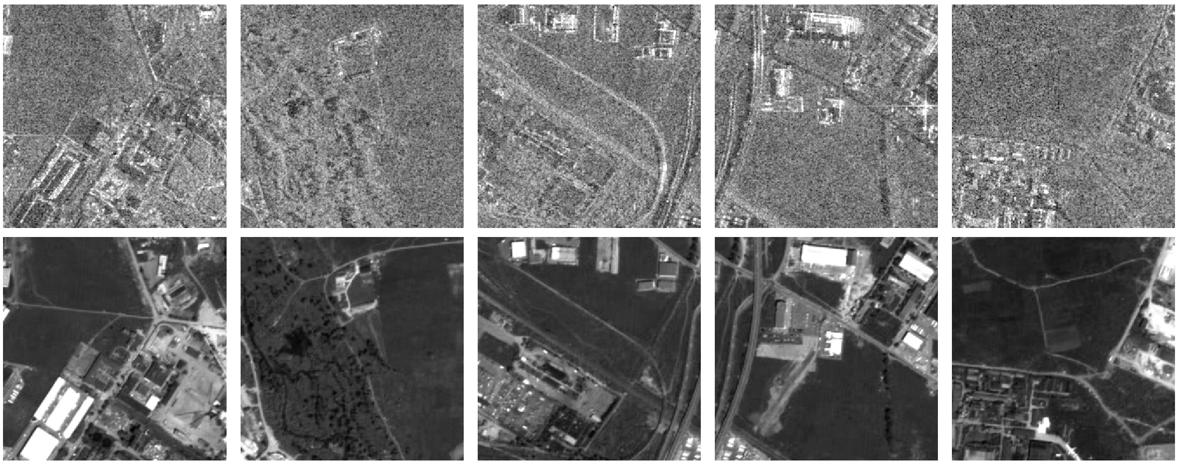


FIGURE 3.4: Some exemplary patch-pairs derived from the UrbanAtlas dataset. Top row: TerraSAR-X patches, bottom row: PRISM optical patches. The ground sampling distance of the patches is 2.5m.

3.1.2 Pseudo-Siamese Architecture for SAR-Optical Similarity

The seminal deep matching approach proposed by Merkle, Luo, et al. (2017) was based upon the successes of two-stream networks within the realm of conventional image matching problems (Zagoruyko et al., 2015). However, matching using two-stream networks with shared weights, such as siamese architectures, is inadvisable as they are based in the assumption that the features of the imagery fall within a common manifold. This assumption does not hold true for SAR and optical imagery, which have vastly different radiometric and geometric properties.

Thus it was proposed by (Mou et al., 2017) to first transform the imagery to a common feature manifold using independent CNNs, before matching. Based on this, a pseudo-siamese network architecture with two identical, yet separate streams, and a spatially-aware fusion network was proposed within the frame of this thesis. This architecture constrains the network to learning meaningful representations, of the input SAR and optical imagery, which fall within a common manifold that is matchable by the fusion network. The proposed network architecture is depicted in Figure 3.5.

The architecture of each stream is based on the well-known VGG16 architecture, proposed by the Oxford Visual Geometry Group (Simonyan et al., 2015). Each stream is consists of a series of 3×3 convolutional kernels followed by batch normalization and Rectified Linear Unit (ReLU) activation. The use of stacked layers of small kernels, to describe larger receptive fields, increases the non-linearity and thus descriptive nature of the network. Padding is added at all stages throughout the network to preserve the spatial dimensions of the features maps through the convolution operator. Additionally, 2×2 max-pooling operators, with a stride of 2, are added at various points in the network to reduce the spatial dimensionality of the feature maps. The fusion network operates on the concatenation of the reduced and transformed SAR and optical feature maps, and consists of two consecutive convolutional laters, followed by two fully-connected layers. The convolutional layers follow the same 3×3 kernel structure, but make use of a stride of 2, rather than a max-pooling, to further reduce the spatial

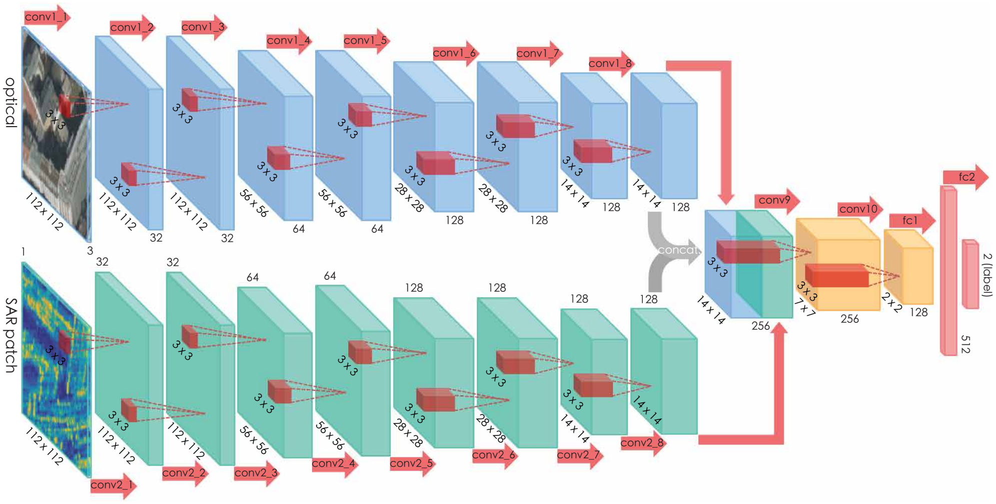


Figure 3.5: Pseudo-siamese CNN architecture for SAR-optical image matching. The optical stream is shown in blue, and the SAR stream in green, while the fusion network is depicted by the yellow convolutional layers and red fully connected layers.

dimensionality of the combined feature maps. The final stage of the fusion network consists of a 512-channel, followed by a 2-channel fully connected layer.

The network was trained as a one-hot encoded binary classification problem, using (non-)corresponding SAR-optical patch pairs, from a deterministically partitioned subset of the SARptical dataset (Wang et al., 2018), and a Binary Cross Entropy (BCE) loss. Thus the output of the final layer at inference time can be directly interpreted as a measure of similarity between the input patches.

When evaluated on an independent subset of the SARptical dataset, with a patch size of 112×112 pixels, the networks was able to achieve an accuracy of 77% with a fixed false positive rate of 5%. This performance was found to degrade rapidly if smaller patches were used, as depicted in Figure 3.6. Thus highlighting the importance of spatial context in SAR-optical matching. Exemplary results highlighting the performance of the pseudo-siamese network for matching high-resolution SAR and optical patches are presented in Figure 3.7.

While these first results showed great promise, further investigation highlighted significant shortcomings of the SARptical dataset and lead to the author questioning the suitability of this dataset for deep learning applications. Thus within the frame of this thesis the pseudo-siamese architecture was retrained and evaluated on the Urban Atlas datasets presented in Section 3.1.1. These results are presented in Chapter 5, within the scope of a comparative evaluation of the deep matching methodologies described as part of this thesis.

3.1.3 Multi-Scale Feature Space Matching

The pseudo-siamese architecture, as well as many other dual-stream SAR-optical matching approaches, are designed such that they can be used as a replacement for conventional similarity metrics such as NCC and MI. However, when applied in this manner,

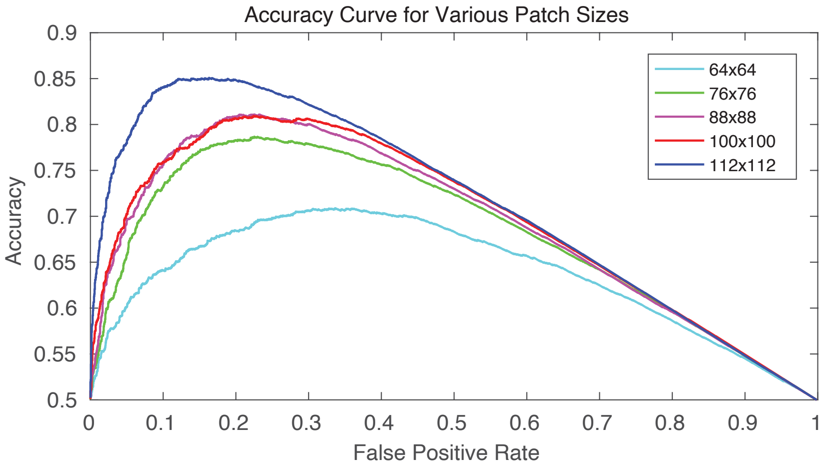


FIGURE 3.6: The performance of the pseudo-siamese architecture in relation to the size fo the input SAR and optical patches.

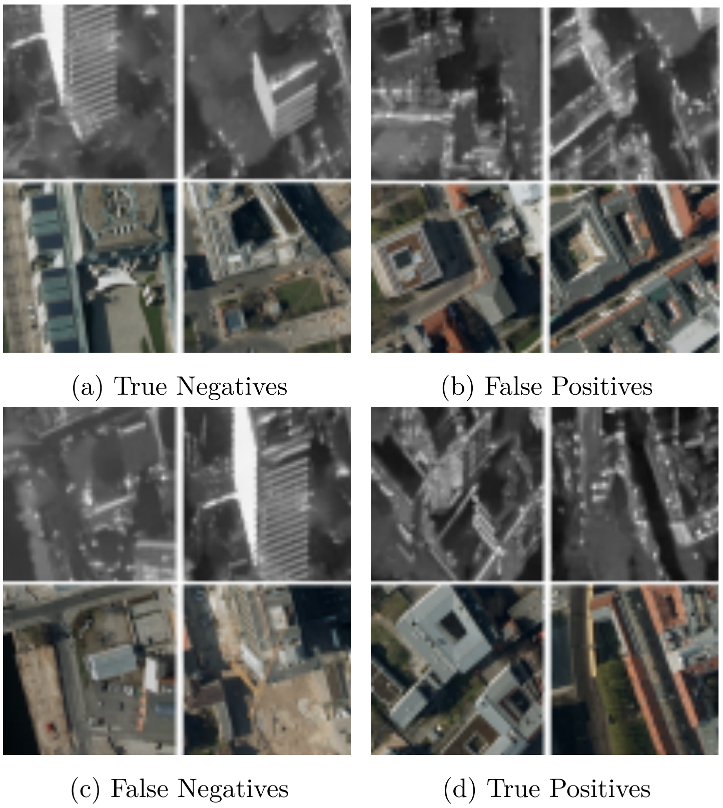
they are computationally expensive as they need to be evaluated at each offset within the search window, as described by Figure 2.6. Furthermore, existing approaches to SAR-optical matching primarily rely on features extracted from the final layers of deep CNNs. These features contain rich global semantic information; however, this comes at the cost of being low resolution and translation invariant. Thus it is argued that they lack the fine detailed features needed to capture minor offsets, which are imperative to determining accurate correspondence between high-resolution SAR and optical imagery.

Thus an alternative architecture which formulates the SAR-optical correspondence problem as a multi-scale search problem was proposed, herein referred to as the CorrASL network. Under this formulation, the goal of the network is to determine the most likely point of correspondence for the center pixel of an optical template patch within a larger SAR search region.

The architecture follows a pseudo-siamese design and is based around the concept of convolutional hypercolumns (Hariharan et al., 2015). Each stream consists of a feature extraction and feature reduction sub-network, and matching is performed using a feature-space correlation operator.

The modality specific hypercolumns are constructed by upsampling, using bi-linear interpolation, and stacking the feature maps extracted from each of the layers of the feature extraction network. The depth of the hypercolumn is then reduced to 256 channels, using a series of three 1×1 convolutional layers. To allow for the accentuation of salient features, a spatial attention map (as proposed by Woo et al., 2018) is created and applied to each hypercolumn. Finally, the hypercolumns are normalized along the channel dimension using L_2 normalization.

The template hypercolumn is then matched within the search hypercolumn using a feature space correlation operation with *valid* padding. The resultant correspondence map is then upsampled and zero padded to match the extent of the search window. The final output of the network is a heatmap, for which the maximum value represents the point of correspondence of the center pixel of the template patch within the search



(a) True Negatives

(b) False Positives

(c) False Negatives

(d) True Positives

Figure 3.7: Exemplary positive and negative predictions of correspondence, achieved using the proposed pseudo-siamese CNN.

region. The full architecture of the correspondence network, as well as the input and output datum, is depicted in Figure 3.8.

The network was trained using 256 × 256 pixel SAR search patches and cropped, 128 × 128 pixel, optical template patches from the Urban Atlas dataset proposed in Section 3.1.1. To prevent overfitting, and to better simulate conditions of misalignment, the optical template patches were cropped, from their respective full-size optical patches, using a random offset from the center pixel. A spatial softmax operator is then applied to the output heatmap, and a weighted mean squared error (MSE) loss is computed between the activated heatmap and a 2D Kronecker delta function. Additionally, an L_1 regularization term is included in the objective function to encourage sparsity in the correspondence heatmap. An example of the training inputs is depicted in Figure 3.9. Similarly, the process of determining correspondence from the network output is shown in Figure 3.10.

When evaluated on independent patch pairs extracted from SAR and optical imagery of 8 spatially diverse cities, see Section 3.1.1, the network was able to accurately (within 1 pixel) determine the point of correspondence 46.9% of the time with an average matching error of 2.1 pixels, and mean average precision of 2.62 pixels. Some exemplary matching results are depicted in Figure 3.11.

While the overall matching accuracy might initially appear low, it should be recalled that the approach used in creating the dataset could not provide guarantees of joint feature visibility. Thus the dataset is likely to include several SAR patches which

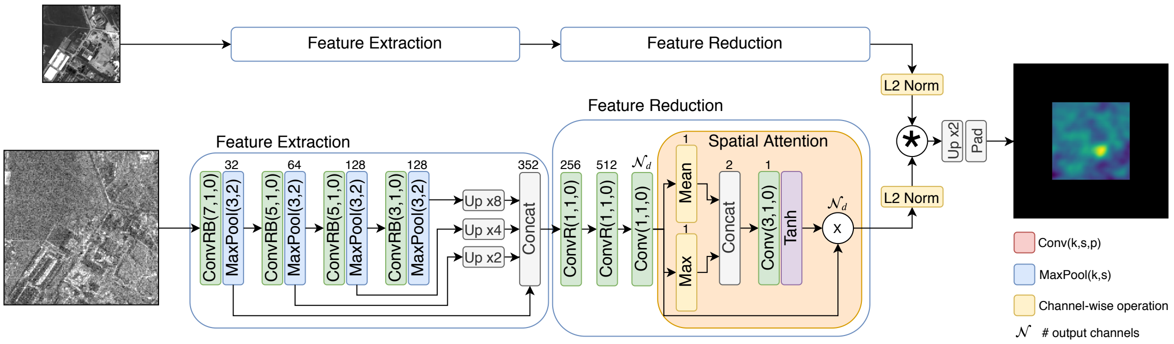
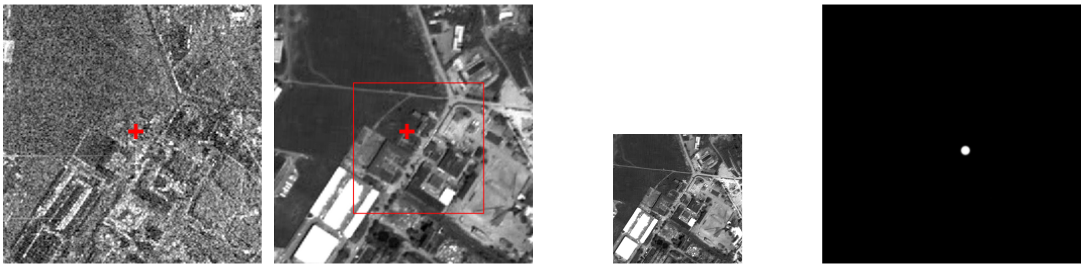


FIGURE 3.8: The network architecture showing the layer details for the SAR branch with Conv(k, s, p) and MaxPool(k, s), representing a convolutional layer, and pooling layer, with a kernel of size k, stride of s, and padding of p, respectively. Convolution followed by ReLU is represented as ConvR(k, s, p), and the addition of batch normalization as ConvRB(k, s, p).



(a) (b) (c) (d)

FIGURE 3.9: A single training sample created from the Urban Atlas dataset. (a) The SAR search patch cropped around the location of the optical Harris corner (represented by the red cross), (b) the optical patch from which the template patch (depicted by the red box) is extracted with a random offset during training, (c) The extracted template patch, and (d) the 2D Kronecker delta based ground truth label representing the true point of correspondence.

contain no salient features, and thus cannot be matched. However, as the network directly outputs a heatmap, it is theorised that the shape of the correspondence surface captured by the heatmap can be used to filter out failed matching attempts. This hypothesis was investigated within the frame of this thesis, and a deep learning-based solution is presented in Section 3.3.3.

3.1.4 Summary

In this section two large-scale SAR-optical datasets, the medium resolution SEN1-2 dataset and the high-resolution Urban Atlas-based dataset, were presented. These datasets were created with the application of SAR-optical deep matching in mind. However, their applicability goes beyond this application and they are suited for a multitude of SAR-optical deep learning-based data fusion endeavours.

Furthermore, two novel deep matching architectures were proposed for the matching

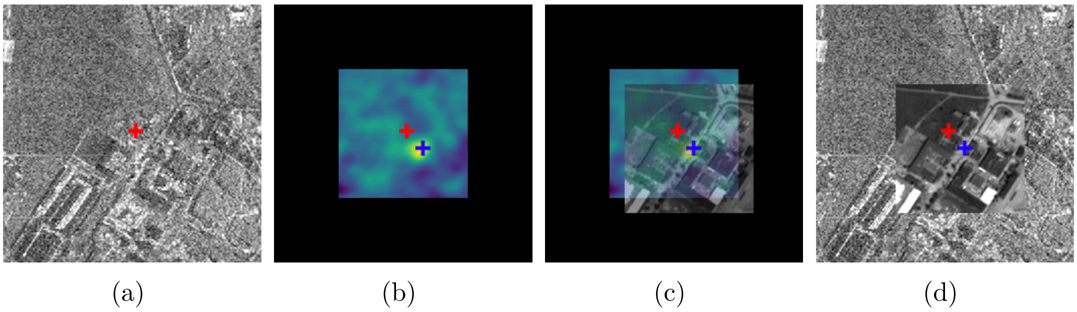


FIGURE 3.10: The process by which the correspondence heatmap can be used to determine the corresponding point for the center pixel of the optical template patch. (a) The search window with its center pixel marked by a red plus, (b) the resultant heatmap from the correspondence network with its center pixel aligned to that of the search window, and the peak point of correspondence depicted by a blue plus. (c) The center of the optical template patch is aligned to the peak point of correspondence, (d) the final alignment of the optical template patch, with the located point of correspondence marked by the blue plus.

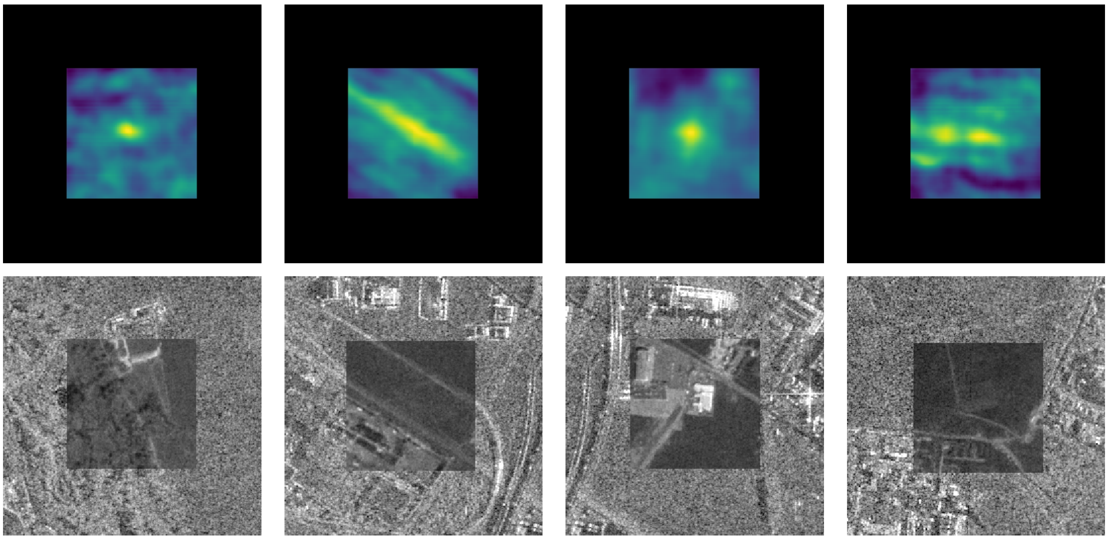


FIGURE 3.11: Accurately matched results achieved using the correlation-based matching network (CorrASL). The first row shows the correspondence heatmaps for each results. While the second row shows the optical template patch overlaid within the SAR search window by aligning the center pixel with the point of maximal correspondence.

of high-resolution SAR and optical imagery. The pseudo-siamese architecture framed the problem as a similarity metric problem whereby the network was trained to learn a generalizable similarity metric for comparing SAR and optical image patches. In this configuration the pseudo-siamese architecture is best suited towards feature point matching applications, whereby the assumption is made that an intersection exists between the sets of detected feature points in each modality, or that an ideal multi-modal feature point detector exists. As at this time neither of these assumptions is robust enough for operational application, thus the pseudo-siamese network is best applied as a replacement for conventional similarity metrics in a search-based matching framework. However, when applied in this manner it is computationally expensive, which can be prohibitive to matching across large regions.

To combat some of the short-comings of pre-existing approaches, as well as the pseudo-siamese approach, an alternative multi-scale feature space deep matching architecture was proposed. This architecture is based on multi-scale convolutional hypercolumns and incorporates the search process as part of the matching network through the inclusion of a feature-space correlation operator. Thus the architecture is much more computationally performant than the pseudo-siamese architecture. However, this formulation of the matching problem relies strongly on assumptions that the point of correspondence is within the search window, and that the supporting region is unambiguous. Furthermore, the pixel values in the produced correspondence heatmaps do not represent an absolute score of similarity, but rather a relative one. This means the maximum value of the heatmap alone cannot be used to determine if the match was successful or not, unlike the in the heatmaps produced when applying the pseudo-siamese network.

3.2 Matching with Scarce Data

Peer-Reviewed Publications Related to this Section

Hughes, L. H., & Schmitt, M. (2019). A semi-supervised approach to SAR-optical image matching. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *IV-2/W7*, 71–78.

Hughes, L. H., Schmitt, M., & Zhu, X. X. (2018). Mining hard negative samples for SAR-optical image matching using generative adversarial networks. *Remote Sensing*, *10*(10), 1552.

The large diversity of high-resolution SAR and optical sensors available, and the complexity involved in the creation of large-scale labelled datasets, means the learning of generalizable matching strategies for SAR-optical imagery is likely to remain a scarce data problem for many years to come. While this statement might appear contradictory in relation to the new datasets and methods presented in Section 3.1, it does not negate the value of these approaches. However, methods trained on a dataset containing only data from one optical and one SAR sensor are unlikely to be able to generalize to data from sensors with different properties. This is both due to the visible differences between the modalities becoming more extreme as spatial resolution increases, as well as due to the inherent nature of learning models from data which restricts the applicability of those models to the same data manifold. Thus it is important to investigate the effects of small data on matching, and develop techniques which enable the learning of useful models in the presence of small data.

To this end this section presents two novel deep learning-based approaches for application in small data environments. The first approach is focussed on reducing over-fitting and improving performance of models when limited labelled training data is available,

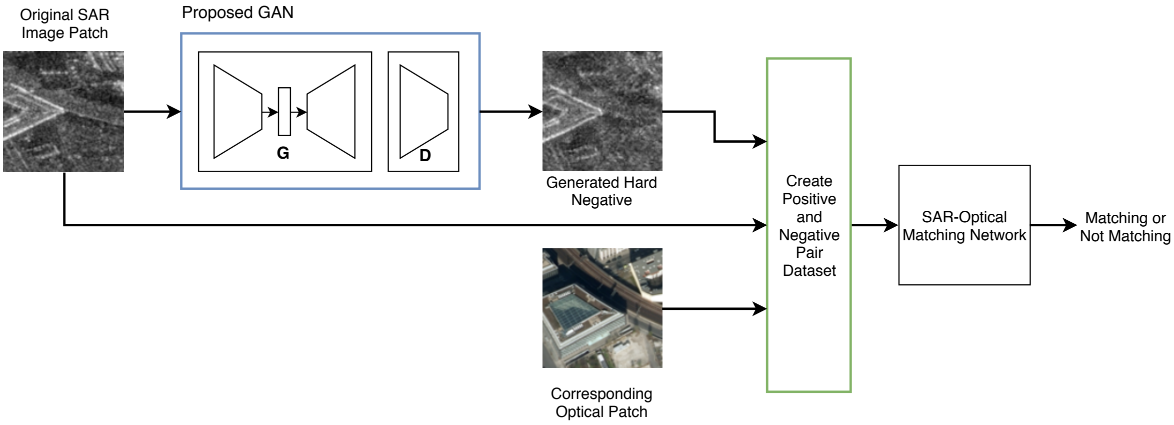

Original SAR Image Patch
Proposed GAN
G
D
Generated Hard Negative
Create Positive and Negative Pair Dataset
SAR-Optical Matching Network
Matching or Not Matching
Corresponding Optical Patch

FIGURE 3.12: The proposed generative hard-negative mining framework. The GAN is trained to create hard-negative samples based on the input image patch. Together with the original corresponding optical patch, these samples are then used to train the SAR-optical matching network.

while the second approach is aimed at exploiting the vast archives of unlabelled training data to train a deep matching architecture with a limited number of high accuracy labels.

3.2.1 Hard Negative Mining with Generated Samples

Networks trained on small datasets tend to overfit the training data, and thus lose their ability to generalize to unseen data, such as the test set. This effect, in turn, leads to a large number of false-positive (FP) or false-negative (FN) results in the final classification output, as the network is biased towards one prediction output. In the domain of SAR-optical matching, the reduction of FP correspondences is an essential task as many downstream data fusion tasks are more sensitive to incorrect correspondences than to a small number of correspondences.

A common approach to improving the discriminability between classes in classification tasks, and thus reducing the false-positive rate (FPR), is hard negative mining (Felzenszwalb et al., 2008). However, the standard implementation of hard negative mining relies on the assumption that during training, as the FPR decreases, sufficient negative samples exist that hard negative samples can still be mined. While this is generally true for large datasets, in the case of small datasets, this assumption breaks down quickly, and the training procedure collapses back to the random selection of negative samples. Thus a generative approach to hard negative mining was proposed. In this approach, a Generative Adversarial Network (GAN) architected such that it could be trained to generate the SAR patches in the training set. The continuous latent space created through this training procedure could then be sampled to create a set of hard negatives. The overall framework is depicted in Figure 3.12.

To enable the generation of high-resolution samples which are slightly augmented versions of an input sample, three main functions needed to be achieved by the GAN. Firstly, a GAN architecture and training mechanism capable of producing high-fidelity imagery is required. Secondly, the output of the GAN needs to be conditioned on the input imagery. Finally, the latent space produced by the conditioning needs to be locally smooth such that it can be sampled to allow for output to create variations of the input imagery.

To this end, the ProGAN architecture proposed by Karras et al. (2018) was extended, by converting the generator network into a Variational Autoencoder (VAE). In this manner, the input image patch is used to learn a smooth latent distribution for the training data, which is then sampled to obtain a nearby latent vector. This latent vector is then used by the generator to create a novel variant of the input patch which is suitable for use as a hard negative sample. This style of network architecture is also commonly known as a VAEGAN or Adversarial VAE (Larsen et al., 2016).

The decoder network of the VAE is directly based off of the ProGAN generator network, and consists of a fully-connected bottleneck layer followed by a series of convolutional modules. Each convolutional modules consists of a nearest-neighbour upsampling layer, a convolutional layer, a leaky ReLU (LReLU) activation function and a pixel-wise feature vector normalization layer (PixelNorm) (Karras et al., 2018).

The encoder network is constructed by mirroring the structure of the previously described decoder network. Thus the upsampling layers are replaced with average pooling layers which act to downsample the feature maps. Additionally, two fully-connected layers are added to the head of the network, after the existing bottleneck layer, which are used to represent the latent space as a mean and standard deviation. These layers are required for imposing a prior distribution on the latent space, as well as to reparameterize the sampling operation such that it is differentiable.

The discriminator has an equivalent structure to the encoder network, with two minor adaptations. The first is the inclusion of a mini-batch standard deviation layer, which adds an additional feature map to the second last convolutional layer. This was done in order to increase variation in the network, and prevent overfitting of the discriminator. The second modification is the replacement of the bottleneck layer with a fully-connected layer which reduces the output of the last convolutional layer to a single scalar value. This scalar value represents a score of the ’realness’ of the evaluated image.

The network was trained using the progressive growing procedures proposed by Karras et al. (2018) and the dual GAN and VAE losses defined by Larsen et al. (2016). The training procedure started with an image resolution of 4 pixels and gradually increased this by a factor of 2 as the losses stabilized at the current resolution. This process was continued until the final resolution of 128 pixels was achieved. The use of subnetworks which share a common number of layers with a similar structure reduced the complexity involved in transitioning between resolutions during training. An example of the outputs of the generator network, as the training progressed, are depicted in Figure 3.13.

In order to use the proposed generative network to create hard negative samples, it is trained on all the SAR patches contained in the dataset which is to be used to train the final matching architecture, in this case, the pseudo-siamese architecture presented in Section 3.1.2. In doing so the encoder learns the latent distribution of the training

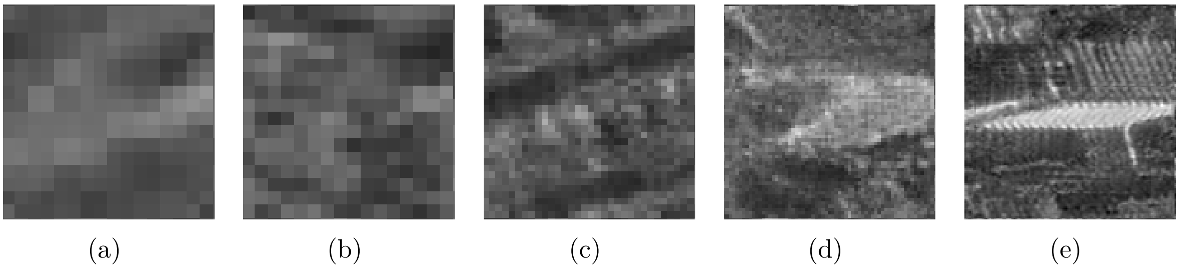


FIGURE 3.13: An example of progressively grown images taken at increasing image resolutions during the training processes. The images have a resolution of (a) 8 × 8 pixels, (b) 16 × 16 pixels, (c) 32 × 32 pixels, (d) 64 × 64 pixels and (e) 128 × 128 pixels, respectively.

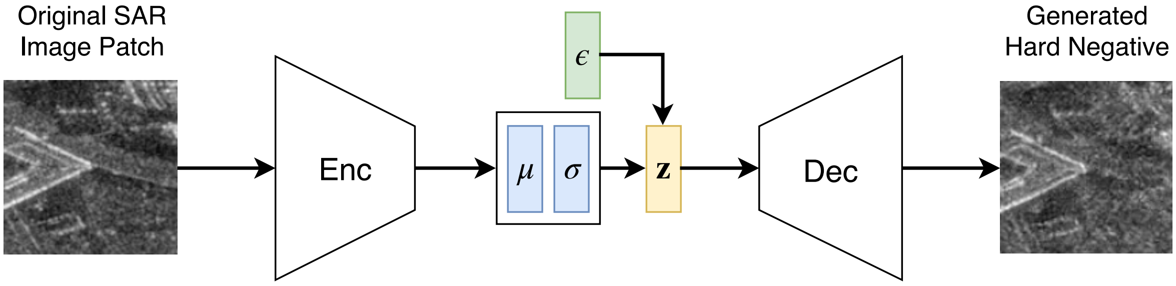


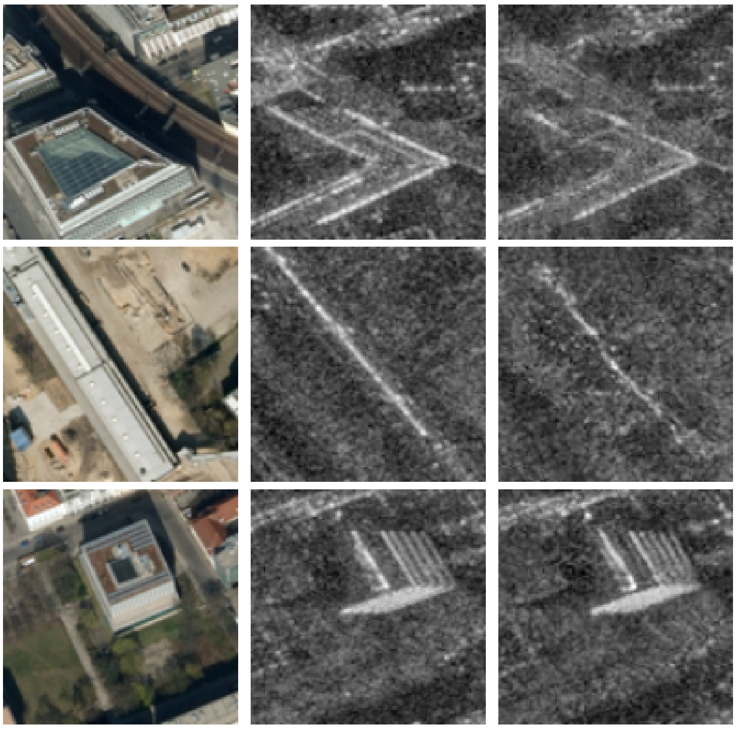
FIGURE 3.14: The inference network used to generate hard negative samples. The position of the input patch within the latent space is parameterized by a Gaussian distribution, this distribution is then sampled to create a latent code $\mathbf{z}$ which describes a nearby, but novel data point.

samples and the decoder learns to reconstruct the input imagery from this distribution. As the assumption is that the dataset is small, the network will likely overfit the training set. However, as the goal is not to generate completely novel samples, this is of little consequence within this application and can in some cases be beneficial to the realism of the generated samples.

Post training, the discriminator network is discarded and the VAE is used to generate hard negative SAR samples for each SAR image patch in the original training dataset. As the latent space is continuous and follows a unit normal distribution, a slightly augmented version of the input patch can be created by sampling the latent distribution near to the location of the encoded input patch within the latent space. This process is depicted in Figure 3.14.

The set of generated SAR-like images are then combined with the original dataset, such that each corresponding SAR-optical patch pair is extended to include a generated SAR-like hard-negative sample, which is labelled as non-corresponding. A few samples of an extended dataset are shown in Figure 3.15.

Finally, the extended dataset is used to train the SAR-optical deep matching network. This is done by combining the appending the extended dataset to the original training dataset, such that each training sample appears twice; once with a random non-corresponding SAR patch and once with a generated non-corresponding SAR-like patch. Table 3.1 summarizes the results of applying GAN based hard negative mining



(g) Optical (h) SAR (i) Generated

FIGURE 3.15: Generation of SAR-like image patches with the hard negative GAN. The original corresponding pair of (a) optical and (b) SAR patches is extended to include a (c) generated hard negative sample.

TABLE 3.1: Details of SAR-optical matching results under the application of various hard-negative training strategies and at different false positive rates (FPR)

Method	**Precision**	**Recall**	**Acc.** (5% FPR)	**Max Acc.**	**FPR** (Max Acc.)
Random	**0.83**	0.84	0.76	0.83	0.16
Nearest Neighbour	0.77	**0.96**	0.70	0.85	0.21
Traditional Hard Neg.	0.79	0.89	0.72	0.83	0.19
Proposed Approach	**0.83**	0.87	**0.81**	**0.86**	**0.13**

to the training of the pseudo-siamese architecture presented in Section 3.1.2. The results clearly show that generative hard negative mining leads to a significant decrease in the false positive rate (or an increase in accuracy when fixing the FPR to 5%) over the baseline as well as alternative negative mining strategies. Thereby, highlighting the value of the proposed approach in training robust classifiers with scarce data.

3.2.2 Semi-Supervised SAR-Optical Matching

Although the hard-negative mining approach significantly improves the robustness of models learnt on small data, it does not enhance their generalizability. The reason for this is that models learnt directly from data are only as diverse as the dataset

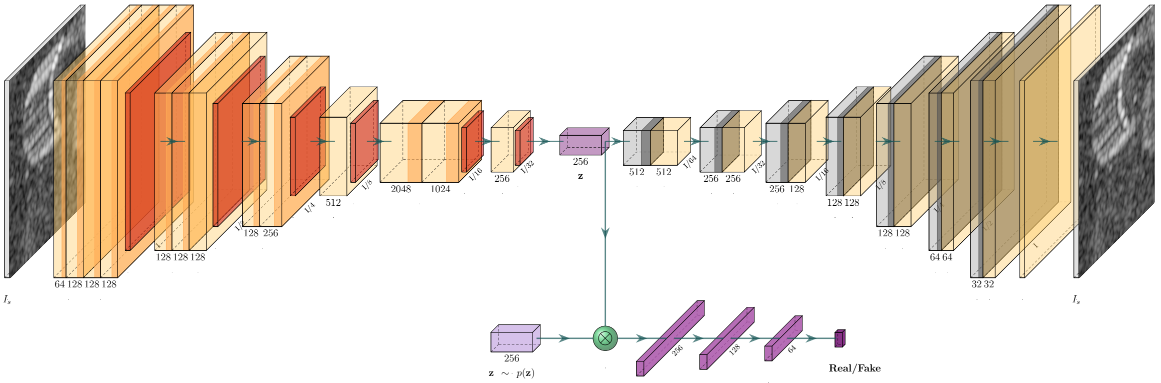


FIGURE 3.16: The SAR stream of the proposed semi-supervised matching network architecture. The autoencoder learns to generate a diverse latent space **z**, through self-supervised reconstruction of the input. The discriminator network is used to condition the latent distribution to an arbitrary prior, using an adversarial training scheme. The optical stream mirrors the SAR stream, and the discriminator is shared between the two.

itself. Thus to learn more robust, and diverse models under small data conditions, the SAR-optical matching problem was reformulated as a semi-supervised learning task.

Semi-supervised learning constitutes a set of techniques for exploiting stores of unlabelled data to support learning diverse models in data constrained environments (Chapelle et al., 2006). Thus the goal behind this reformulation is to utilize the vast stores of unlabelled data, which exist in the domain of remote sensing, to support the learning of diverse and generalizable SAR and optical image representations. In this manner, the SAR-optical matching task can be learnt, based on these representations, using a limited number of labelled training samples.

Taking inspiration from the supervised matching networks developed by Liu et al. (2018) and Mukherjee et al., 2017, a dual autoencoder (AE) architecture was proposed. This architecture allows for the use of unlabelled and unpaired data to train the domain-specific autoencoders. Furthermore, the latent code generated in the AE bottleneck is a natural feature descriptor which is used for the SAR-optical matching task. Alignment between the SAR and optical latent spaces is achieved using a supervised matching task, based on a small dataset of labelled correspondences, and a joint adversarial loss which is implemented using a discriminator network. The overall structure of the architecture is depicted in Figure 3.16.

The encoder network is based on the VGG11 (Simonyan et al., 2015) architecture, and thus follows the structure of the pseudo-siamese architecture presented in Section 3.1.2. Similarly, it consists of blocks of 3×3 convolutions, batch normalization, LRelU activation and max-pooling. The decoder network mirrors the encoder network shape, but instead consists of blocks of 3×3 transposed convolutions, for upsampling the feature maps, followed by 3×3 convolutional layers, and ReLU activation throughout.

The autoencoders were trained on alternating batches of labelled corresponding images pairs, and unlabelled image pairs. For batches containing unlabelled data, the training was supervised by a MSE reconstruction loss computed between the input image and the image reconstructed from the generated latent code. In doing so the aim was to ensure the latent code could accurately represent the key features of the input image. For labelled batches an additional matching loss was included as part of the training process. The matching loss took the form of a contrastive matching loss (Hadsell et al., 2006), computed between the pair of latent codes generated by the (non-)corresponding input pair. The contrastive loss encourages the network to learn a latent space where corresponding pairs are near to each other, while non-corresponding pairs have a squared norm distance of at least m (Chopra et al., 2005).

While this training strategy encourages the network to align the latent space for corresponding samples, it does not guarantee the smoothness or alignment of the manifold beyond these samples. Thus a discriminator network was introduced to impose a continuous prior distribution, a multivariate normal distribution in this case, on both the SAR and optical latent spaces (Makhzani et al., 2016). The discriminator network consists of three fully-connected layers, two of which are followed by a LReLU activation, while the final layer uses a sigmoid activation. A single discriminator network is shared between the SAR and optical AEs, such that it further encourages the latent spaces to follow a similar distribution.

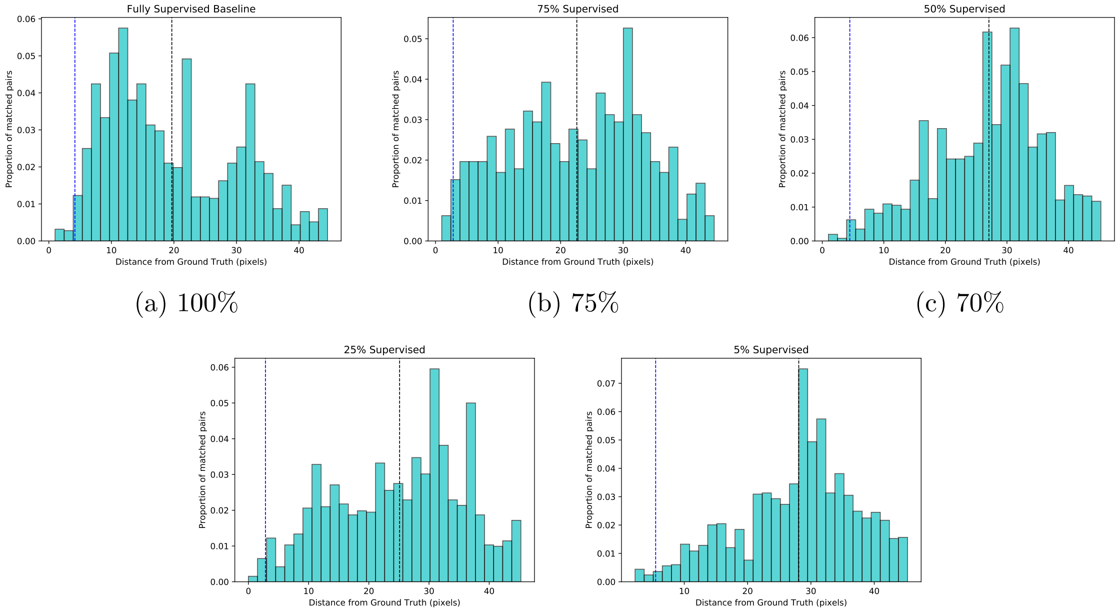
The encoder and discriminator network form the bases of a GAN, whereby the *real* samples are drawn from a prior distribution and the *fake* samples are made up of the latent codes generated by the SAR and optical encoders. The generative loss was included in the autoencoder loss function used to supervise the encoder network while the discriminator was trained independently on alternating batches.

The networks were trained using the 128×128 pixel patches extracted from the high-resolution Urban Atlas dataset, presented in Section 3.1.1. The training dataset was split into supervised and unsupervised subsets with varying amounts of supervised data, namely 100%, 75%, 50%, 25% and 5%. After training the decoder and discriminator networks are discarded, and the encoder networks are used to match an optical template patch within a SAR search window using the cosine-distance between the descriptors at each location.

The matching performance for each scenario is depicted, in Figure 3.17, as histograms/density functions of the pixel distance between the detected point of correspondence and the ground truth location. Furthermore, Figure 3.18 depicts matching heatmaps obtained over a variety of scenes containing varying building density and difficulty.

From Figure 3.17 it is clear that the 1-percentile performance of all the approaches is relatively similar. However, beyond that it is clear that a non-linear relationship exists between the amount of supervision and the accuracy of the obtained matches. This is further clarified by the smoothness, and consistency of the heatmaps presented in Figure 3.18. The heatmaps for the 50% supervision task are significantly noisier than in the case of 25% or 5% supervision, which more often manage to obtain the correct point of correspondence for the optical template.

At first glance, this behaviour seemed counter-intuitive, however, an analysis of the literature (Dai et al., 2017) lead to the hypothesis that the unsupervised reconstruction loss and supervised matching loss are orthogonal to some degree. Thus, by optimizing



(a) 100%

(b) 75%

(c) 70%

(d) 25%

(e) 5%

Figure 3.17: Histograms reflecting the precision of the determined matched point when compared to the ground truth location for varying degrees of supervision. The dashed black line represents the mean matching distance while the dashed blue line represents the 1-percentile matching distance.

for both losses in the baseline method, the network ends up in a local minimum which is not necessarily well suited to either task. Furthermore, the reduction in labelled data can be interpreted reweighting the supervised and unsupervised terms of the loss function, thus explaining why, in some cases, lower amounts of supervision appear to provide better matching results.

Although the overall accuracy, and number of matching is lower than that achieved for the multi-scale matching approach presented in Section 3.1.3, the overall result still provides key insights. Firstly, that semi-supervised techniques, with very sparse data, are still able to produce successful matches. Secondly, it highlights the difficulties of matching high-resolution SAR and optical data without spatial context (i.e. using only feature descriptors). Finally, it reinforces the hypothesis that the heatmaps produced for successful matches appear to be visually separable from those of failed matches, and thus could be filtered out using deep learning approaches.

3.2.3 Summary

In this section, two different approaches to enabling matching under scarce data were presented. The first approach presented a generative approach to the problem of hard negative mining. In doing so, artificial SAR images were used to extend a small training dataset with hard negative samples, such that the network learned more robust decision making features. This provided a significant improvement to the robustness of the pseudo-siamese matching network, trained on a small dataset, without affecting the

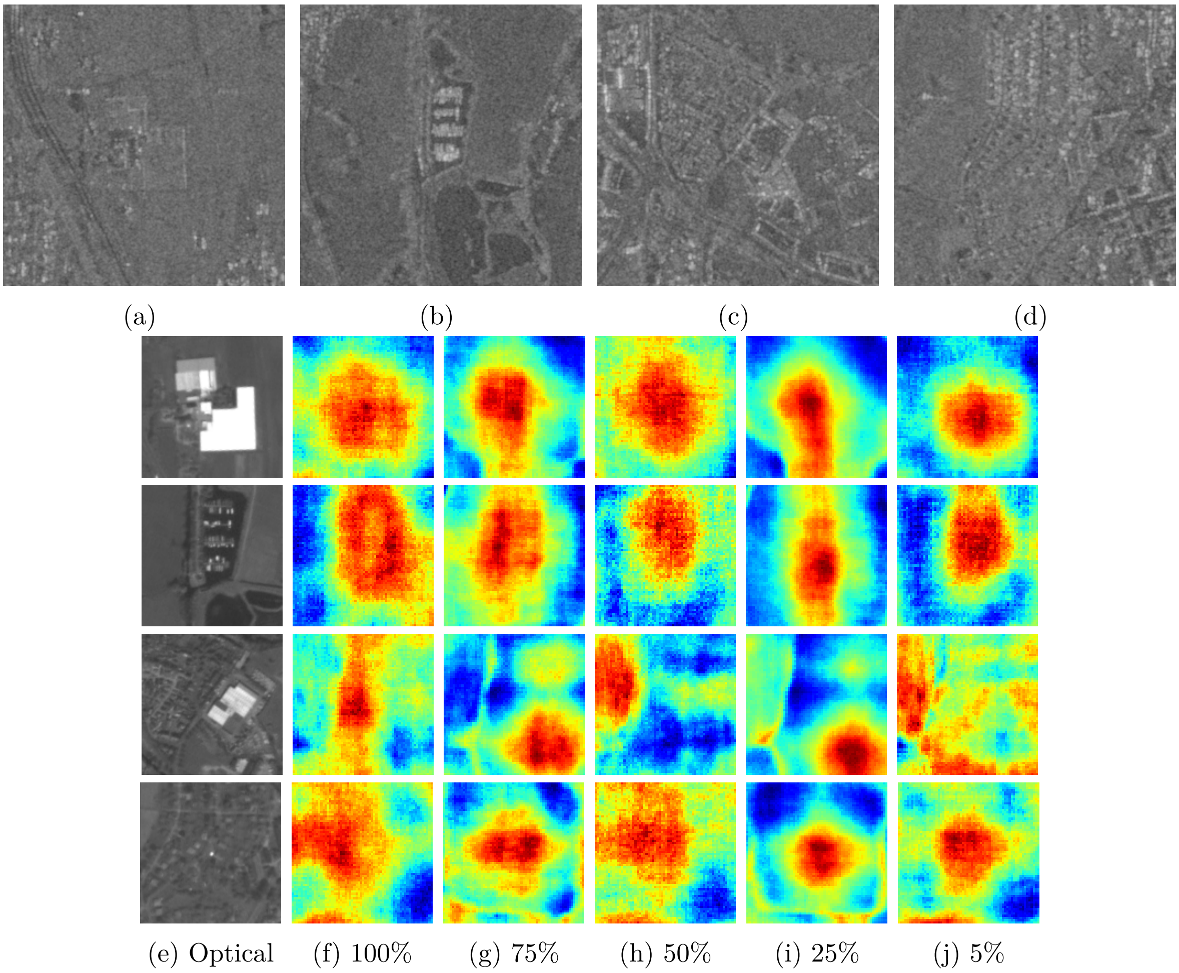


Figure 3.18: Correspondence maps produced under varying conditions of data scarcity, on example scenes of differing density. **(a-d)** exemplary SAR test scenes, corresponding rows depicting **(e)** optical image patch, and **(f - j)** correspondence maps when trained with supervision percentage of 100%, 75%, 50%, 25% and 5% respectively.

overall accuracy. However, this method should not be interpreted as a mechanism for improving network generalization, but rather as a means to enhance the performance of a specialized network (a network trained and applied on a specific set of data, in a particular region).

The second approach focussed on evaluating the use of unlabelled data to learn generic feature representations, which could be jointly trained for matching using a small number of labelled samples. Although the network performance was limited, even when trained in a fully supervised manner, some key observations were made. Firstly, it was shown that even with minimal training data, the network is still able to determine correspondences between certain patches (mainly in sparse suburban regions) with reasonable accuracy. Secondly, the complex dynamics between unsupervised and supervised learning tasks were uncovered, and the incompatibility between features required for reconstruction and those needed for matching were highlighted. Finally, the hypothesis that correspondence heatmaps hold essential information for the automated removal of outliers was further corroborated.

Due to the need for additional supporting networks, these approaches are significantly more computationally intensive to train than fully supervised matching networks. However, they provide useful mechanisms for achieving/supporting SAR-optical matching in when obtaining additional labelled training data is intractable due to cost, time or accessibility constraints. Furthermore, these techniques are applicable beyond the scope of SAR-optical matching and could provide useful insights and mechanisms to other fields suffering from scarce data, or complex labelling tasks.

3.3 A Comprehensive Framework for SAR-Optical Matching

Peer-Reviewed Publications Related to this Section

Hughes, L. H., Auer, S., & Schmitt, M. (2018). Investigation of joint visibility between SAR and optical images of urban environments. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *4*(2), 129–136.

Hughes, L. H., Marcos, D., Lobry, S., Tuia, D., & Schmitt, M. (2020). A framework for sparse matching of SAR and optical imagery [Under Review]. *ISPRS Journal of Photogrammetry and Remote Sensing.*

While the SAR-optical matching sub-task has been the focus of much previous research, both as part of this thesis as well as in the broader community, it does not constitute the entire matching process. In general, matching methodologies need to be initialised with feature points or search regions to make the matching task computationally feasible and improve the likelihood of finding corresponding points. The quality of these feature points, or regions, is thus critical to the success of the matching process, as low-quality initialisations will inevitably lead to a higher number of inaccurate correspondences. Furthermore, even the most sophisticated matching methodologies will produce outliers in the form of incorrect, or inaccurate correspondences. Thus the addition of an outlier removal task is of equal importance in guaranteeing the success of the matching subtask.

Due to the inherent complexities involved in determining jointly visible features across vastly heterogenous data sources, such as SAR and optical imagery, previous matching pipelines have almost exclusively relied on features detected in a single modality. These approaches make the assumption that specific types of features have a higher likelihood of being jointly visible in both modalities, and are thus often limited to a single type of feature such as road intersections (Merkle, 2018) or strong scatterers (Bürgmann et al., 2019). Although these assumptions hold in some cases they are an over simplification of the factors involved in a feature being jointly visibility, and their hand-crafted nature limits their scope of application to specific types of scenes.

To this end an investigation into joint feature visibility is conducted, in order to determine the expected upper bound of feature visibility under typical image acquisition geometry. Based on these insights a novel approach to determining regions which are likely to contain jointly visible salient features is proposed. Furthermore, as outliers are inevitable in any matching approach, a deep learning-based approach to outlier detection is proposed which operates directly on correspondence heatmaps to determine the probability successful matching. Finally, these two novel mechanisms are combined with a correspondence network, see Section 3.1, in order to create a comprehensive matching framework which allows for the determination of correspondences across large and diverse scenes without strong assumptions about scene or acquisition geometry.

3.3.1 Joint Feature Visibility

Although joint feature visibility is affected by both radiometric and geometric differences between SAR and optical imagery, it is the geometric differences which are hardest to compensate for in the feature detection process. The reason for this is that the unknown scene geometry plays a significant role in which features are jointly visible. Thus before deciding on a feature detection strategy an investigation into the expected upper bound of joint visibility was performed, to gain an understanding of the various contributing factors and the expected types and density of detectable features. In this context the term *joint visibility* refers to the number of pixels which are imaged and corresponding in both the SAR and optical modalities.

In order to perform a detailed analysis of a scene in terms of joint visibility, a framework for pixel-wise modelling and interpretation of a scene was required. The SimGeoI framework (Auer et al., 2017) is an object-level simulation framework which enables automated alignment and interpretation of SAR and optical remote sensing images. Thus extensions to this framework were made to enable pixel-level alignment and simulation of the scene. To achieve this an additional, iterative pixel modelling and ray-tracing procedure was added to the original SimGeoI pipeline. These extensions are depicted in Figure 3.19.

As depicted in Figure 3.19, the pixel-level simulation occurs by modelling each pixel in a normalized Digital Surface Model (nDSM) by a small sphere in world coordinates according to its relative (X, Y) coordinates and respective DSM height. Each sphere model is then independently projected into the image frame using a modality specific ray-tracing engine, with the camera defined according to the SAR or optical sensor metadata as per the standard SimGeoI procedure. The location of the sphere in world coordinates, as well as its projected position in the sensor specific image frame is then used to sample the various interpretation layers generated during the object-level simulation. In doing so the image to DSM correspondence for each pixel, in each modality is known, along with its modality specific interpretation class. Based on this data, and a set of filtering operations, it is trivial to determine pixel-wise correspondences and visibility between the SAR and optical image domains. In doing so cross-modal and joint visibility maps can be produced for a scene.

Cross-modal visibility maps indicate which pixels in the present modality are not visible in the corresponding modality. For instance, the optical cross-modal visibility map would depict which pixels in the optical domain correspond with radar shadow in the SAR domain, and similarly for the SAR cross-modal map. While the joint visibility

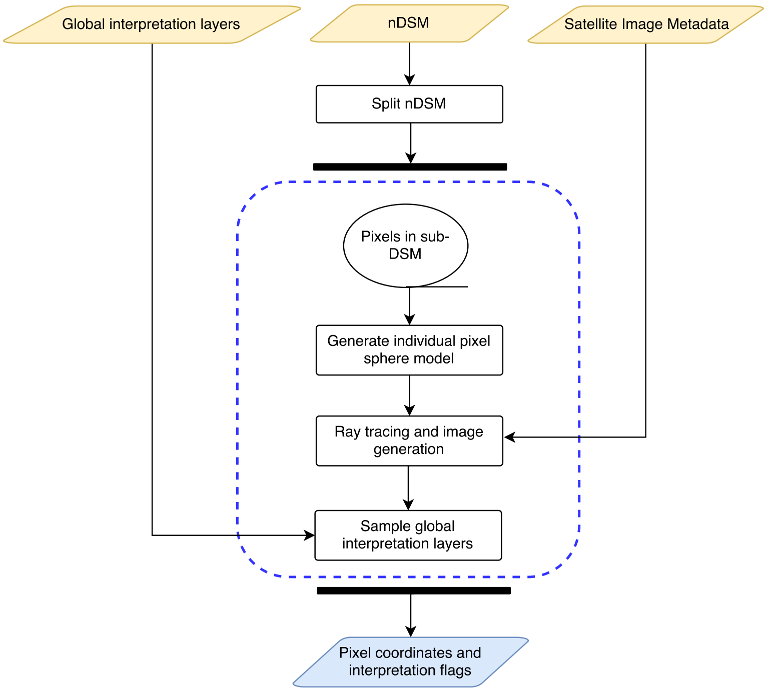

Global interpretation layers
nDSM
Satellite Image Metadata
Split nDSM
Pixels in sub-DSM
Generate individual pixel sphere model
Ray tracing and image generation
Sample global interpretation layers
Pixel coordinates and interpretation flags

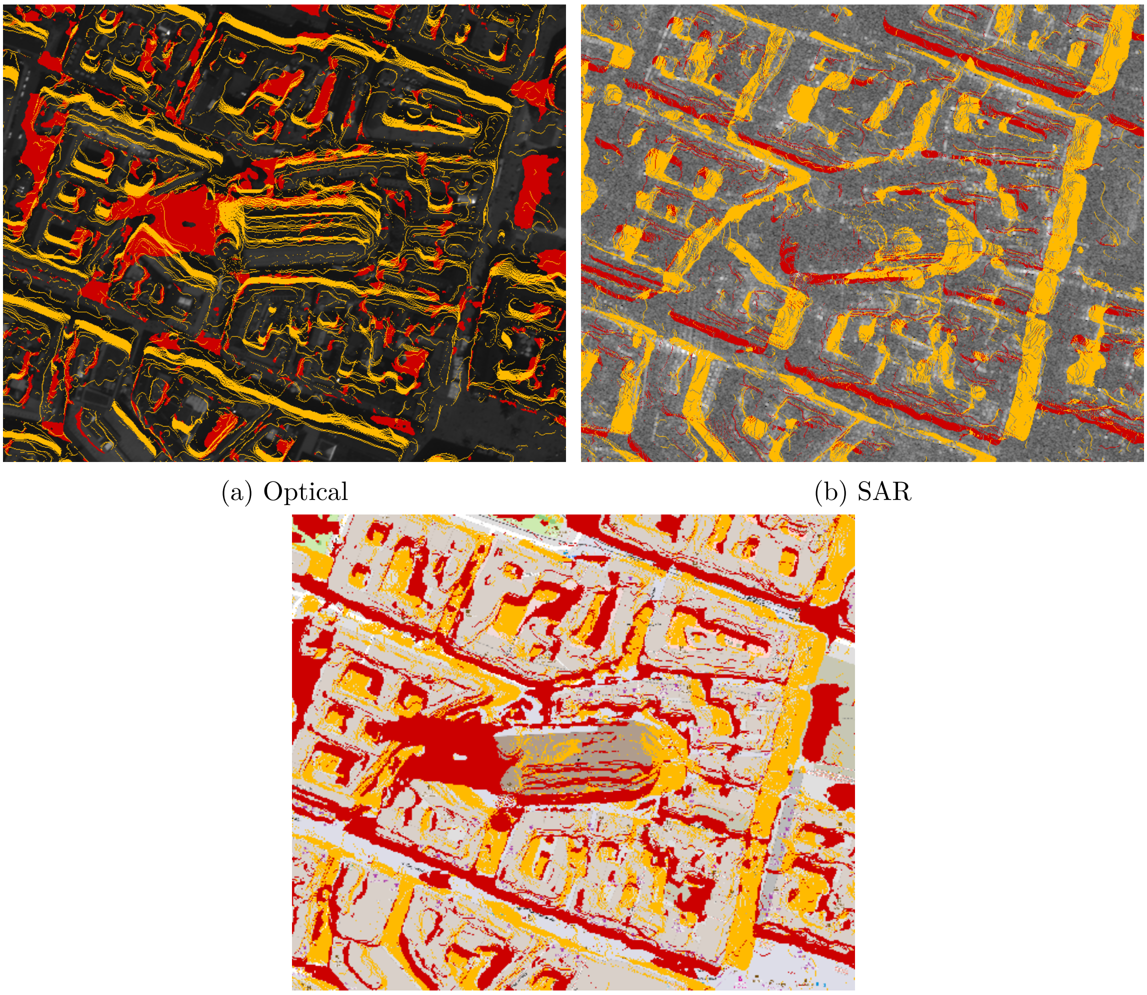
FIGURE 3.19: Extension of the SimGeoI framework to pixel-wise simulation and interpretation. The process encapsulated in the dashed blue area is run in parallel and is independent for each sub-DSM. The results are collated at the end, into a single results file for the specified image. Yellow: inputs which are obtained from the original SimGeoI pipeline in Fig. 1. Blue: Collated output file containing pixel-wise results for a single satellite image.

map combines these cross-modal maps to indicate which pixels are visible in both acquisitions. Examples of each, for a subset of a scene, are presented in Figure 3.20.

The red pixels in Figure 3.20 represent pixels which are not visible in both modalities due to geometric reasons. The yellow pixels in the visibility maps describe regions in the image whose visibility is dependent on the spatial relationship between the sensors and the geometric distortions which occur during imaging. All other pixels represent regions which are jointly imaged by both sensors, and are unlikely to be subject to geometric distortion.

Based on these maps a quantitative estimate of the upper bound for joint feature visibility was carried out by determining the ratios of visible and non-visible pixels. Thus as urban regions constitute the most complex geometric differences between SAR and optical modalities and analysis was run on typical inner city scene, based on Munich, Germany. The results of this analysis are presented in Table 3.2.

Thus based on purely geometric constraints it can be seen that the upper bound of feature visibility in a typical urban environment, is relatively low - being around 55% in an urban environment which largely consists of mid-rise buildings. Furthermore, it can be seen that building rooftops and flat ground regions constitute the regions of highest joint visibility and thus provide the most likely candidate regions for finding matchable features. However, it should be noted that the joint feature visibility is



(a) Optical

(b) SAR

(c) Joint Visibility

FIGURE 3.20: Cross-modal visibility maps for optical (a) and SAR (b) images; and joint visibility map (c), of Frauenkirche (church) Munich. (a) Red: radar shadow extent; yellow: building facades. (b) Red: optical occlusions; yellow: layover facade extent. (c) Red: Not jointly visible points; yellow: uncertain vertical points (i.e. facades).

TABLE 3.2: Breakdown of the scene coverage of various layers in the cross-modal and joint visibility maps.

Image Type	**Not Jointly Visible**	**Uncertain**	**Jointly Visible**
SAR Image	9.50%	17.77%	72.73%
Optical Image	14.53%	14.73%	70.74%
Scene	25.89%	18.89%	55.22%

likely to be significantly lower that 55% in reality as this bound does not account for feature saliency and radiometric differences which can further affect feature visibility across modalities.

3.3.2 Finding Good Points to Match

The task of designing a multi-modal feature point detector to detect intersecting sets of salient feature points across modalities is highly nuanced and dependant on many interrelated factors, as discussed in Section 3.3.1. This is especially true at high-resolution, where both geometric and radiometric effects play a role in the appearance of features and the likelihood of those features being jointly visible.

However, by incorporating prior information, in the form of geo-referencing, the SAR-optical matching task can be formulated as a correspondence search problem, see Section 3.1.3. Under this formulation, there is no need for a precise multi-modal feature detector, as the correspondence task relies on support regions with a greater spatial extent than in conventional feature matching tasks. Thus a deep learning-based architecture was proposed to guide the selection of multi-modal search and template regions, such that the likelihood of determining correspondence between these regions is increased.

The proposed architecture is comprised of two independent CNNs, one for each modality, and a simple fusion operator. The output of each CNN is a modality-specific map which indicates the likelihood of a region being matchable in the other modality. These maps can then be merged into a *cross-modality scene goodness map* using the fusion operator.

To account for possible geo-registration errors, and geometric distortions, which lead to offsets between corresponding points across modalities, the *domain-specific goodness maps* are created at a low spatial resolution. This ensures that the maximum expected offset between the modalities is collapsed into a single pixel in the output maps. Furthermore it guarantees that the domain-specific goodness maps are aligned before fusion, so that common points of high goodness can be extracted.

The domain-specific networks consist of the first four convolutional layers of the VGG11 architecture (Simonyan et al., 2015), followed by two additional 3×3 convolutional layers, each with a stride of 2. The number of feature channels is then reduced using two fully connected layers implemented using 1×1 convolutional blocks. Finally, an average pooling layer with a kernel size of 4 and unity stride ensures the network has a receptive field size of 128×128 pixels. Thus the proposed good regions can encapsulate a maximum offset of 64 pixels between the modalities, with a granularity of 32 pixels. These specific values were chosen based on the maximum expected offset for the Urban Atlas dataset presented in Section 3.1.1, as well as the insights gained into template patch size from the design of correspondence networks, see Section 3.1. The full architecture can be seen in Figure 3.21.

The domain-specific goodness networks were trained using corresponding SAR-optical patch pairs from the high-resolution Urban Atlas dataset (Section 3.1.1). The goodness problem was then formulated as a binary classification task using a BCE loss, and binary goodness labels for each patch. Due the complexity of creating ground

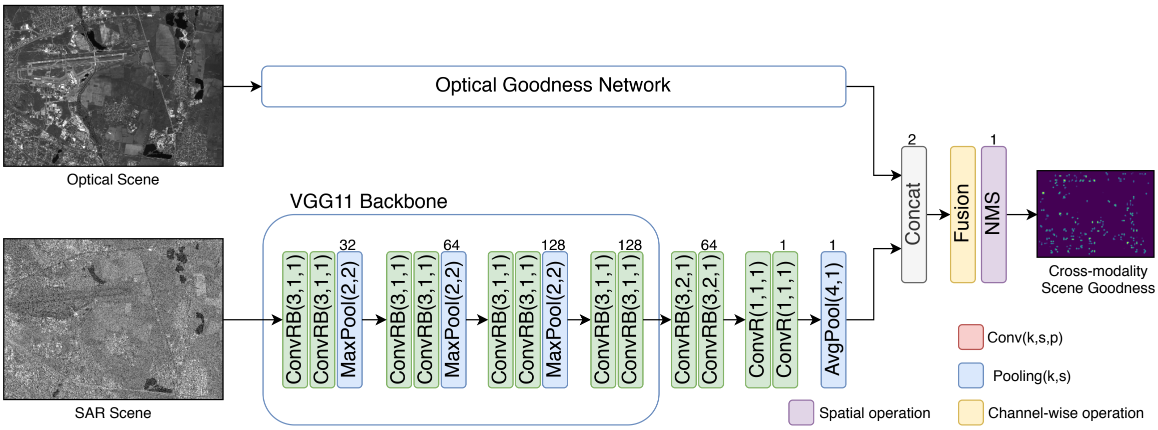


FIGURE 3.21: The goodness network architecture showing the layer details for the SAR branch with Conv(k, s, p) and (Max/Avg)Pool(k, s), representing a convolutional layer, and pooling layer, with a kernel of size k, stride of s, and padding of p, respectively. ReLU and batch normalization are represented by suffixing R and B to Conv, while non-maximal suppression is represented as NMS.

TABLE 3.3: The binary classification performance of the goodness networks with respect to the patches in the test dataset.

Modality	Accuracy	Precision	Recall
SAR	63.6	68.9	69.0
Optical	65.1	69.8	71.3

truth labels for goodness, a weakly-supervised approach was taken. Thus the goodness labels were derived from the matching loss, between the SAR and optical patch, when matched using the *CorrASL* SAR-optical correspondence network presented in Section 3.1.3. Taking this approach presented two main benefits: firstly it removed human bias from the labelling of good regions for matching, and secondly, it encouraged the goodness networks to learn which types of features in one modality generally transfer to matchable features in the other modality. The classification accuracy of the domain specific goodness networks on the test dataset is presented in Table 3.3.

The results of the classification task presented in Table 3.3 highlights the complexity of determining matchable regions across vastly heterogeneous domains, such as SAR and optical. However, the low classification accuracy in the test dataset can also be a side-effect of the feature selection mechanism used during dataset creation (as only optical corner points were considered). In order to assess the benefit of the goodness network to matching, it was applied within that context for the selection of candidate patches.

The first step to using the trained goodness networks to extract candidate search and template regions is to fuse the SAR and optical goodness maps into a cross-modality scene goodness map. This fusion is performed using a minimum operator, which selected the minimum response for each pixel between the two domain-specific goodness maps. Secondly, a localized spatial non-maximum suppression (Dusmanu et al., 2019) was applied, using a 3×3 kernel, to suppress secondary peaks and thus highly overlapping regions. The goodness maps at each stage of the fusion process are depicted in Figure 3.22.

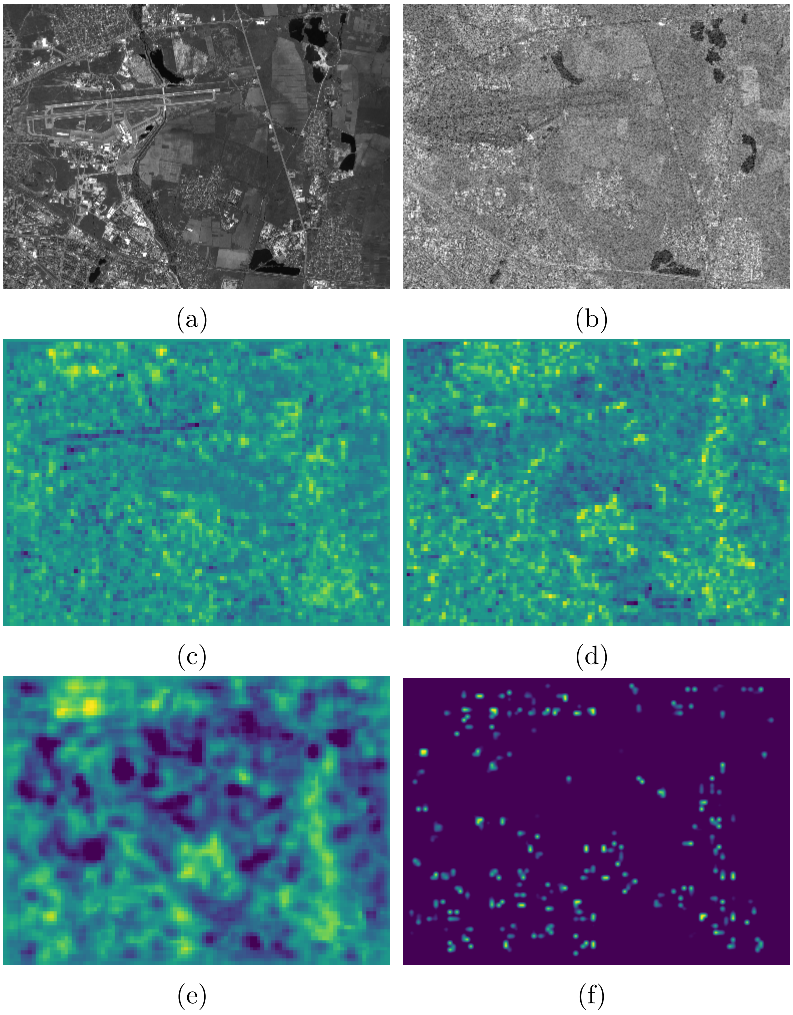


FIGURE 3.22: Example of scene goodness maps produced by the domain specific networks, and the final, fused goodness map. (a) and (b) are the optical and SAR images of the scene. (c) and (d) are the respective domain-specific goodness maps. (e) is the cross-modality goodness map created by minimum fusion of (c) and (d). The final scene goodness, after non-maximal suppression is shown in (f), whereby bright pixels represent identified regions of high goodness.

Finally, candidate patches are extracted around the points of high goodness by transforming these point locations into the original image space. Some examples of regions with high and low goodness as well as misclassified regions are depicted in Figure 3.23. These examples are derived from the corresponding SAR-optical patch pairs in the test dataset.

From Figure 3.23 it can be seen that regions identified as having high goodness contain strong, unambiguous and discriminable features in both modalities, while the low goodness regions lack these properties. False positive and false negative regions share

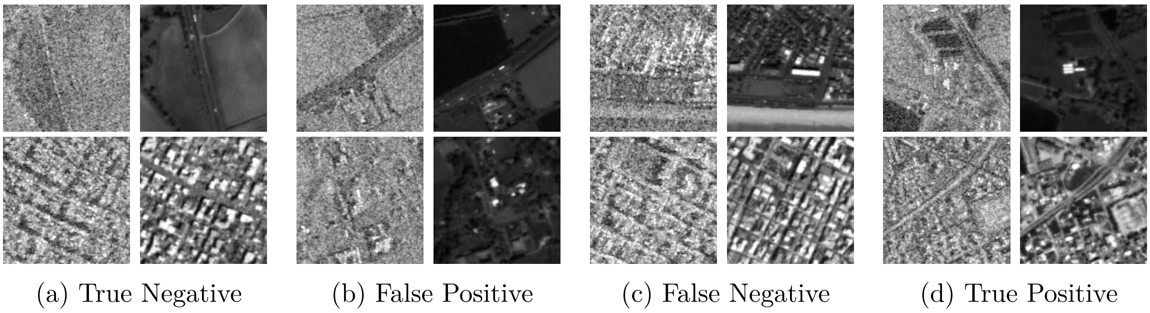


FIGURE 3.23: Examples of regions of low and high goodness (a) and (d) respectively, along with misclassified regions (b) and (c). The SAR patch is shown on the left, and optical on the right for each of the patch pairs.

similar properties, whereby strong features exist in both domains, however, the features contain some ambiguity with respect to matching. In these cases the goodness network appears to fallback on the scene object density as a measure of goodness.

When matching regions of high goodness using the CorrASL network, described in Section 3.1.3, the number of accurately matched points increased by approximately 13%-points to an accuracy of 59% with a mean matching distance of 1.62 pixels. However, the total number of candidate patches decreased substantially, and thus the final set of correspondences was smaller than when only using optical domain features. However, many downstream data fusion do not require large sets of correspondences but rather require that correspondences are accurate and spatially diverse (Bagheri et al., 2018; Müller et al., 2012; Qiu et al., 2018).

3.3.3 Removing Outliers

Even in the best case scenario, where the proposed candidate patches meet all the requirements for increasing the likelihood of matching, outliers and incorrect matches will still exist. This is especially true in the case of matching under extreme heterogeneity, where joint feature visibility is cannot be guaranteed due to the significant differences in radiometric properties, and the complex geometric distortions present in the imagery. Thus the final stage of the feature matching pipeline is to detect and remove as many outliers as possible, such that the final set of correspondences contains only points which are most likely to be inliers.

In classical computer vision the task of identifying and removing outliers has largely been based around statistical methods, such as the Random Sampling and Consensus (RANSAC) algorithm (Fischler et al., 1981). However, these approaches rely on models for the expected transfer of feature between images. In the case of SAR-optical matching, these models are difficult to construct due to the mathematically complex nature of the SAR imaging process, as described in Section 2.1.1. Thus many of the previously discussed SAR-optical matching approaches rely on filtering out matches based on their similarity scores. While this approach works for obvious incorrect matches, it fails to filter out incorrect matches which occur due to spatial symmetry (i.e. two nearby objects with a similar structure) or ambiguities within in the scene.

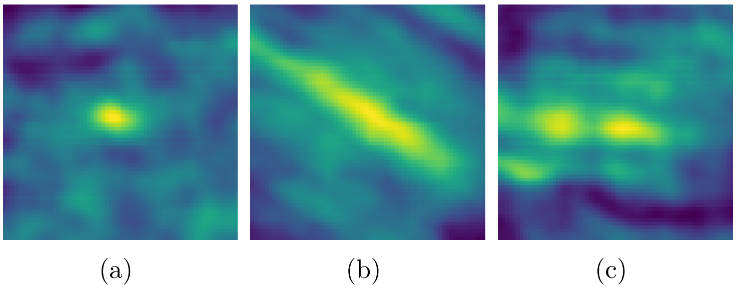


FIGURE 3.24: Examples of common patterns seen in the correspondence heatmaps. For brevity only the *valid* region of the heatmap is depicted. (a) Single, strong response with a low spread. (b) A matching ambiguity exists along a single direction (c) A strongly multimodal response, with a wide spread.

Based on previous observations made within the scope of this thesis, it was hypothesized that the correspondence heatmaps produced by these networks can provide insights into the quality of a matching result. This hypothesis was based on the observation that good matches tend to exhibit a single narrow peak, while incorrect matches are often multi-modal, ambiguous, or have a wide spread. Examples of various trends seen within correspondence heatmaps are presented in Figure 3.24.

Thus a CNN architecture was proposed, and the outlier removal task was formulated as a binary classification problem, whereby the probability of a matching result representing an inlier was predicted based on the correspondence heatmap.

The backbone of the proposed Outlier Reduction Network (ORN) borrows from the architecture of the feature extraction network proposed in Section 3.1.3. However, the first convolutional layer makes use of instance normalization, as opposed to batch normalization. This adaptation was made as the heatmaps produced by the correspondence network have a variable dynamic range and thus cannot be considered to come from the same distribution. Furthermore, as the outlier identification task is a binary classification problem, the head of the network was also adapted to be more suited to this task. This modification included the addition of an adaptive average pooling layer (AdaptAvgPool), which pools the entire spatial extent to produce a single value. Furthermore, a sigmoid activation was applied to the final layer of the network such that the output score directly represents the probability that a heatmap was generated by a successful matching process. The full architecture details are presented in Figure 3.25.

The ORN was trained using correspondence heatmaps created during the training of a SAR-optical matching network, the CorrASL network in this case. Similarly, the training labels were derived as a binarization of the heatmaps respective matching loss. Using these data, the network was trained in a supervised manner using a BCE loss function. In this way, human bias is mostly removed from the labelling process, and the network has the flexibility to learn which heatmap features best represent successful matching.

Evaluating the trained ORN on a test dataset, produced in the same manner as the training dataset, a binary classification accuracy of 81%, with a precision of 76.1% and

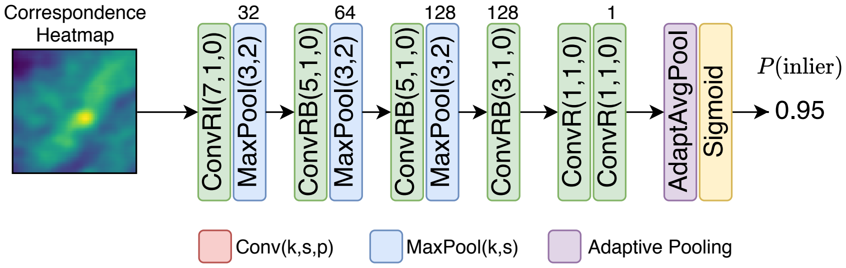
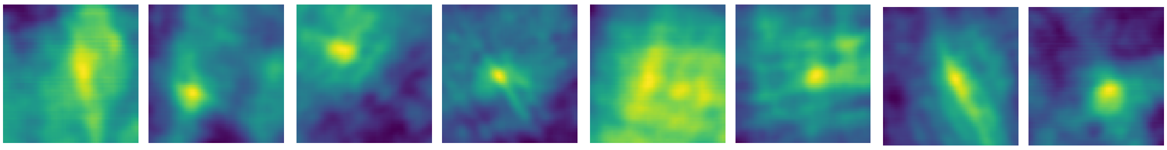


FIGURE 3.25: The architecture of the Outlier Reduction Network (ORN). Conv(k, s, p) and MaxPool(k, s), representing a convolutional layer, and a max-pooling layer, with a kernel of size k, stride of s, and padding of p. ReLU, instance normalization and batch normalization are represented by the suffixes R, I and B, respectively.



(a) True Negative (b) False Positive (c) False Negative (d) True Positive

FIGURE 3.26: Examples of heatmaps corresponding to incorrectly (a) and correctly (d) matched regions, along with mis-classified correspondence heatmaps (b) and (c). The ORN only makes use of the *valid* region of the heatmap for classification.

a recall of 89.5%, was achieved. These results confirm the hypothesis that the correspondence surface holds key information for determining whether or not the matching process was successful. Figure 3.26, provides visual examples of both positive and negative classification results.

The classification results, depicted in Figure 3.26, show that the classification of successful correspondences relies upon more than just the local peak characteristics, although this is an essential factor. Furthermore, these results highlight why classification based purely on the maximum response within the heatmap may lead to a higher number of outliers.

The addition of the ORN to the matching pipeline was further shown to improve the overall accuracy and precision of the final correspondence set, irrespective of the features used to initialize the matching task. Thus highlighting the importance of the outlier removal process, and the proposed approach in performing this task.

3.3.4 Matching High-Resolution SAR and Optical Imagery

Although recently proposed deep matching architectures, within the domain of optical computer vision, are trained in an end-to-end manner (DeTone et al., 2018; Dusmanu et al., 2019), the initial approaches were based on a modular structure consisting of many sub-networks which were chained together (Yi et al., 2016). This modular approach was initially proposed in order to simplify the matching process into manageable subproblems, each with clearly defined data boundaries and goals.

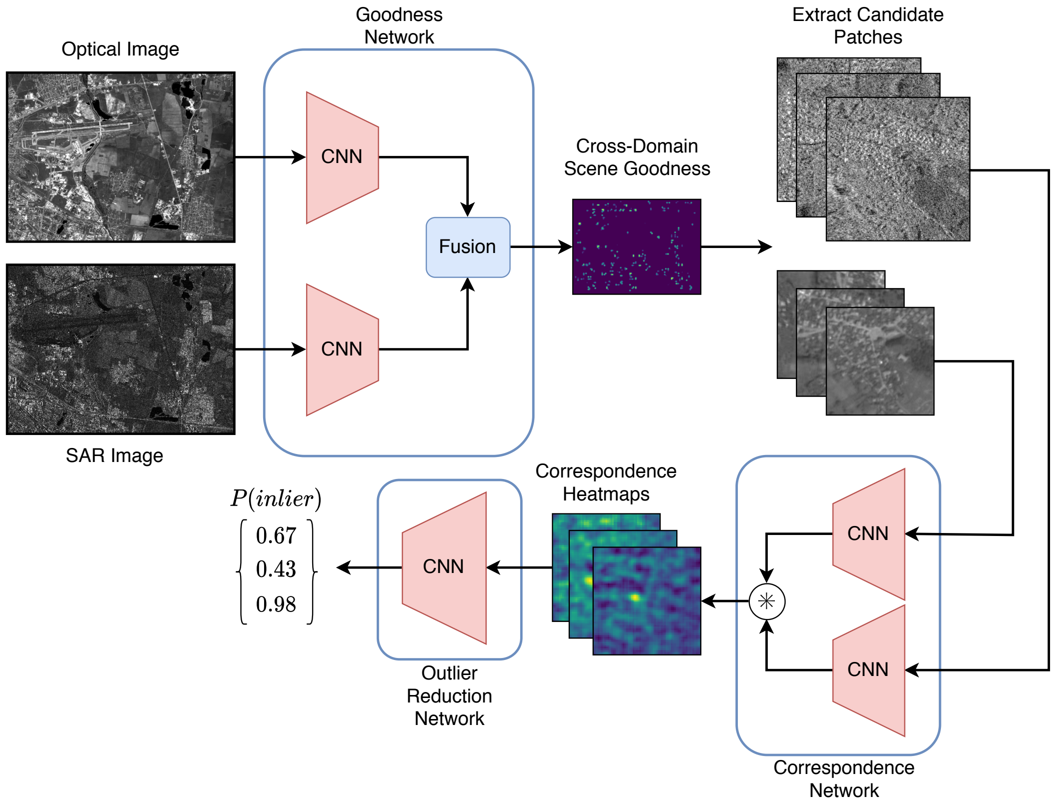


Figure 3.27: Architecture of the proposed comprehensive SAR-optical matching framework. SAR and optical images of the same scene are provided as inputs. The *goodness* network creates a cross-domain scene goodness map which is used to guide the selection of candidate patches. These patches are then matched using a correspondence network, and outliers identified and removed using an outlier reduction network.

Given the recency of SAR-optical deep matching and the lack of prior investigations into deep learning-based feature detection and outlier removal for SAR-optical matching applications, this thesis has approached the development of a comprehensive SAR-optical matching framework from a modular point of view. To this end various deep neural networks have been proposed to address the sub-tasks of feature detection, correspondence and outlier removal.

Thus it was proposed to link these individual solutions to form a comprehensive deep learning-based SAR-optical matching framework which is suited to determining correspondences across large, high-resolution SAR and optical scenes. An overview of the sub-components and the proposed framework architecture is depicted in Figure 3.27.

From Figure 3.27, the three main components of the proposed framework can be seen. The cross-domain scene goodness map generated by the goodness network (Section 3.3.2) is used to extract candidate patch pairs for matching. Each candidate pair consists of a SAR search extent and an optical template patch. These candidate pairs are then matched by a correspondence network which formulates the matching task as a search task and provides as output a correspondence heatmap. Within the frame of this thesis, multiple SAR-optical correspondence architectures have been proposed (Section 3.1 and Section 3.2.2), all of which are capable formulating the correspondence task in this manner. Finally, the resultant correspondence heatmaps are classified according to the likelihood that they represent a successful correspondence. A threshold is applied to filter out low probability correspondences. Finally, the identified set of inliers is used to create the correspondence set by assigning correspondence between the center pixel of the optical template patch and the point of maximal response in the correspondence heatmap - which is spatially aligned to the SAR search extent.

Although end-to-end trainable models have a number of benefits related to their ability to jointly optimize over all sub-problems, these benefits come with significantly more complex data requirements and a lack of flexibility in the network at inference time. Thus the modular approach employed in this thesis allows for each of the sub-networks to be replaced or adapted according to the specific data available and the requirements of the downstream data fusion tasks. Furthermore, it allows for the independent evaluation of each of the sub-tasks such that future research can be better focussed to specific pain points within the SAR-optical matching problem.

3.3.5 Summary

In this section an investigation into the effects of scene and imaging geometry on joint feature visibility was conducted. The results of this investigation highlighted the low level of joint visibility in urban environments, and thus the difficulty of designing generalizable SAR-optical feature point detectors. Based on these findings, a novel deep learning-based approach to multi-modal feature detection was proposed. The proposed approach reformulated feature point detection problem as a *good* region detection problem whereby the goal was to identify image regions in each modality which exhibited a structure and saliency that had a high likelihood of being matchable in the other modality.

Furthermore, a novel approach to outlier detection was proposed, which does not rely on robust statistics, nor mathematical models of feature transfer between images. Instead, it uses a CNN to directly identify successful matches from the structure of the correspondence heatmap surface. In this way, it can be used to filter out correspondence results which have a higher uncertainty or may be ambiguous.

Finally, it was described how to combine this goodness network and outlier removal network with an existing correspondence network (such as those presented within the frame of this thesis), to form a comprehensive SAR-optical matching framework.

Individually the goodness network and the ORN each lead to substantial improvements to the accuracy and precision of the matching results produced by the correspondence network. However, these improvements come at the cost of a reduced number of correspondences in the final set. While this is not always ideal, it is deemed acceptable for a large number of downstream data fusion tasks, such as co-registration (Merkle, Luo, et al., 2017; Müller et al., 2012; Suri & Reinartz, 2010) and SAR-optical stereogrammetry (Bagheri et al., 2018; Qiu et al., 2018), which favour high accuracy and spatial diversity over the total number of correspondences.

4. Publications Supporting this Thesis

This dissertation is founded in the framework of *thesis by publication*, and thus is comprised of several of peer-reviewed publications. These publications provide an in-depth look into the various challenges and potential deep learning-based solutions to the SAR-optical image matching problem, as discussed in Chapter 3.

In the spirit of research and the doctoral process, the core publications of this thesis are supplemented by additional publications which do not directly contribute to the scope of this thesis, or were not subject to the peer-review process. However, they were created during this period of doctoral research and thus further highlight the authors contributions to the advancement of science, well-beyond the scope of the thesis. They are listed at the end of this chapter for completeness.

4.1 Investigation of Joint Visibility Between SAR and Optical Images of Urban Environments

ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume IV-2, 2018
ISPRS TC II Mid-term Symposium “Towards Photogrammetry 2020”, 4–7 June 2018, Riva del Garda, Italy

INVESTIGATION OF JOINT VISIBILITY BETWEEN SAR AND OPTICAL IMAGES OF URBAN ENVIRONMENTS

L. H. Hughes[1], S. Auer[2], M. Schmitt[1]

[1] Signal Processing in Earth Observation, Technical University of Munich (TUM), Munich, Germany - (lloyd.hughes, m.schmitt)@tum.de
[2] Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Oberpfaffenhofen, Germany - stefan.auer@dlr.de

Commission II, WG II/4

KEY WORDS: Synthetic aperture radar (SAR), optical remote sensing, feature visibility, data fusion

ABSTRACT:

In this paper, we present a work-flow to investigate the joint visibility between very-high-resolution SAR and optical images of urban scenes. For this task, we extend the simulation framework SimGeoI to enable a simulation of individual pixels rather than complete images. Using the extended SimGeoI simulator, we carry out a case study using a TerraSAR-X staring spotlight image and a Worldview-2 panchromatic image acquired over the city of Munich, Germany. The results of this study indicate that about 55% of the scene are visible in both images and are thus suitable for matching and data fusion endeavours, while about 25% of the scene are affected by either radar shadow or optical occlusion. Taking the image acquisition parameters into account, our findings can provide support regarding the definition of upper bounds for image fusion tasks, as well as help to improve acquisition planning with respect to different application goals.

1. INTRODUCTION

One of the most important examples for the exploitation of complementary information from different remote sensing sensors is the joint use of synthetic aperture radar (SAR) and optical data (Tupin, 2010, Schmitt et al., 2017). While SAR measures the physical properties of an observed scene and can be acquired independently of daylight and cloud coverage, optical sensors measure chemical characteristics, and require both daylight and clear environmental conditions. Nevertheless, optical data is significantly easier to interpret for human operators and usually provides more details at a similar resolution. In contrast to this, SAR data not only includes amplitude information, but phase too, which enables a high-precision measurement of three-dimensional scene topography and the deformations thereof.

The challenge of fusing SAR and optical data is greatest when data of very high spatial resolutions covering complex built-up areas are to be fused. One example for this is very-high-resolution (VHR) multi-sensor stereogrammetry as discussed by (Qiu et al., 2018). In this application sparse tie-point matching is combined with estimation of the corresponding 3D point coordinates. While the study demonstrated the general feasibility of sparse SAR-optical stereogrammetry of urban scenes, it also brought to light the difficulties involved with robust tie-point matching in the domain of VHR remote sensing imagery. These difficulties, which had also been discussed by (Zhang, 2010, Dalla Mura et al., 2015, Schmitt and Zhu, 2016) before, are caused by the vastly different imaging geometries of SAR and optical images. This difference hinders any straight-forward alignment by exploiting the image geo-coding or classical image-to-image registration methods, and makes prior information about the acquisition and 3D scene geometry a necessity. Even with the use of prior 3D scene knowledge, SAR and optical image tie-point matching still relies on image based multi-modal matching methods. However, these methods are not robust to artefacts caused by the fundamental nature of the imaging geometries (Dalla Mura et al., 2015). For example, multi-path signals, speckle and layover in SAR images can create visual features which have no valid correspondence in the optical image. Nevertheless, image similarity metrics might still detect structurally similar areas in the optical image which then leads to incorrectly matched tie-points. Similarly, points visible in the SAR image might be occluded in the optical image and thus could end up incorrectly matched. These incorrectly matched pixels will lead to a degraded, and sometimes meaningless, fusion product.

In order to be able to develop more sophisticated fusion techniques, it is imperative that the causal effects between scene geometry, imaging modality and acquisition parameters are fully understood, such that an intuition can be built up as to what scene parts are jointly visible between SAR and optical images of complex urban scenes.

In this paper we make use of a remote sensing simulation framework in order to get a feeling for the smallest common denominator, i.e. to produce joint visibility maps for VHR SAR and optical images. Using these maps we aim to provide a better understanding of the causal relationships between the various imaging factors and their effects on the upper bound of possible fusion products. For this task, we first extend the SimGeoI simulation framework (Auer et al., 2017) to allow for dense, pixel-wise simulation of SAR and optical images. Using this extended framework, we develop a processing chain to create easily interpretable joint visibility maps of VHR SAR-optical images. Finally, we produce such joint visibility maps for a test dataset consisting of a TerraSAR-X staring spotlight and a Worldview-2 image acquired over the city of Munich, Germany, to provide the first educated estimation regarding the limitation of SAR-optical data fusion for urban scenes.

The remainder of this paper is structured as follows: Section 2 describes our adaptions to the SimGeoI simulation framework, while Section 3 explains how the adapted framework can be

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-IV-2-129-2018 | © Authors 2018. CC BY 4.0 License.

used to generate joint visibility maps. Section 4 shows the results achieved on real experimental SAR and optical very-high-resolution imagery. Finally, we discuss our findings in Section 5 and provide a conclusion in Section 6.

2. EXTENSION OF SIMGEOI FOR JOINT VISIBILITY MAPPING

2.1 The SimGeoI Simulation Framework

SimGeoI (Auer et al., 2017) is an object-level simulation framework which enables automated alignment and interpretation of SAR and optical remote sensing images. The SimGeoI framework makes use of prior scene knowledge, remote sensing image metadata and a ray-tracing procedure in order to simulate the remote sensing images, and derive object level interpretation layers of the scene from these images. The SimGeoI work-flow is summarized in the flowchart shown in Fig. 1.

The prior scene knowledge is defined by a digital surface model (DSM) provided in UTM coordinates. The DSM is represented by a raster file with pixel values describing the height of each point in the scene. The second input, the image metadata, is extracted directly from the original remote sensing images, which also have to be geo-coded to a UTM coordinate system. The image metadata and geometric prior knowledge in the same coordinate system allow for automated alignment of remote sensing images based on simulation techniques.

The first stage of the process consists of filtering and decomposing the raw DSM in order to create a digital terrain model (DTM) and a normalized DSM (nDSM) (Ilehag, 2016). DTM and nDSM are then triangulated in order to form a closed 3D scene model from the 2.5D DSM data. The next stage is to extract sensor parameters from the image metadata. These parameters include sensor perspective, image properties and average scene height and are used to define signal source, sensor perspective, and image parameters for the ray tracing procedure. Surface parameters are defined appropriately in order to separate object (white) from background (black) in generated images. The image simulation then takes place using a sensor specific ray-tracing engine, GeoRaySAR (Tao et al., 2011) for SAR and GeoRayOpt (Auer et al., 2017) for optical images, and the defined scene model and sensor. This ray tracing step is repeated for the DSM, nDSM and DTM, respectively. Finally the simulated images are geo-coded by rotating the images to a north-east orientation, and then correcting for the constant shift caused by different imaging planes between the original image and the simulated images. With this the simulated images are geo-coded into the UTM coordinate system and aligned with both the DSM and the original image data.

Using the simulated images from the DSM, DTM and nDSM, SimGeoI is able to create various object-level interpretation layers of the scene (Auer et al., 2017). These layers include: ground and vegetation extent; as well as shadow and layover in the case of SAR images; and sun shadow and building extent in the optical case. As the simulated images have been aligned to the DSM and are geo-coded in the same coordinate frame as the original images, these interpretation layers can be used to extract and compare object-level features between remote sensing images of the same scene, from different view points or imaging modalities.

While SimGeoI provides accurate image alignment, and various interpretation layers to aid in understanding SAR and optical images, these insights are only applicable to the object-level of a scene. However, to fully understand all the factors involved in joint visibility of image parts and features across multi-modal remote sensing data, and to build up an intuition of the upper bound of fusion products we require a more fine-grained interpretation of the scene.

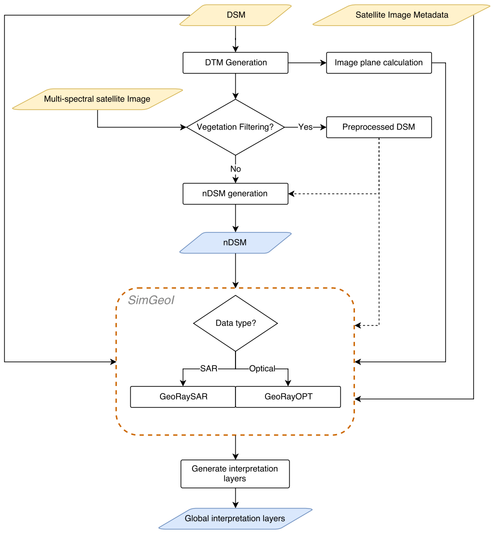


Figure 1. Automated simulation and alignment of remote sensing images with SimGeoI. The red framed section represents the core of SimGeoI which is responsible for the ray-tracing of the DTM, DSM, and nDSM and geo-coding of the resulting images. Yellow: user provided inputs, blue: output products which are used in later processes.

2.2 Extension of SimGeoI for the Simulation of Individual Pixels

In order to perform a detailed analysis of the scene in terms of joint visibility, and uncertainty with respect to artefacts and imaging modality, we extend the SimGeoI framework to enable pixel-level alignment and simulation of the scene. To achieve this pixel-level simulation we add an iterative pixel modelling and ray-tracing procedure as an additional stage to the original SimGeoI pipeline. These additions are depicted in Fig. 2.

Our pixel-level simulation starts by segmenting the preprocessed, non-triangulated nDSM into sub-DSMs using a grid based system. This is done in order to ensure large scenes can be processed in a parallel manner, as each sub-DSM is independent in the ray-tracing phase. Each sub-DSM is then processed in a pixel-wise manner, where each DSM pixel is modelled as a small sphere with its original X, Y coordinates, and a height corresponding to the DSM height at that point. It should be noted that each pixel is used to create a separate 3D model, such that only a single sphere exists in each model. These pixel-wise models are then fed into the ray-tracing engine, along with the camera definition which was created as per the standard SimGeoI simulation procedure. The simulated image, which contains only a single activated pixel, for each pixel-wise model is then geo-coded and aligned with the original remote sensing image. The location of the activated pixel, in UTM coordinates, is then extracted and used to sample the various interpretation layers generated during the object-level simulation. By doing so we are able to not only

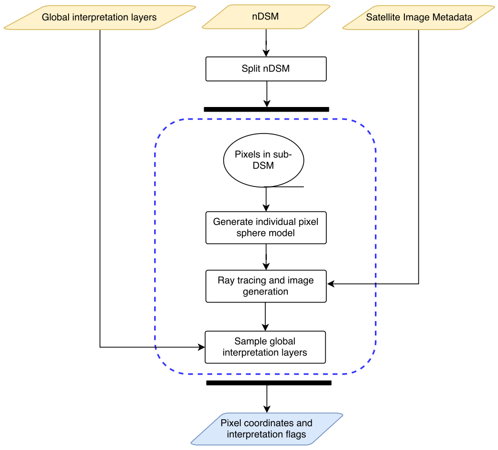

Global interpretation layers
nDSM
Satellite Image Metadata
Split nDSM
Pixels in sub-DSM
Generate individual pixel sphere model
Ray tracing and image generation
Sample global interpretation layers
Pixel coordinates and interpretation flags

Figure 2. Our extension of the SimGeoI framework to pixel-wise simulation and interpretation. The process encapsulated in the dashed blue area is run in parallel and is independent for each sub-DSM. The results are collated at the end, into a single results file for the specified image. Yellow: inputs which are obtained from the original SimGeoI pipeline in Fig. 1. Blue: Collated output file containing pixel-wise results for a single satellite image.

obtain a pixel-wise correspondence between the multi-modal remote sensing images, as well as image pixel to DSM correspondence, but also a pixel-level interpretation of the scene. The DSM pixel coordinate, simulated image pixel coordinates, and pixel interpretation flags for each pixel are then collated and stored in a tabular format.

It should be noted that due to the DSM being a 2.5D raster representation of the scene, vertical regions in the DSM appear as discontinuities when converted to a 3D point cloud representation. Thus our simulation process is unable to obtain pixel correspondences, and interpretation of the facade regions of buildings. These vertical discontinuities can be seen more clearly in Fig. 3.

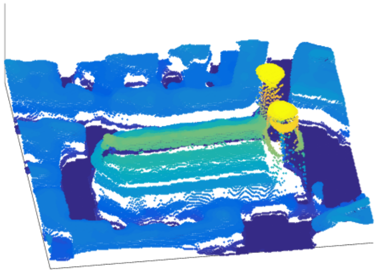


Figure 3. An exemplary point cloud which was extracted from a DSM. The vertical discontinuities are clearly visible as white patches in the point cloud.

Furthermore, as we simulate the DSM pixels individually, imaging effects such as occlusion and radar shadow are not accounted for during simulation. Thus we are able to obtain the theoretical image pixel coordinates for every DSM pixel, irrespective of its true visibility in the original remote sensing image.

3. GENERATING JOINT VISIBILITY MAPS

Using the outputs of the extended SimGeoI framework described in Section 2 for both the SAR and optical images, we are able to derive joint visibility maps for the scene. However, as the DSM pixels are simulated independently, we first need to apply a sensor specific post-processing stage to the results in order to generate additional interpretation layers. These layers are used to impose the original scene geometry constraints on the simulation results. The results from post-processing can then be fused into a final dataset which is used to generate the joint visibility maps. The post-processing and merging process is depicted in Fig. 4 and described in detail below.

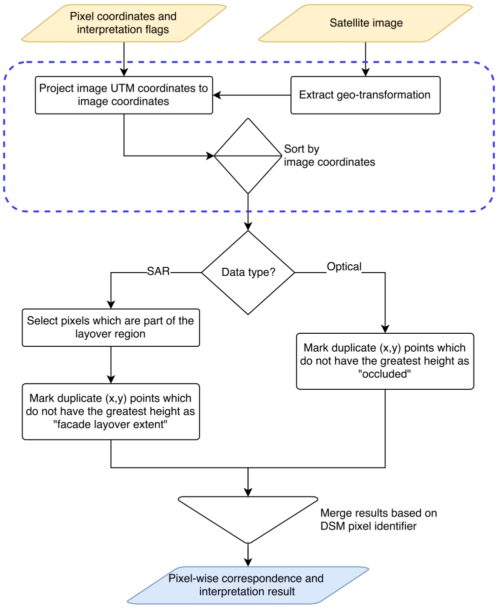

Pixel coordinates and interpretation flags
Satellite image
Project image UTM coordinates to image coordinates
Extract geo-transformation
Sort by image coordinates
Data type?
SAR
Optical
Select pixels which are part of the layover region
Mark duplicate (x,y) points which do not have the greatest height as "facade layover extent"
Mark duplicate (x,y) points which do not have the greatest height as "occluded"
Merge results based on DSM pixel identifier
Pixel-wise correspondence and interpretation result

Figure 4. Our post-processing and merging stage. The process highlighted in blue is run separately for both the SAR and optical pixel-wise results. The projected and sorted image coordinates for both the SAR and optical simulation are then processed to enforce geometric constraints and finally merged into a single output result. Yellow: inputs from previous stages of the pipeline, Blue: the final merged and post-processed pixel-wise interpretation and correspondence dataset which is used to create our joint visibility maps.

3.1 Post-Processing

As the simulation results do not account for the geometric constraints of the scene, we use a post-processing step to add additional interpretation flags to each pixel. These flags specify whether the pixel is subject to any geometric constraints. As these constraints are different between SAR and optical images we require a sensor specific approach to post-processing.

In the case of the optical image simulation, as all the DSM pixels are simulated independently it is possible that many co-linear points exist. Co-linear points are points in the 3D scene which line along the same line of projection, and thus are not truly visible as only the point closest to the camera will be seen. The other points along this line of projection will be occluded. For this reason we add an additional interpretation flag to the optical simulation results specifying whether a simulated pixel is occluded or not. In order to determine co-linear points we make use of a simple strategy which does not require storing intermediate ray-tracing products. Firstly the geo-coded image pixel coordinates are converted to image (x, y)-coordinates such that co-linear points have the same (x, y)-coordinates in the image space. We then select the image pixel which has the greatest corresponding DSM pixel height as the visible pixel, and define all other pixels as being occluded. This strategy holds due to the fact that the remote sensing images we are using are guaranteed to be taken from an aerial vantage point within a relatively small range of image incidence angles. A visual description of why this assumption and technique works can be seen in Fig. 5.

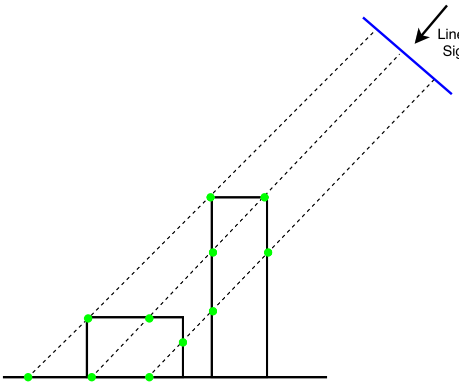


Figure 5. Simplified imaging geometry of an optical satellite sensor. It can be seen that colinear spheres (green), will project to the same image plane (blue) coordinates. However, only the sphere with the greatest height will truly be visible on the image plane.

For SAR images, post-processing is used to determine the extent of facade layover in the image. While SimGeoI provides a layover interpretation layer, this layer masks all scene object-pixels which are subject to layover. However, as the roof structure remains the same and is not often heavily distorted by layover we wish to exclude it from this mask. The reasons for excluding the roof region of buildings is that this region is often jointly visible and may contain important features. Layover pixels are additive in nature and contain, for instance, signal components from both the ground and a building. For this reason we wish to only mask the layover regions which contain ground signal and signal from the facade of the building, not the roof. this is achieved by converting the geo-coded image coordinates to (x, y)-pixel coordinates, and then extracting the pixel with the greatest height to be the building roof. The duplicate pixels are then defined as the layover extent of the building facade. This strategy holds as only a direct signal response occurs on the surface of the modelled DSM pixel sphere. A visual argument for this post-processing stage is depicted in Fig. 6.

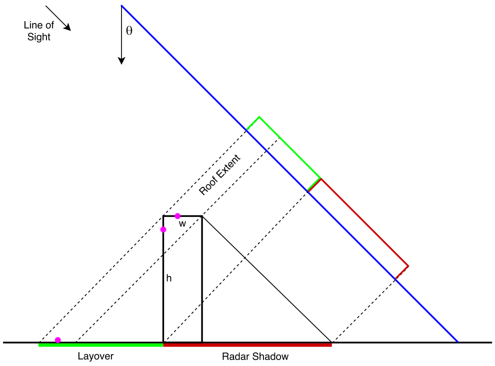


Figure 6. Simplified model of a SAR sensor, and the formation of layover (green) and shadow (red) in a simple scene. The magenta spheres will map to the same image coordinates in the layover region. However, we can ignore the point on the building facade as it is not modelled due to the DSM being 2.5D. Thus by selecting the point with the greatest height we are able to extract the roof extent of the layover, and thereby can obtain the extent of the facade.

3.2 Merging SAR and Optical Simulations

As the SAR and optical images are simulated independently of each other, it is required that we merge their simulation files in order to be able to assess joint visibility between the original images. When we split the nDSM into sub-DSMs we make use of a grid based strategy, such that each grid block can be assigned a unique identifier. Furthermore, when processing the individual pixels in each sub-DSM, the pixels are labelled and processed in a left to right, top to bottom manner. This ensures that each DSM pixel has a unique identifier. Additionally, the SAR and optical simulations make use of the same DSM, thus the DSM identifiers in the SAR and optical image simulation results are equivalent and can be matched by a simple inner join on the data. This enables us to easily determine corresponding pixels between the original SAR and optical images as well as the joint visibility of pixels based on filtering the merged result set by features described in the various interpretation layers and marking the appropriate pixels in the original images. For exemplary demonstration, a small subsection of a final merged simulation result set is presented in Tab. 1.

Table 1. An example of a merged simulation output. Note: the UTM coordinates have been reduced in precision for formatting reasons.

block_id **point_id**	B2674 P186	B2593 P889	B2594 P341
UTMx_sar	691489.874	691481.371	691477.364
UTMy_sar	5334883.531	5334887.031	5334881.032
height_sar	655.292	655.290	655.282
shadow_sar	0	0	0
layover_sar	1	1	1
ground_sar	0	0	0
facade_sar	True	False	False
UTMx_opt	691414.419	691405.919	691401.919
UTMy_opt	5334878.698	5334882.199	5334876.201
height_opt	655.292	655.290	655.282
shadow1_opt	0	0	0
layover1_opt	274	0	498
ground1_opt	0	0	0
layover2_opt	274	0	498
shadow2_opt	0	0	0
ground2_opt	0	0	0
occluded_opt	False	True	False

ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume IV-2, 2018
ISPRS TC II Mid-term Symposium "Towards Photogrammetry 2020", 4–7 June 2018, Riva del Garda, Italy

In order to generate the joint visibility maps we use the merged pixel-wise simulation product, as well as the original remote sensing images. Using these data, generating joint visibility maps for both the SAR and optical images becomes a trivial task. By filtering the dataset to only include the points which make up a specific interpretation layer in either the SAR or optical image, we are able to exploit the list of corresponding SAR and optical image coordinates and plot the extent of this interpretation layer in both images. An example of a joint interpretation layer generated in this manner is depicted in Fig. 7.

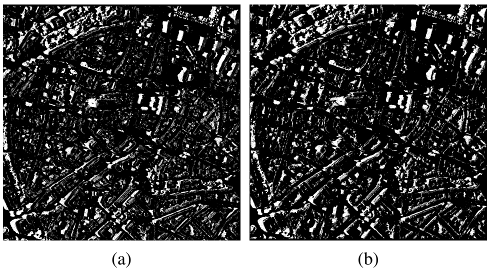


Figure 7. An example of an interpretation layer mask. The extent of radar shadow in the SAR image (a), as well as the extent of shadowed pixels in the optical image (b), is shown in white. Black pixels are unaffected by radar shadow.

4. EXPERIMENT AND RESULTS

4.1 Test Data

For our experiments we make use of a dataset consisting of VHR optical and SAR images, as well as a DSM of the city of Munich, Germany. The DSM of the Munich scene was derived from a Worldview-2 stereo image pair and has a horizontal resolution of $0.5m$ and vertical resolution of $1m$. The details of the remote sensing images are summarized in Tab. 2.

Table 2. Parameters of the test images over Munich, Germany

Data	WorldView-2	TerraSAR-X
Acquisition Date	12/07/2010	07/06/2008
Imaging Mode	panchromatic	staring spotlight
Off-nadir angle (at scene center)	14.5°	49.9°
Orbit	770km	515km
Heading angle	189.0°	188.3° descending
Pixel spacing (east, north)	0.5m	0.5m

4.2 Joint Visibility Map Results

In order to understand which pixels are visible in both the SAR and optical images, we propose the concept of cross-modal and joint visibility maps. These maps describe which pixels can be seen in both images, and thus which pixels are appropriate for matching and fusion applications such as stereogrammetry or tie point detection for image registration.

By masking the facade layover extent and optical occlusion interpretation layers in the SAR image, and the radar shadow and facade extent layers in the optical image, cross-modal joint visibility maps are generated for the scene described in Section 4.1. These cross-modal joint visibility maps can be seen in Figs. 8 and 9. A cropped area around the Frauenkirche (church) is depicted in Fig. 10. For easier reference, the extent of this area is marked by a white frame in Figs. 8 and 9.

In addition to these cross-modal joint visibility maps, we create a joint visibility map which is the projection of both cross-modal joint visibility maps onto an ortho-image, in our case an OpenStreetMap layer. This joint visibility map, seen in Fig. 11, represents the full extent of visible, non-visible and uncertain regions of the scene with respect to both sensors.

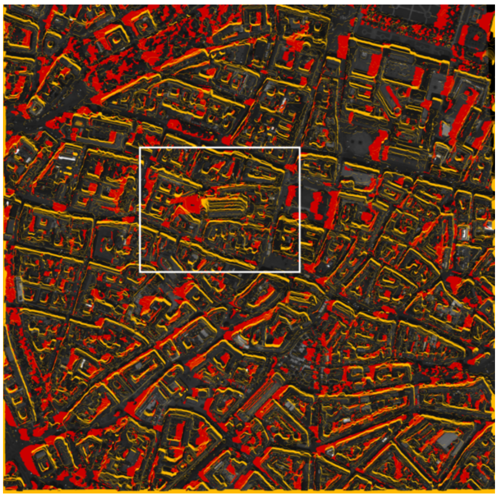


Figure 8. Cross-modal joint visibility map of Munich projected onto the WorldView-2 image. Red: radar shadow extent; yellow: building facades.

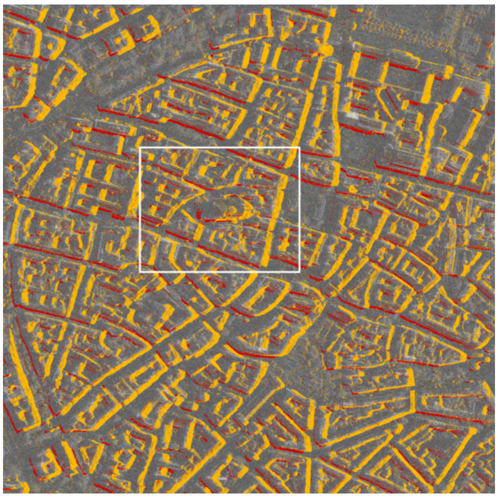
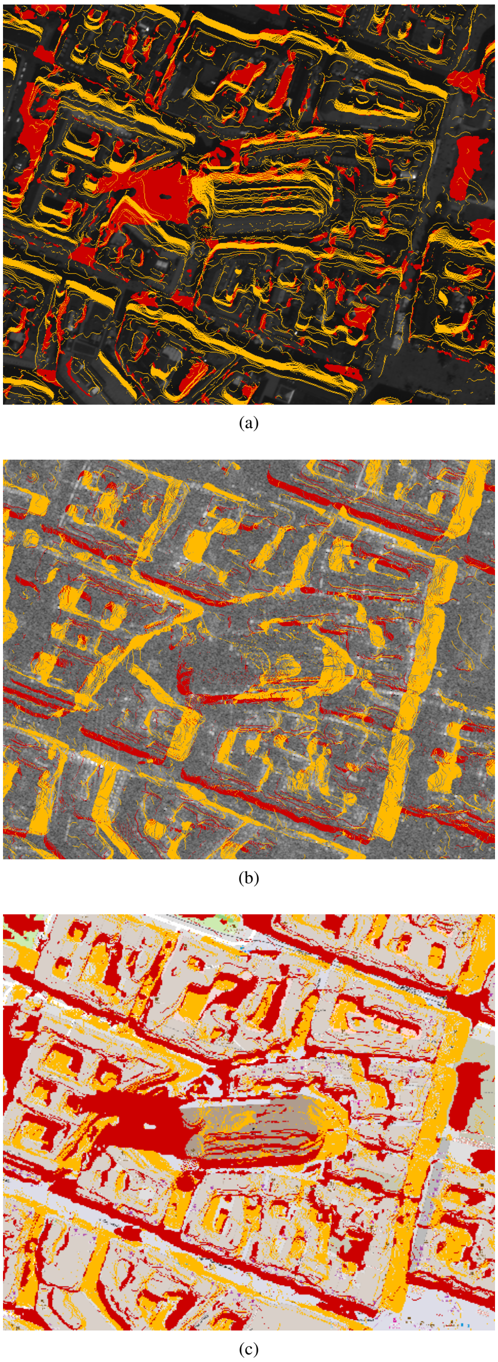


Figure 9. Cross-modal joint visibility map of Munich projected onto the TerraSAR-X image. Red: optical occlusions; yellow: facades layover extent.

The red pixels in these figures represent regions of each image which are not visible in the other modality. For example, in the

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-IV-2-129-2018 | © Authors 2018. CC BY 4.0 License.



(a)

(b)

(c)

Figure 10. Cross-modal joint visibility maps for optical (a) and SAR (b) images; and joint visibility map (c), of Frauenkirche (church) Munich. (a) Red: radar shadow extent; yellow: building facades. (b) Red: optical occlusions; yellow: layover facade extent. (c) Red: Not jointly visible points; yellow: uncertain vertical points (i.e. facades).

case of the optical joint visibility map shown in Fig. 8, the red pixels represent areas of the optical image which cannot be seen in the SAR image due to radar shadow. The yellow pixels in the joint visibility maps describe regions in the image which have high uncertainty with respect to matching, or whose visibility is dependent on the spatial relationship between the sensors and the geometric distortion effects which occur during imaging. For instance, in the SAR visibility map (Fig. 9), the yellow regions represent the extent of building facade in the layover region, while the yellow in the optical visibility map describes the extent of the facade in the optical image. In the case of the joint visibility map, Fig. 11, the red and yellow pixels are formed by combining the results of the cross-modal joint visibility maps described above.

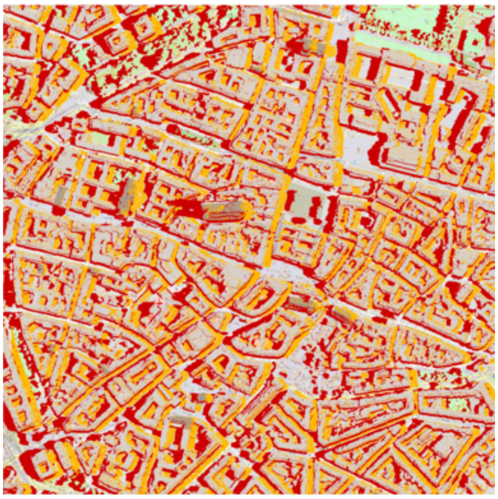


Figure 11. Joint visibility map of Munich projected onto an OpenStreetMap layer. Red: image parts that are not jointly visible due to radar shadow or optical occlusion; yellow: uncertain vertical areas (e.g. facades).

The regions in red cannot be matched and thus do not contribute to the fusion product as they are only visible in one of the images. In contrast, the areas in yellow can still provide useful data, and high quality matching results, if the imaging parameters and scene structure are such that:

- the SAR and optical sensors image the same facade,
- the layover of the facade does not overlay another area with prominent signal response,
- the image matching technique does not rely purely on image geo-coding for defining search areas,
- the scene structure is such that the building facades produce matchable features in both the SAR and optical domain.

5. DISCUSSION

The results presented in Section 4 show that even when accounting for imaging effects such as radar shadow, optical occlusions and facade uncertainty, a significant portion of the scene remains jointly visible, even in complex urban scenes. However, many effects such as sensor baselines, scene geometry, and sensor viewing angels affect the extent of non-visible and uncertain pixels. In this section the effects of these factors on the joint visibility of the scene will be discussed.

5.1 Effect of Sensor Baseline

The baseline between the SAR and optical sensors determines the extent of the scene which is imaged. From our test scene we can see how a relatively wide baseline, coupled with different viewing directions, leads to the SAR and optical sensors capturing different building facades. Furthermore, this non-zero baseline also introduces larger regions of non-jointly visible points as the radar shadow and optical occlusions do not overlap, as is clear when comparing the cross-modal joint visibility maps (Figs. 8 and 9) to the final joint visibility map (Fig. 11).

As it was shown by (Qiu et al., 2018) in order to have favourable conditions for stereogrammetry, the baseline between the sensors should be as small as possible. This small baseline is also favourable for joint visibility. It ensures that the radar shadow (red pixels in Fig. 8) overlaps with the points which are occluded in the optical images (red pixels in Fig. 9), thus decreasing the non-visible regions.

However, a small baseline is not favourable for SAR-optical image matching as the layover of the building falls towards the sensors on the SAR image plane, while the building extent in the optical image falls away from the sensor. Thus it increases the number of uncertain (yellow) pixels in our joint visibility map. Furthermore, building facade images appear mirrored with respect to each other, while the roof structure remains in the same orientation, thus making purely image-based matching approaches more difficult. While prior information about the scene can assist in determining search regions to find corresponding features, and can provide information as to flips and rotations required for patch comparison, the matching of these features remains a difficult task.

5.2 Effect of Viewpoint and Scene Geometry

The viewing angle of the sensors on the scene combined with the scene geometry have the largest part to play in the joint visibility of scene parts. From the results presented in Fig. 10a we are able to see how the high Frauenkirche building causes a large number of pixels to be lost in the optical image due to the extensive radar shadow experienced at a viewing angle of $\theta = 49.9^\circ$. The extent of the radar shadow can be reduced by decreasing the viewing angle. However, this is at the cost on increasing the extent of the layover. In order to ensure that the layover does not fall on nearby buildings, and thereby cause interference with other feature rich areas, it is beneficial to ensure that the extent of the radar shadow is larger than the extent of the layover. From our test scene and resulting joint visibility map, Fig. 11, we can observe that the layover region is smaller than the shadow region, as there is little overlap between red and yellow pixels. This favorable condition is always true for incidence angles greater than $\theta = 45^\circ$.

In the optical case we see that it is preferable to have a viewing angle as close to nadir as possible. In doing so the number of ground points which are occluded by building structures (red pixels in Fig. 9) is minimized. Furthermore, a small viewing angle also reduces the extent of the building facade seen (yellow pixels in Fig. 8) and thus the uncertainty in matching facade regions. Unlike the SAR imaging case, there is no trade-off between a large and small viewing angle in the optical case. For our optical test data, a small viewing angle of $\theta = 14.5^\circ$ was used, and the resulting cross-modal joint visibility map depicts this in the small extent of the facade and occluded regions.

In order to decrease the number of not jointly visible pixels, the smallest viewing angle obtainable by the SAR sensors should be utilized (20° for TerraSAR-X). However, while this provides the greatest joint visibility the extent of the uncertain regions will be large, and thus could degrade matching accuracy and fusion products. For this reason we argue that the optimal viewing angle needs to be considered with reference to the application and scene structure at hand, in order to ensure accurate feature matching can occur but also that a large enough number of pixels remain available to produce a meaningful fusion product.

5.3 Upper Bound of Data Fusion

Apart from developing an intuition as to how scene geometry, viewing angles and sensor baseline play a role in joint scene visibility, we can also extract a theoretic upper bound for data fusion from our joint visibility maps. In order to do this we obtain quantitative results as to the coverage of the scene, when viewed from a nadir angle. These results are presented in Table 3, both from the point of view of the individual images as well as regarding the full scene extent.

Table 3. Breakdown of the scene coverage of various layers in the cross-modal and joint visibility maps.

Image Type	**Not Jointly Visible**	**Uncertain**	**Jointly Visible**
SAR Image	9.50%	17.77%	72.73%
Optical Image	14.53%	14.73%	70.74%
Scene	25.89%	18.89%	55.22%

From the breakdown in Table 3 it is clear that in our test scene only slightly more than half of its extent is jointly visible in both the SAR and the optical satellite image, while the rest is either missing because of optical occlusion or radar shadowing, or uncertain because of belonging to vertical surfaces (i.e. facades). As the imaging geometries of our test scene are typical and are not extreme in viewing angle, scene geometry nor sensor baseline, it can be inferred that this upper bound is likely achievable for scenes of a similar nature.

5.4 Simulation Limitations

As our simulation process makes use of a 2.5D DSM, several limitations exist in our output data. The main limitation is that we cannot obtain pixel-level correspondences on building facades, even when both sensors image the same facade. This leads to building facade pixels being missing from the final merged simulation results, and thus we cannot draw precise conclusions as to the level of joint visibility present in the facade regions. However, we can infer the possibility of joint visibility based on our joint visibility maps, and the sensors viewing angles of the scene.

Furthermore, due to not modelling facades, it is possible that incorrect correspondences can occur when the visible co-linear point lies on a building facade and occluded points on a building rooftop or on the ground. We can see this situation by observing the scene in Fig. 5 and noting how the ray passing through all the facades of the tall building may land upon the roof of the lower building and thus provide an incorrect response. However, due to optical remote sensing data having a look angle of less than 45°, and more commonly less than 25°, as well as the optical scene being modelled using an nDSM, such a situation will only occur with closely spaced buildings of significantly different heights. Furthermore, in order for such an error to have a negative influence on the accuracy of the joint visibility maps, the incorrectly labelled point needs to occur with the same incorrect label in both the SAR and optical cross-modal visibility maps.

In the case of the SAR simulation, the effects of not modelling the building facades are not as apparent. This is due to the fact that the rooftop and facade points layover onto the ground, and while the facade pixels are not simulated these ground and rooftop pixels are, thus encapsulating the full extent of the layover.

6. CONCLUSION AND OUTLOOK

Through our experiments for the first time a strong intuition on the bounds of joint visibility in multi-modal remote sensing was gained – backed by quantitative results. To achieve this, we developed a framework which allows for pixel-wise correspondence to be determined between multi-modal remote sensing images. This framework can provide the basis for many other applications involving the investigation of joint-visibility as well as for data acquisition in applications where high quality labelled data and correspondence information is required, such as training deep matching algorithms.

We further developed an intuition as to the appearance and effect of the various factors involved in the imaging of the scene. We were able to show why a small baseline between the sensors is favourable for stereogrammetry applications. We further described the trade-off between non-visible regions and uncertain regions and present an argument for why the selection of the scene viewing angle is mainly dependent on factors influencing the SAR image. Our results further describe the joint visibility for our test scene is around 55%, even without any optimizing of viewing angle or sensors baselines. This number can serve as an approximate upper bound for matching and image fusion endeavours. Since our test scene was fairly typical, it can be expected that this upper bound approximately extends to scenes with a similar structure and imaging geometry.

In future work the simulation of the building facades will be included in order to gain a more accurate understanding of the nature of uncertain areas in the image, and to what degree these areas remain uncertain and difficult to match. An investigation into the visibility of strong feature points, and their transferability between the SAR and optical domain will be discussed, with the aim of assisting in the selection of high quality feature points and regions to aid matching in SAR-optical stereogrammetry. We will further present a mathematical framework to allow for easier selection of an optimal viewing angle and baseline for use in matching and SAR-optical stereogrammetry data acquisition.

ACKNOWLEDGEMENTS

This work is supported by the German Research Foundation (DFG) under grant SCHM 3322/1-1.

REFERENCES

Auer, S., Hornig, I., Schmitt, M. and Reinartz, P., 2017. Simulation-based interpretation and alignment of high-resolution optical and SAR images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 10(11), pp. 4779–4793.

Dalla Mura, M., Prasad, S., Pacifici, F., Gamba, P., Chanussot, J. and Benediktsson, J. A., 2015. Challenges and opportunities of multimodality and data fusion in remote sensing. *Proceedings of the IEEE* 103(9), pp. 1585–1601.

Ilehag, R., 2016. Exploitation of digital surface models from optical satellites for the identification of buildings in high resolution SAR imagery. Master's thesis, KTH, Sweden.

Qiu, C., Schmitt, M. and Zhu, X. X., 2018. Towards automatic SAR-optical stereogrammety over urban areas using very high resolution images. *ISPRS Journal of Photogrammetry and Remote Sensing* 138, pp. 218–231.

Schmitt, M. and Zhu, X. X., 2016. On the challenges in stereogrammetric fusion of SAR and optical imagery for urban areas. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 41(B7), pp. 719–722.

Schmitt, M., Tupin, F. and Zhu, X. X., 2017. Fusion of SAR and optical remote sensing data – challenges and recent trends. In: *Proceedings of IEEE International Geoscience and Remote Sensing Symposium*, Fort Worth, TX, USA, pp. 5458–5461.

Tao, J., Palubinskas, G., Reinartz, P. and Auer, S., 2011. Interpretation of SAR images in urban areas using simulated optical and radar images. In: *Proceedings of Joint Urban Remote Sensing Event*, pp. 41–44.

Tupin, F., 2010. Fusion of optical and SAR images. In: *Radar Remote Sensing of Urban Areas*, Springer, pp. 133–159.

Zhang, J., 2010. Multi-source remote sensing data fusion: status and trends. *International Journal of Image and Data Fusion* 1(1), pp. 5–24.

4.2 Identifying Corresponding Patches in SAR and Optical Images with a Pseudo-Siamese CNN

784 IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, VOL. 15, NO. 5, MAY 2018

Identifying Corresponding Patches in SAR and Optical Images With a Pseudo-Siamese CNN

Lloyd H. Hughes, *Student Member, IEEE*, Michael Schmitt, *Senior Member, IEEE*, Lichao Mou,
Yuanyuan Wang, *Member, IEEE*, and Xiao Xiang Zhu, *Senior Member, IEEE*

***Abstract*— In this letter, we propose a pseudo-siamese convolutional neural network architecture that enables to solve the task of identifying corresponding patches in very high-resolution optical and synthetic aperture radar (SAR) remote sensing imagery. Using eight convolutional layers each in two parallel network streams, a fully connected layer for the fusion of the features learned in each stream, and a loss function based on binary cross entropy, we achieve a one-hot indication if two patches correspond or not. The network is trained and tested on an automatically generated data set that is based on a deterministic alignment of SAR and optical imagery via previously reconstructed and subsequently coregistered 3-D point clouds. The satellite images, from which the patches comprising our data set are extracted, show a complex urban scene containing many elevated objects (i.e., buildings), thus providing one of the most difficult experimental environments. The achieved results show that the network is able to predict corresponding patches with high accuracy, thus indicating great potential for further development toward a generalized multisensor key-point matching procedure.**

***Index Terms*— Convolutional neural networks (CNNs), data fusion, deep learning, deep matching, image matching, optical imagery, synthetic aperture radar (SAR).**

I. Introduction

THE identification of corresponding image patches is used extensively in computer vision and remote sensing-related image analysis, especially in the framework of stereoapplications or coregistration issues. While many successful handcrafted approaches, specifically designed for the matching of optical images, exist [1], to this date, the matching of images acquired by different sensors still remains a widely unsolved challenge [2]. This particularly holds for a joint exploitation of synthetic aperture radar (SAR) and optical imagery caused by two completely different sensing modalities: SAR imagery collects information about the physical properties of the scene and follows a range-based imaging geometry, while optical imagery reflects the chemical characteristics of the scene and follows a perspective imaging geometry. Hence, structures elevated above the ground level, such as buildings or trees, show strongly different appearances in both SAR and optical images (see Fig. 1), in particular when dealing with very high-resolution (VHR) data.

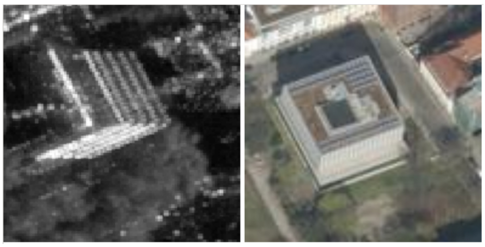


Fig. 1. Simple detached multistory building as (Left) SAR amplitude image and (Right) optical photograph.

In order to deal with the problem of multisensor key-point matching, several sophisticated approaches have been proposed, e.g., exploiting phase congruency as a generalization of gradient information [3]. However, even sophisticated handcrafted descriptors reach their limitations for highly resolving data showing densely built-up urban scenes, which—in the SAR case—is often difficult to interpret even for trained experts.

Therefore, this letter aims at learning a multisensor correspondence predictor for SAR and optical image patches of the state-of-the-art VHR data. Inspired by promising results achieved in the context of stereomatching for optical imagery [4], [5], we also make use of a convolutional neural network (CNN). The major difference of this letter to these purely optical approaches is that we focus on the aforementioned, distinctly more complicated multisensor setup and, therefore, design a specific pseudo-siamese network architecture with two separate, yet identical convolutional streams for processing SAR and optical patches in parallel instead of a weight-shared siamese network in order to deal with the heterogeneous nature of the input imagery.

II. Network Architecture

A. Pseudo-Siamese Convolutional Network

Since SAR and optical images lie on different manifolds, it is not advisable to compare them directly by descriptors

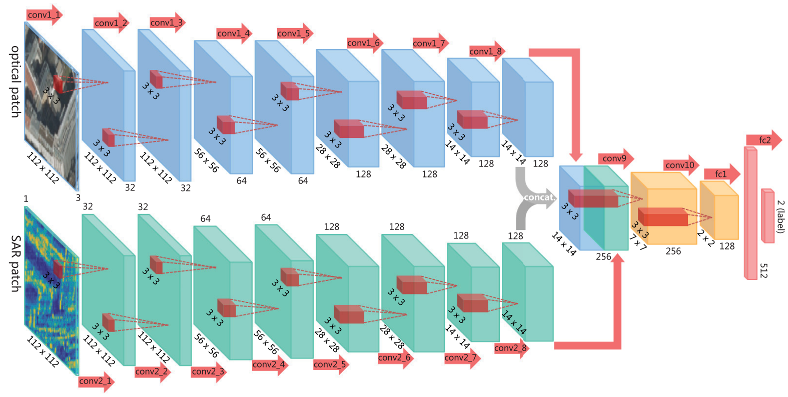
Manuscript received June 30, 2017; revised September 18, 2017, October 30, 2017, and December 19, 2017; accepted January 24, 2018. Date of publication March 12, 2018; date of current version April 20, 2018. This work was supported in part by the China Scholarship Council, in part by the European Research Council under the EU Horizon 2020 Research and Innovation Program under Grant ERC-2016-StG-714087(So2Sat), in part by the Helmholtz Association under the framework of the Young Investigators Group SiPEO under Grant VH-NG-1018, and in part by the German Research Foundation under Grant SCHM 3322/1-1. *(Corresponding author: Xiao Xiang Zhu.)*

L. H. Hughes, M. Schmitt, L. Mou, and Y. Wang are with Signal Processing in Earth Observation, Technical University of Munich, 80333 Munich, Germany (e-mail: lloyd.hughes@tum.de; m.schmitt@tum.de; lichao.mou@dlr.de; yuanyuan.wang@dlr.de).

X. X. Zhu is with Signal Processing in Earth Observation, Technical University of Munich, 80333 Munich, Germany, and also with the Remote Sensing Technology Institute, German Aerospace Center, 82234 Wessling, Germany (e-mail: xiao.zhu@dlr.de).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LGRS.2018.2799232



Details for Config.	SAR Conv. Stream	Optical Conv. Stream
conv. 3x3 (l2 norm), stride 1, padding 1	Conv 1_1 + ReLU + BN	Conv 2_1 + ReLU + BN
conv. 3x3 (l2 norm), stride 1, padding 1	Conv 1_2 + ReLU + BN	Conv 2_2 + ReLU + BN
stride 2	Max Pooling 1	Max Pooling 1
conv. 3x3 (l2 norm), stride 1, padding 1	Conv 1_3 + ReLU + BN	Conv 2_3 + ReLU + BN
conv. 3x3 (l2 norm), stride 1, padding 1	Conv 1_4 + ReLU + BN	Conv 2_4 + ReLU + BN
stride 2	Max Pooling 2	Max Pooling 2
conv. 3x3 (l2 norm), stride 1, padding 1	Conv 1_5 + ReLU + BN	Conv 2_5 + ReLU + BN
conv. 3x3 (l2 norm), stride 1, padding 1	Conv 1_6 + ReLU + BN	Conv 2_6 + ReLU + BN
stride 2	Max Pooling 3	Max Pooling 3
conv. 3x3 (l2 norm), stride 1, padding 1	Conv 1_7 + ReLU + BN	Conv 2_7 + ReLU + BN
conv. 3x3 (l2 norm), stride 1, padding 1	Conv 1_8 + ReLU + BN	Conv 2_8 + ReLU + BN
	Concatenation Layer	
conv. 3x3 (l2 norm), stride 2, padding 1	Conv 9 + ReLU + BN	
conv. 3x3 (l2 norm), stride 2, padding 1	Conv 10 + ReLU + BN	
stride 2	Max Pooling 4	
	FC 1 + ReLU	
	FC 2 + Softmax	
	Loss	

Fig. 2. Proposed pseudo-siamese network architecture and layer configuration.

designed for matching optical patches. Siamese CNN architectures are not suitable for this task either, as weights are shared between the parallel streams, thus implying the inputs share similar image features. In order to cope with the strongly different geometric and radiometric appearances of SAR and optical imagery, in [6], we proposed a pseudo-siamese network architecture with two separate, yet identical convolutional streams, which process the SAR patch and the optical patch in parallel and only fuse the resulting information at a later decision stage. Using this architecture, the network is constrained to first learn meaningful representations of the input SAR patch and the optical patch separately and to combine them on a higher level. The work presented in this letter is an extension of [6] by improving the fusion part of the network architecture, using a different training strategy, and resorting to nonlocally prefiltered SAR patches instead of temporal mean maps. In addition, we evaluate the network on a deterministically partitioned data set instead of a randomly partitioned one, as random partitioning will always cause positively biased results due to overlapping regions in patches.

The architecture of the proposed network is shown in Fig. 2. It is mainly inspired by the philosophy of the well-known VGG Nets [7]. The SAR and optical image patches are passed through a stack of convolutional layers, where we make use of convolutional filters with a very small receptive field of 3×3 rather than using larger ones, such as 5×5 or 7×7. The reason is that 3×3 convolutional filters are the smallest kernels to capture patterns in different directions, such as center, up/down, and left/right, but still have an advantage: the use of small convolutional filters will increase the nonlinearities inside the network and thus make the network more discriminative.

The convolution stride in our network is fixed to 1 pixel; the spatial padding of convolutional layer input is such that the spatial resolution is preserved after convolution, i.e., the padding is 1 pixel for the all 3×3 convolutional layers in our network. Spatial pooling is achieved by carrying out seven max-pooling layers, which follow some of the convolutional layers. They are used to reduce the dimensionality of the feature maps. Max pooling is performed over 2×2 pixel windows with stride 2.

The fusion stage of our proposed network is made up of two consecutive convolutional layers, followed by two fully connected layers. The convolutional layers consist of 3×3 filters, which operate over the concatenated feature maps of the SAR and optical streams, in order to learn a fusion rule which minimizes the final loss function. Max pooling is omitted after the first convolutional layer in the fusion stage, and a stride of 2 is used in order to downsample the feature maps while preserving the spatial information [8]. The use of 3×3 filters and the absence of max pooling after the first convolution allow the fusion layer to learn a fusion rule, which is somewhat invariant to spatial mismatches caused by the difference in imaging modalities. This is due to the fact that the fusion layer uses 3×3 convolutions to learn relationships between the features while preserving nearby spatial information. The lack of max pooling means that these learned spatial relationships are preserved as not only the maximal response is considered, while the stride of 2 is used to reduce the feature size. The final stage of the fusion network consists of two fully connected layers: the first of which contains 512 channels; while the second, which performs one-hot binary classification, contains 2 channels.

In a nutshell, the convolutional layers in our network apart from the fusion layer generally consist of 3×3 filters and follow two rules: 1) the layers with the same feature map size have the same number of filters and 2) the number of feature maps increases in the deeper layers, roughly doubling after each max-pooling layer (except for the last convolutional stack in each stream). All layers in the network are equipped with a rectified linear unit as an activation function, except the last fully connected layer, which is activated by a softmax function. Fig. 2 shows the schematic of the configuration of our network.

B. Loss Function

We make use of the binary cross-entropy loss for training our network. Let $X = (x_1^{\text{sar}}, x_1^{\text{opt}}), (x_2^{\text{sar}}, x_2^{\text{opt}}), \ldots, (x_n^{\text{sar}}, x_n^{\text{opt}})$ be a set of SAR-optical patch pairs, where $x_i^{\text{sar}}, x_i^{\text{opt}} \in R^{D \times D}, \forall_i = 1, \ldots, n$ and $\mathbf{y}_i$ is the one-hot label for the pair $(x_i^{\text{sar}}, x_i^{\text{opt}})$ (with $[1, 0]$ denoting a dissimilar pair, and $[0, 1]$ denoting a similar pair). We then seek to minimize the binary cross-entropy loss

$$E = \frac{1}{n}\sum_{i=1}^{n} \mathbf{y}_i \cdot \log f\left(x_i^{\text{sar}}, x_i^{\text{opt}}, \theta\right) \tag{1}$$

where $f(x_i^{\text{sar}}, x_i^{\text{opt}}, \theta)$ denotes the output vector of the softmax function when comparing the input pair $(x_i^{\text{sar}}, x_i^{\text{opt}})$ with the current network parameters θ.

C. Configuration Details

Fig. 2 shows the full configuration of our network. Apart from the previously discussed architecture, we also make use of batch normalization after the activation function of each convolutional layer. This leads to an increase in the training speed and reduces the effects of internal covariate shift. In order to reduce overfitting during training, we made use of L_2-regularization, with $\lambda = 0.001$, for the convolution kernels of the SAR and optical streams, and dropout with a rate of 0.7 for the first fully connected layer.

III. Automatic Patch Pool Generation

For training and testing purposes, a large pool of corresponding and noncorresponding SAR and optical image patches is needed. While the classical work on deep matching for optical imagery can usually rely on easy-to-achieve optical patch pools (see, for example, the *Phototourism Patch data set* [4], [9]), annotating corresponding patches in VHR optical and SAR imagery of complex urban scenes is a highly nontrivial task even for experienced human experts. Thus, one of the contributions of this letter is the introduction of a fully automatic procedure for SAR-optical patch pool generation.

A. "SARptical" Framework

In order to solve the challenge of automatic data set generation, we resort to the so-called "SARptical" framework of Wang *et al.* [10], an object-space-based matching procedure developed for mapping textures from optical images onto 3-D point clouds derived from SAR tomography. The core of this algorithm is to match the SAR and optical images in 3-D space in order to deal with the inevitable differences caused by different geometrical distortions. Usually, this would require an accurate digital surface model of the area to link homologue image parts via a known object space. In contrast, the approach in [10] creates two separate 3-D point clouds, one from SAR tomography and one from optical stereo matching, which are then registered in 3-D space to form a joint ("SARptical") point cloud, which serves as the necessary representation of the object space. The flowchart of the approach can be seen in Fig. 3. In order to estimate the 3-D positions of the individual pixels in the images, the algorithm requires an interferometric stack of SAR images as well as at least a pair of optical stereoimages. The matching of the two point clouds in 3-D guarantees the matching of the SAR and the optical images. Finally, we can project the SAR image into the geometry of the optical image via the "SARptical" point cloud and vice versa.

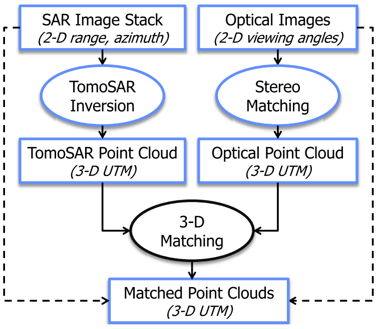


Fig. 3. Flowchart of the patch-pool generation procedure.

B. Data Preparation

For the work presented in this letter, we made use of a stack of 109 TerraSAR-X high-resolution spotlight images of the city of Berlin, acquired between 2009 and 2013 with about 1-m resolution, and of nine UltraCAM optical images of the same area with 20-cm ground spacing. After the reconstruction of the "SARptical" 3-D point cloud, 8840 pixels with high SNR (>5 dB) were uniformly sampled from the nonlocally filtered master SAR amplitude image and projected into the individual optical images, yielding a total of 10 108 corresponding optical pixels. The reason for the difference in pixel numbers is that each of the nine optical multiview stereoimages is acquired from a different viewing angle, making it possible for each SAR image pixel to have up to nine corresponding optical image pixels. The actual number of corresponding optical pixels is dependent on the visibility of the SAR pixel from the respective optical point of view.

All SAR patches are centered at their corresponding SAR image pixels. Their size is fixed at 112 × 112 pixels with a pixel spacing of about 1 m. In analogy, the optical patches are centered at the corresponding optical pixels. After resampling to adjust the pixel spacing, the SAR patches were rotated, so that both patches align with each other as a first approximation.

In order to reduce bias when training our network, we randomly selected a single correct optical correspondence for each SAR image patch during the final data set preparation. In addition, we randomly assign one wrong optical correspondence to each SAR patch in order to create negative examples. Thus, eventually, we end up with 17 680 SAR-optical patch pairs (see Fig. 1 for an example of the class of correct matches).

As final preprocessing steps, the optical patches were converted to gray scale, and all patches were normalized [11] to a radiometric range of [0; 1] with subsequent subtraction of their means.

C. Patch Pool Partitioning

In order to provide a fair experimental design, we partition the patch pool in the following manner: 9724 (55%) of the patch pairs are used as training data set, 2652 (15%) as validation set, and 5304 (30%) as test data set. It has to be

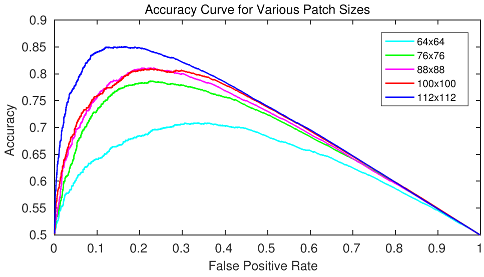


Fig. 4. Comparison of different patch sizes.

noted that we do not partition the patch pool on a purely randomized basis but rather resort to a deterministic partitioning method in order to avoid positively biased test results. The full extent SAR and optical images are first deterministically partitioned and then each partition is processed to generate positive and negative samples for training, validation, and testing, respectively.

IV. Experiments and Results

A. Training Details

The network was trained using the Adam [12] optimization algorithm as it is computationally efficient and exhibits faster convergence than standard stochastic gradient descent methods. The optimization hyperparameters are fixed to $\beta_1 = 0.9$ and $\beta_2 = 0.999$ with a learning rate of $\alpha_t = 0.0009$. The learning rate was found via a grid search method on our training and validation data, while the $\beta-$parameters were kept at their recommended values. Prior to training the network, weight vectors were initialized using the truncated uniform distribution described in [13], and the bias vectors were initialized with zero values. Training was conducted with 2 Nvidia TitanX GPUs using class balanced, minibatches of 64 SAR-optical patch pairs (32 corresponding and 32 noncorresponding pairs) over 30 epochs; training took on average 25 min with a single forward pass taking around 3 ms to complete.

We trained five versions of our proposed network, each at a different patch size, in order to evaluate the effect of patch size on classification accuracy. Patch cropping was done on-the-fly with the new patch being cropped from the center of a larger patch—this was done as the center pixel is the point of correspondence between the SAR and optical patch. Furthermore, we seeded our random number generator with a fixed value of 0, at the start of training for each patch size, in order to prevent the randomization effects between networks.

B. Evaluation Results

We evaluate the proposed network with different input patch sizes using our testing patch pool (described in Section III), which has further been cropped around the center pixel to produce new testing pools with different patch sizes.

The *accuracy versus false positive rate* curves corresponding to different patch sizes can be seen in Fig. 4. Table I reports the corresponding confusion matrix values for our proposed network evaluated with each patch size; it is to be noted that the confusion matrix is the reflective of the network at the point of highest overall performance for each patch size.

TABLE I
Confusion Matrix Values for Different Patch Sizes

	64	76	88	100	112
TP	46.6%	61.6%	66.0%	69.8%	82.2%
TN	86.2%	88.0%	88.2%	86.0%	89.8%

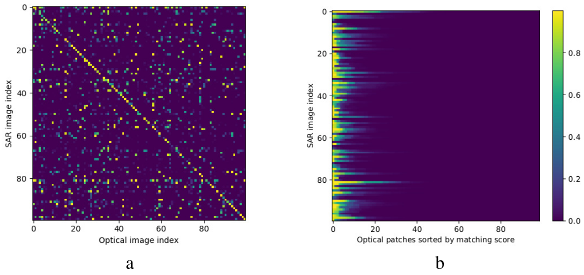


Fig. 5. Results of key-point matching experiment. (a) Confusion matrix showing the matching scores for all SAR and optical key-point patches. (b) Spread of incorrect matches ordered by the similarity score.

C. Key-Point Matching Results

In order to evaluate the proposed network's performance in a real-world, key-point matching scenario, we selected 100 neighboring TomoSAR key-points in the SAR image and extracted the corresponding SAR and optical patch pairs. We selected these key points from a localized area within our test set so as to reproduce the conditions found in a real-world key-point matching application. We then compared every SAR and optical patch in the selected patch set in order to determine the performance of our proposed network in the presence of large numbers of potential mismatches.

In Fig. 5(a), we can see a matrix depicting the similarity scores of the various pair comparisons, where corresponding SAR and optical patches are given the same index number. It should be noted that in determining a binary value for correspondence, a threshold is applied to these similarity scores. Fig. 5(b) shows the sorted scores for all nonsimilar optical patches, making it easier to see the number and strength of incorrect matches in the patch pool.

V. Discussion

Generally, the results summarized in Section IV-B indicate a promising discriminative power of the proposed network. However, the following major points must be considered when interpreting the results.

A. Influence of the Patch Size

As Table I and Fig. 4 clearly indicate, the patch size strongly affects the discriminative power of the network. This result is likely due to the effects of distortions in SAR images, which

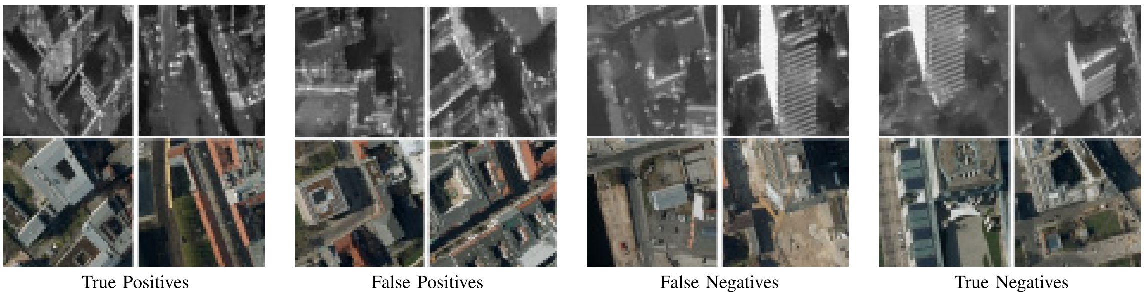


Fig. 6. Exemplary patch correspondence results.

are acquired in a range-based imaging geometry. Thus when patches are cropped to smaller regions, we likely crop out defining features, which are used for matching between the SAR and optical domain. This can be intuitively understood by referring to Fig. 1, where we can see the effects of layover and multipath reflections of the building in the SAR image and a near top down view of the same building in the optical image. Taking away explanatory context will thus render the matching more difficult. All further discussion will be with reference to the results we obtained using the largest patch size, 112 pixels.

B. Comments on the Discriminative Power of the Proposed Network

In summary, our approach obtains an accuracy exceeding 77% on a separate test data set when fixing the false positive rate to 5%, which falls into the same order of magnitude as what can be achieved using the powerful handcrafted HOPC descriptor in combination with an L_2-norm cost function [3].

Furthermore, our approach produced a clear diagonal pattern in Fig. 5(a), which depicts its ability to accurately determine the correct correspondence in a key-point matching scenario. Upon further investigation, it was found that the network achieved 43% top-1 matching accuracy and 74% top-3 accuracy, while 8% of points had no valid matches detected within the key-point set. This was found to be due to large amounts of layover and extreme differences in view point between the SAR and optical patches (see false negatives in Fig. 6).

C. Possible Reasons for False Predictions

From the randomly chosen prediction examples shown in Fig. 6, it can be observed that many of the false positives and false negatives are erroneously matched under extreme differences in viewing angle between the SAR and optical patches. While this may partially be solvable by resorting to larger patch sizes, providing valuable context, there might be a need to exclude image parts with all too strong distortions from further processing.

VI. Conclusion

In this letter, a pseudo-siamese CNN for learning to identify corresponding patches in VHR SAR and optical images in a fully automatic manner has been presented. A first evaluation has shown promising potential with respect to multisensor key-point matching procedures. In order to ensure transferability to other applications not based on key points, e.g., dense matching, we will work on the generation of additional training patches, whose center pixel does not rely on specific key points. In addition, we will test the approach on data coming from a completely different source. In the end, we expect this letter to help paving the way for generalized SAR-optical image matching procedures.

Acknowledgment

The authors would like to thank the NVIDIA Corporation for donating the Titan X Pascal GPU used in this letter.

References

[1] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 10, pp. 1615–1630, Oct. 2005.

[2] M. Schmitt and X. X. Zhu, "Data fusion and remote sensing: An ever-growing relationship," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 4, pp. 6–23, Dec. 2016.

[3] Y. Ye, J. Shan, L. Bruzzone, and L. Shen, "Robust registration of multimodal remote sensing images based on structural similarity," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2941–2958, Mar. 2017.

[4] X. Han, T. Leung, Y. Jia, R. Sukthankar, and A. C. Berg, "MatchNet: Unifying feature and metric learning for patch-based matching," in *Proc. CVPR*, Boston, MA, USA, Jun. 2015, pp. 3279–3286.

[5] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in *Proc. CVPR*, Boston, MA, USA, 2015, pp. 4353–4361.

[6] L. Mou, M. Schmitt, Y. Wang, and X. X. Zhu, "A CNN for the identification of corresponding patches in SAR and optical imagery of urban scenes," in *Proc. JURSE*, Dubai, United Arab Emirates, Mar. 2017, pp. 1–4.

[7] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556

[8] A. Handa, M. Bloesch, V. Pătrăucean, S. Stent, J. McCormac, and A. Davison, "gvnn: Neural network library for geometric computer vision," in *Proc. ECCV Workshops*, Amsterdam, The Netherlands, 2016, pp. 67–82.

[9] S. A. J. Winder and M. Brown, "Learning local image descriptors," in *Proc. CVPR*, Minneapolis, MN, USA, Jun. 2007, pp. 1–8.

[10] Y. Wang, X. Zhu, B. Zeisl, and M. Pollefeys, "Fusing meter-resolution 4-D InSAR point clouds and optical images for semantic urban infrastructure monitoring," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 1, pp. 14–26, Jan. 2017.

[11] Q. Wang, C. Zou, Y. Yuan, H. Lu, and P. Yan, "Image registration by normalized mapping," *Neurocomputing*, vol. 101, pp. 181–189, Feb. 2013.

[12] D. P. Kingma and J. Ba. (2014). "Adam: A method for stochastic optimization." [Online]. Available: https://arxiv.org/abs/1412.6980

[13] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. AISTATS*, Sardinia, Italy, 2010, pp. 249–256.

4.3 The SEN1-2 Dataset for Deep Learning in SAR-Optical Data Fusion

ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume IV-1, 2018
ISPRS TC I Mid-term Symposium "Innovative Sensing – From Sensors to Methods and Applications", 10–12 October 2018, Karlsruhe, Germany

THE SEN1-2 DATASET FOR DEEP LEARNING IN SAR-OPTICAL DATA FUSION

M. Schmitt[1], L. H. Hughes[1], X. X. Zhu[1,2]

[1] Signal Processing in Earth Observation, Technical University of Munich (TUM), Munich, Germany - (m.schmitt,lloyd.hughes)@tum.de
[2] Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Oberpfaffenhofen, Germany - xiao.zhu@dlr.de

Commission I, WG I/3

KEY WORDS: Synthetic aperture radar (SAR), optical remote sensing, Sentinel-1, Sentinel-2, deep learning, data fusion

ABSTRACT:

While deep learning techniques have an increasing impact on many technical fields, gathering sufficient amounts of training data is a challenging problem in remote sensing. In particular, this holds for applications involving data from multiple sensors with heterogeneous characteristics. One example for that is the fusion of synthetic aperture radar (SAR) data and optical imagery. With this paper, we publish the *SEN1-2* dataset to foster deep learning research in SAR-optical data fusion. *SEN1-2* comprises 282,384 pairs of corresponding image patches, collected from across the globe and throughout all meteorological seasons. Besides a detailed description of the dataset, we show exemplary results for several possible applications, such as SAR image colorization, SAR-optical image matching, and creation of artificial optical images from SAR input data. Since *SEN1-2* is the first large open dataset of this kind, we believe it will support further developments in the field of deep learning for remote sensing as well as multi-sensor data fusion.

1. INTRODUCTION

Deep learning has had an enormous impact on the field of remote sensing in the past few years (Zhang et al., 2016, Zhu et al., 2017). This is mainly due to the fact that deep neural networks can model highly non-linear relationships between remote sensing observations and the eventually desired geographical parameters, which could not be represented by physically-interpretable models before. One of the most promising directions of deep learning in remote sensing certainly is its pairing with data fusion (Schmitt and Zhu, 2016), which holds especially for a combined exploitation of *synthetic aperture radar (SAR)* and optical data as these data modalities are completely different from each other both in terms of geometric and radiometric appearance. While SAR images are based on range measurements and observe physical properties of the target scene, optical images are based on angular measurements and collect information about the chemical characteristics of the observed environment.

In order to foster the development of deep learning approaches for SAR-optical data fusion, it is of utmost importance to have access to big datasets of perfectly aligned images or image patches. However, gathering such a big amount of aligned multi-sensor image data is a non-trivial task that requires quite some engineering efforts. Furthermore, remote sensing imagery is generally rather expensive in contrast to conventional photographs used in typical computer vision applications. These high costs are mainly caused by the financial efforts associated to putting remote sensing satellite missions into space. This changed dramatically in 2014, when the SAR satellite Sentinel-1A, the first of the Sentinel missions, was launched into orbit by the European Space Administration (ESA) in the frame of the Copernicus program, which is aimed at providing an on-going supply of diverse Earth observation satellite data to the end user free-of-charge (European Space Agency, 2015).

Exploiting this novel availability of *big remote sensing data*, we publish the so-called *SEN1-2* dataset with this paper. It is comprised of 282,384 SAR-optical patch-pairs acquired by Sentinel-1 and Sentinel-2. The patches are collected from locations spread across the land masses of the Earth and over all four seasons. The generation of the dataset, its characteristics and features, as well as some pilot applications are described in this paper.

2. SENTINEL-1/2 REMOTE SENSING DATA

The Sentinel satellites are part of the Copernicus space program of ESA, which aims to replace past remote sensing missions in order to ensure data continuity for applications in the areas of atmosphere, ocean and land monitoring. For this purpose, six different satellite missions focusing on different Earth observation aspects are put into operation. Among those missions, we focus on Sentinel-1 and Sentinel-2, as they provide the most conventional remote sensing imagery acquired by SAR and optical sensors, respectively.

2.1 Sentinel-1

The Sentinel-1 mission (Torres et al., 2012) consists of two polar-orbiting satellites, equipped with C-band SAR sensors, which enables them to acquire imagery regardless of the weather.

Sentinel-1 works in a pre-programmed operation mode to avoid conflicts and to produce a consistent long-term data archive built for applications based on long time series. Depending on which of its four exclusive SAR imaging modes is used, resolutions down to 5 m with a wide coverage of up to 400 km can be achieved. Furthermore, Sentinel-1 provides dual polarization capabilities and very short revisit times of about 1 week at the equator. Since highly precise spacecraft positions and attitudes are combined with the high accuracy of the range-based SAR imaging principle, Sentinel-1 images come with high out-of-the-box geolocation accuracy (Schubert et al., 2015).

For the Sentinel-1 images in our dataset, so-called ground-range-detected (GRD) products acquired in the most frequently available interferometric wide swath (IW) mode were used. These images contain the σ^0 backscatter coefficient in dB scale for every

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-IV-1-141-2018 |

pixel at a pixel spacing of 5 m in azimuth and 20 m in range. For sake of simplicity, we restricted ourselves to vertically polarized (VV) data, ignoring potentially available other polarizations. Finally, for precise ortho-rectification, restituted orbit information was combined with the 30 m-SRTM-DEM or the ASTER DEM for high latitude regions where SRTM is not available.

Since we want to leave any further pre-processing to the end user so that it can be adapted to fit the desired task, we have not carried out any speckle filtering.

2.2 Sentinel-2

The Sentinel-2 mission (Drusch et al., 2012) comprises twin polar-orbiting satellites in the same orbit, phased at 180° to each other. The mission is meant to provide continuity for multi-spectral image data of the SPOT and LANDSAT kind, which have provided information about the land surfaces of our Earth for many decades. With its wide swath width of up to 290 km and its high revisit time of 10 days at the equator (with one satellite), and 5 days (with 2 satellites), respectively, under cloud-free conditions it is specifically well-suited to vegetation monitoring within the growing season.

For the Sentinel-2 part of our dataset, we have only used the red, green, and blue channels (i.e. bands 4, 3, and 2) in order to generate realistically looking RGB images. Since Sentinel-2 data are not provided in the form of satellite images, but as precisely geo-referenced *granules*, no further processing was required. Instead, the data had to be selected based on the amount of cloud coverage. For the initial selection, a database query for granules with less than or equal to 1% of cloud coverage was used.

3. THE DATASET

In order to generate a multi-sensor SAR-optical patch-pair dataset, a relatively large amount of remote sensing data with very good spatial alignment needs to be acquired. In order to do this in a mostly automatic manner, we have utilized the cloud-based remote sensing platform Google Earth Engine (Gorelick et al., 2017). The individual steps of the dataset generation procedure are described in the following.

3.1 Data Preparation in Google Earth Engine

The major strengths of Google Earth Engine are two-fold from the point of view of our dataset generation endeavour: On the one hand, it provides an extensive data catalogue containing several petabytes of remote sensing imagery – including all available Sentinel data – and other freely available geodata. On the other hand, it provides a powerful programming interface that allows to carry out data preparation and analysis tasks on Google's computing centers. Thus, we have used it to select, prepare and download the Sentinel-1 and Sentinel-2 imagery from which we have later extracted our patch-pairs. The workflow of the GEE-based image download and patch preparation is sketched in Fig. 1. In detail, it comprises the following steps:

3.1.1 Random ROI Sampling In order to generate a dataset that represents the versatility of our Earth as good as possible, we wanted to sample the scenes used as basis for dataset production over the whole globe. For this task, we use Google Earth Engine's `ee.FeatureCollection.randomPoints()` function to randomly sample points from a uniform spatial distribution. Since many remote sensing investigations focus on urban areas and since urban areas contain more complex visual patterns than rural areas, we introduce a certain artificial bias to urban areas by sampling 100 points over all land masses of the Earth and another 50 points only over urban areas. The shape files for both land masses and urban areas were provided by the public domain geodata service `www.naturalearthdata.com` at a scale of 1:50m. If two points are located in close proximity to each other, we removed one of them to ensure non-overlapping scenes.

This sampling process is carried out for four different seed values (1158, 1868, 1970, 2017). The result of the random ROI sampling is illustrated in Fig. 2a.

3.1.2 Data Selection In the second step, we use GEE's tools to filter image collections to select the Sentinel-1/Sentinel-2 image data for our scenes. Since we want to use only recent data acquired in 2017, this first means that we structure the year into the four meteorological seasons: winter (1 December 2016 to 28 February 2017), spring (1 March 2017 to 30 May 2017), summer (1 June 2017 to 31 August 2017), and fall (1 September 2017 to 30 November 2017). Each season is then associated to one of the four sets of random ROIs, thus providing us with the top-level dataset structure (cf. Fig. 3): We structure the final dataset into four distinct sub-groups ROIs1158_spring, ROIs1868_summer, ROIs1970_fall, and ROIs2017_winter.

Then, for each ROI, we filter for Sentinel-2 images with a maximum cloud coverage of 1% and for Sentinel-1 images acquired in IW mode with VV polarization. If no cloud-free Sentinel-2 image or no VV-IW Sentinel-1 image is available within the corresponding season, the ROI is discarded. Thus, the number of ROIs is significantly reduced from about 600 to about 429. For example, all ROIs that were located in Antarctica are rendered obsolete, since the geographical coverage of Sentinel-2 is restricted to 56° South to 83° North.

3.1.3 Image Mosaicking Continuing with the selected image data, we use the Google Earth Engine in-built functions `ee.ImageCollection.mosaic()` and `ee.Image.clip()` to create one single image for each ROI, clipped to the respective ROI extent. The `ee.ImageCollection.mosaic()` function simply composites overlapping images according to their order in the collection in a *last-on-top* sense. As mentioned in Section 2.2, we select only bands 4, 3, and 2 for Sentinel-2 in order to create RGB images.

3.1.4 Image Export Finally, we export the images created in the previous steps as GeoTiffs using the GEE function `Export.image.toDrive` and a scale of 10m. The downloaded GeoTiffs are then pre-processed for further use by cutting the gray values to the $\pm 2.5\sigma$ range, scaling them to the interval $[0; 1]$ and performing a contrast-stretch. These corrections are applied to all bands individually.

3.1.5 First Manual Inspection We then visually inspect all downloaded scenes for severe problems. These can mostly belong to one of the following categories:

- Large *no-data* areas.
  Unfortunately, the `ee.ImageCollection.mosaic()` function does not return any error message if it does not find a suitable image to fill the whole ROI with data. This mostly happens to Sentinel-2, when no sufficiently cloud-free granule is available for a given time period.

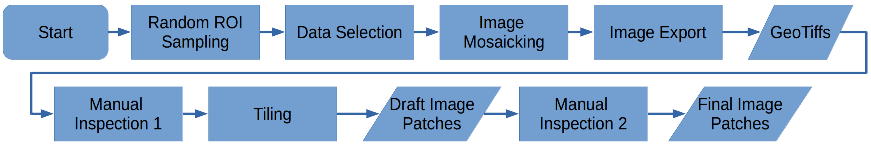
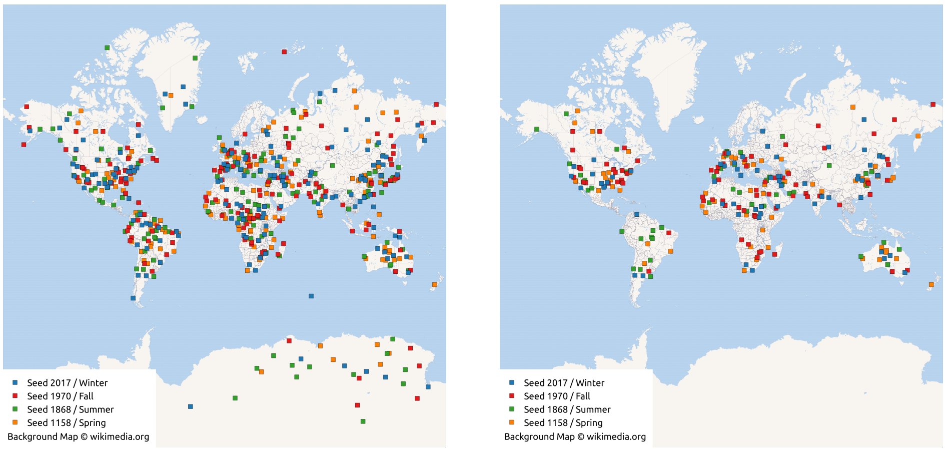


Figure 1. Flowchart of the semi-automatic, Google Earth Engine-based patch extraction procedure.



(a)

(b)

Figure 2. Distribution of the ROIs sampled uniformly over the land masses of the Earth: (a) Original ROIs, (b) final set of scenes after removal of cloud- and/or artifact-affected ROIs.

- Strong cloud coverage.
  The cloud-coverage metadata information that comes with every Sentinel-2 granule is only a global parameter. Thus, it can happen that the whole granule only contains a few clouds, but the part covering our ROI is where all the clouds reside.
- Severely distorted colors.
  Sometimes, we observed very unnatural colors for Sentinel-2 images. Since we want to create a dataset that contains naturally looking RGB images for Sentinel-2, we also removed some Sentinel-2 images with all too strange colors.

After this first manual inspection, only 258 scenes/ROIs remain (cf. Fig. 2b).

3.1.6 Tiling Since our goal is a dataset of patch-pairs that can be used to train machine learning models aiming at various data fusion tasks, we eventually seek to generate patches of 256×256 pixels. Using a stride of 128, we reduce the overlap between neighboring patches to only 50% while maximising the number of independent patches we can get out of the available scenes. We end up with 298,790 Sentinel-1/Sentinel-2 patch-pairs after this step.

3.1.7 Second Manual Inspection In order to remove suboptimal patches that, e.g., still contain small clouds or visible mosaicking seamlines, we have again inspected all patches visually. In this step, 16,406 patch-pairs are manually removed, leaving the final amount of 282,384 quality-controlled patch-pairs. Some examples are shown in Fig. 4.

3.2 Dataset Availability

The *SEN1-2* dataset is shared under the open access license CC-BY and available for download at a persistent link provided by the library of the Technical University of Munich: `https://mediatum.ub.tum.de/1436631`. This paper must be cited when the dataset is used for research purposes.

4. EXAMPLE APPLICATIONS

In this section, we present some example applications, for which the dataset has been used already. These should serve as inspiration for future use cases and ignite further research on SAR-optical deep learning-based data fusion.

4.1 Colorizing Sentinel-1 Images

The interpretation of SAR images is still a highly non-trivial task, even for well-trained experts. One reason for this is the missing color information, which supports any human image understanding endeavour. One promising field of application for the *SEN1-2* dataset thus is to learn to colorize gray-scale SAR images with color information derived from corresponding optical images, as we have proposed earlier (Schmitt et al., 2018). In this approach, we make use of SAR-optical image fusion to create artificial color SAR images as training examples, and of the combination of variational autoencoder and mixture density network proposed by (Deshpande et al., 2017) to learn a conditional color distribution, from which different colorization samples can be drawn. Some first results resulting from a training on 252,384 *SEN1-2* patch pairs are displayed in Fig. 5.

4.2 SAR-optical Image Matching

Tasks such as image co-registration, 3D stereo reconstruction, or change detection rely on being able to accurately determine similarity (i.e. matching) between corresponding parts in different

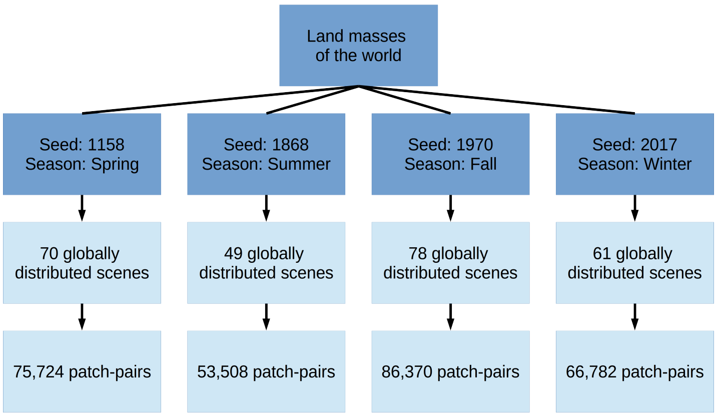

Land masses of the world
Seed: 1158 Season: Spring
Seed: 1868 Season: Summer
Seed: 1970 Season: Fall
Seed: 2017 Season: Winter
70 globally distributed scenes
49 globally distributed scenes
78 globally distributed scenes
61 globally distributed scenes
75,724 patch-pairs
53,508 patch-pairs
86,370 patch-pairs
66,782 patch-pairs

Figure 3. Structure of the final dataset.

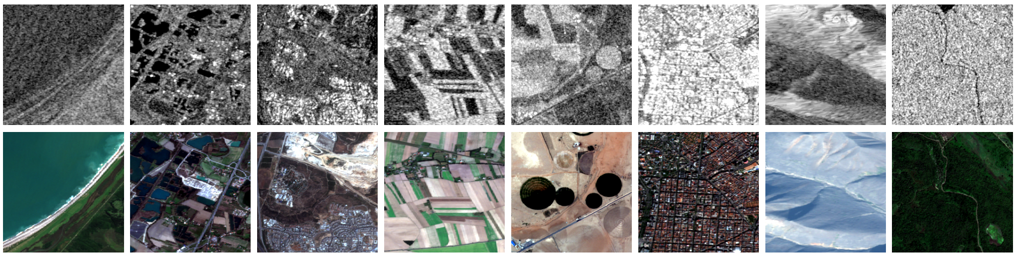


Figure 4. Some exemplary patch-pairs from the *SEN1-2* dataset. Top row: Sentinel-1 SAR image patches, bottom row: Sentinel-2 RGB image patches.

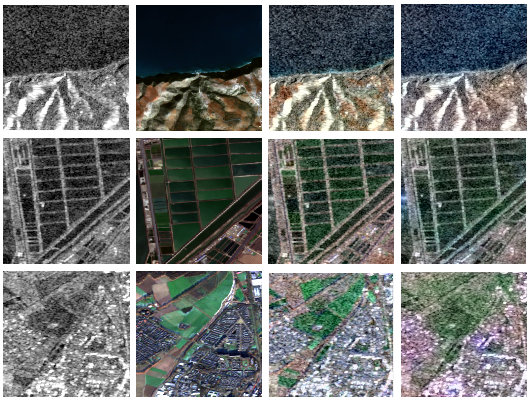


Figure 5. Some results for colorized SAR image patches. In each row, from left to right: original Sentinel-1 SAR image patch, corresponding Sentinel-2 RGB image patch, artificial color SAR patch based on color-space-based SAR-optical image fusion, artificial color SAR image predicted by a deep generative model.

images. While well-established methods and similarity measures exist to achieve this for mono-modal imagery, the matching of multi-modal data remains challenging to this day. The *SEN1-2* dataset can assist in creating solutions in the field of multi-modal image matching by providing the large quantities of data required to exploit modern *deep matching* approaches, such as proposed by (Merkle et al., 2017) or (Hughes et al., 2018): Using a pseudo-siamese convolutional neural network architecture, corresponding SAR-optical image patches of a *SEN1-2* test subset can be identified with an accuracy of 93%. The confusion matrix for the model of (Hughes et al., 2018) trained on 300,000 corresponding and non-corresponding patch pairs created from a *SEN1-2* training subset can be seen in Tab. 1. Furthermore, some exemplary matches achieved on the test subset are shown in Fig. 6.

Table 1. Confusion Matrix for Pseudo-siamese patch matching trained on *SEN1-2*

$\hat{y}/y$	non-match	match
non-match	93.84%	6.16%
match	6.02%	93.98%

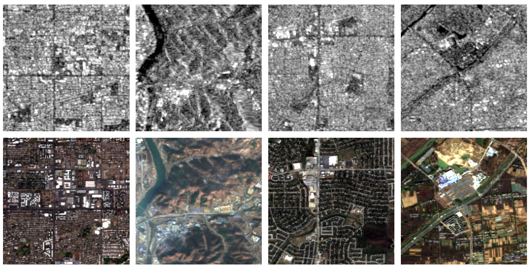


Figure 6. Some true positives achieved in SAR-optical image matching. The first row depicts the Sentinel-1 SAR image patch, while the second row depicts the corresponding Sentinel-2 optical patch as predicted by a pseudo-siamese convolutional neural network.

ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume IV-1, 2018
ISPRS TC I Mid-term Symposium “Innovative Sensing – From Sensors to Methods and Applications”, 10–12 October 2018, Karlsruhe, Germany

4.3 Generating Artificial Optical Images from SAR Inputs

Another possible field of application of the *SEN1-2* dataset is to train generative models that allow to predict artificial SAR images from optical input data (Marmanis et al., 2017, Merkle et al., 2018) or artificial optical imagery from SAR inputs (Wang and Patel, 2018, Ley et al., 2018, Grohnfeldt et al., 2018). Some preliminary examples based on the well-known generative adversarial network (GAN) `pix2pix` (Isola et al., 2017) trained on 108,221 *SEN1-2* patch pairs are shown in Fig. 7.

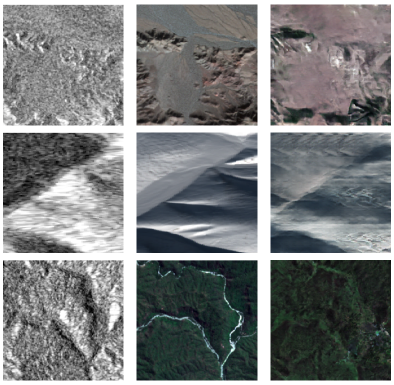


Figure 7. Some preliminary examples for the prediction of artificial optical images from SAR input data using the `pix2pix` GAN. In each row, from left to right: original Sentinel-1 SAR image patch, corresponding Sentinel-2 RGB image patch, artificial GAN-predicted optical image patch.

5. STRENGTHS AND LIMITATIONS OF THE DATASET

To our knowledge, *SEN1-2* is the first dataset providing a really large amount ($>$ 100,000) of co-registered SAR and optical image patches. The only other existing dataset in this domain is the so-called SARptical dataset published by (Wang and Zhu, 2018). In contrast to the *SEN1-2* dataset, it provides very-high-resolution image patches from TerraSAR-X and aerial photogrammetry, but is restricted to a mere 10,000 patches extracted from a single scene, which is possibly not sufficient for many deep learning applications – especially since many patches show an overlap of more than 50%. With its 282,384 patch-pairs spread over the whole globe and all meteorological seasons, *SEN1-2* will thus be a valuable data source for many researchers in the field of SAR-optical data fusion and remote sensing-oriented machine learning. A particular advantage is that the dataset can easily be split into various deterministic subsets (e.g. according to scene or according to season), so that truly independent training and testing datasets can be created, supporting unbiased evaluations with regard to unseen data.

However, also *SEN1-2* does not come without limitations: For example, we restricted ourselves to RGB images for the Sentinel-2 data, which is possibly insufficient for researchers working on the exploitation of the full radiometric bandwidth of multi-spectral satellite imagery. Furthermore, at the time we carried out the dataset preparation, GEE stocked only Level-1C data for Sentinel-2, which basically means that the pixel values represent top-of-atmosphere (TOA) reflectances instead of atmospherically corrected bottom-of-atmosphere (BOA) information. We are planning to extend the dataset for a future version 2 release accordingly.

6. SUMMARY AND CONCLUSION

With this paper, we described and released the *SEN1-2* dataset, which contains 282,384 pairs of SAR and optical image patches extracted from versatile Sentinel-1 and Sentinel-2 scenes. We assume this dataset will foster the development of machine learning, and in particular, deep learning approaches in the field of satellite remote sensing and SAR-optical data fusion. For the future, we plan on releasing a refined, second version of the dataset, which contains not only RGB Sentinel-2 images, but full multi-spectral Sentinel-2 images including atmospheric correction. In addition, we might add coarse land use/land cover (LULC) class information to each patch-pair in order to foster also developments in the field of LULC classification.

ACKNOWLEDGEMENTS

This work is jointly supported by the Helmholtz Association under the framework of the Young Investigators Group SiPEO (VH-NG-1018, www.sipeo.bgu.tum.de), the German Research Foundation (DFG) as grant SCHM 3322/1-1, and the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement ERC-2016-StG-714087, Acronym: *So2Sat*).

REFERENCES

Deshpande, A., Lu, J., Yeh, M.-C., Chong, M. J. and Forsyth, D., 2017. Learning diverse image colorization. In: *Proc. CVPR*, Honolulu, HI, USA, pp. 6837–6845.

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P. et al., 2012. Sentinel-2: ESA’s optical high-resolution mission for GMES operational services. *Remote sensing of Environment* 120, pp. 25–36.

European Space Agency, 2015. Sentinels: Space for Copernicus. `http://esamultimedia.esa.int/multimedia/publications/sentinels-family/`. .(Accessed July 30, 2018).

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D. and Moore, R., 2017. Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment* 202, pp. 18–27.

Grohnfeldt, C., Schmitt, M. and Zhu, X., 2018. A conditional generative adversarial network to fuse SAR and multispectral optical data for cloud removal from Sentinel-2 images. In: *Proc. IGARSS*, Valencia, Spain. in press.

Hughes, L. H., Schmitt, M., Mou, L., Wang, Y. and Zhu, X. X., 2018. Identifying corresponding patches in SAR and optical images with a pseudo-siamese CNN. *IEEE Geoscience and Remote Sensing Letters* 15(5), pp. 784–788.

Isola, P., Zhu, J.-Y., Zhou, T. and Efros, A. A., 2017. Image-to-image translation with conditional adversarial networks. In: *Proc. CVPR*, Honolulu, HI, USA, pp. 1125–1134.

Ley, A., d’Hondt, O., Valade, S., Hänsch, R. and Hellwich, O., 2018. Exploiting GAN-based SAR to optical image transcoding for improved classification via deep learning. In: *Proc. EUSAR*, Aachen, Germany, pp. 396–401.

Marmanis, D., Yao, W., Adam, F., Datcu, M., Reinartz, P., Schindler, K., Wegner, J. D. and Stilla, U., 2017. Artificial generation of big data for improving image classification: a generative adversarial network approach on SAR data. In: *Proc. BiDS*, Toulouse, France, pp. 293–296.

Merkle, N., Auer, S., Müller, R. and Reinartz, P., 2018. Exploring the potential of conditional adversarial networks for optical and SAR image matching. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. in press.

Merkle, N., Wenjie, L., Auer, S., Müller, R. and Urtasun, R., 2017. Exploiting deep matching and SAR data for the geolocalization accuracy improvement of optical satellite images. *Remote Sensing* 9(9), pp. 586–603.

Schmitt, M. and Zhu, X., 2016. Data fusion and remote sensing – an ever-growing relationship. *IEEE Geosci. Remote Sens. Mag.* 4(4), pp. 6–23.

Schmitt, M., Hughes, L. H., Körner, M. and Zhu, X. X., 2018. Colorizing Sentinel-1 SAR images using a variational autoencoder conditioned on Sentinel-2 imagery. In: *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, Vol. XLII-2, pp. 1045–1051.

Schubert, A., Small, D., Miranda, N., Geudtner, D. and Meier, E., 2015. Sentinel-1a product geolocation accuracy: Commissioning phase results. *Remote Sensing* 7(7), pp. 9431–9449.

Torres, R., Snoeij, P., Geudtner, D., Bibby, D., Davidson, M., Attema, E., Potin, P., Rommen, B., Floury, N., Brown, M. et al., 2012. GMES Sentinel-1 mission. *Remote Sensing of Environment* 120, pp. 9–24.

Wang, P. and Patel, V. M., 2018. Generating high quality visible images from SAR images using CNNs. *arXiv:1802.10036*.

Wang, Y. and Zhu, X. X., 2018. The SARptical dataset for joint analysis of SAR and optical image in dense urban area. *arXiv:1801.07532*.

Zhang, L., Zhang, L. and Du, B., 2016. Deep learning for remote sensing data. *IEEE Geoscience and Remote Sensing Magazine* 4(2), pp. 22–40.

Zhu, X. X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F. and Fraundorfer, F., 2017. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine* 5(4), pp. 8–36.

4.4 Mining Hard Negative Samples for SAR-Optical Image Matching Using Generative Adversarial Networks

remote sensing

MDPI

Article

Mining Hard Negative Samples for SAR-Optical Image Matching Using Generative Adversarial Networks

Lloyd Haydn Hughes [1], Michael Schmitt [1] and Xiao Xiang Zhu [1,2,*]

[1] Signal Processing in Earth Observation, Technical University of Munich (TUM), 80333 Munich, Germany; lloyd.hughes@tum.de (L.H.H.); m.schmitt@tum.de (M.S.)

[2] Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), 82234 Wessling, Germany

* Correspondences: xiaoxiang.zhu@dlr.de; Tel.: +49-89-289-22657

Received: 13 August 2018; Accepted: 25 September 2018; Published: 27 September 2018

check for updates

Abstract: In this paper, we propose a generative framework to produce similar yet novel samples for a specified image. We then propose the use of these images as hard-negatives samples, within the framework of hard-negative mining, in order to improve the performance of classification networks in applications which suffer from sparse labelled training data. Our approach makes use of a variational autoencoder (VAE) which is trained in an adversarial manner in order to learn a latent distribution of the training data, as well as to be able to generate realistic, high quality image patches. We evaluate our proposed generative approach to hard-negative mining on a synthetic aperture radar (SAR) and optical image matching task. Using an existing SAR-optical matching network as the basis for our investigation, we compare the performance of the matching network trained using our approach to the baseline method, as well as to two other hard-negative mining methods. Our proposed generative architecture is able to generate realistic, very high resolution (VHR) SAR image patches which are almost indistinguishable from real imagery. Furthermore, using the patches as hard-negative samples, we are able to improve the overall accuracy, and significantly decrease the false positive rate of the SAR-optical matching task—thus validating our generative hard-negative mining approaches' applicability to improve training in data sparse applications.

Keywords: synthetic aperture radar; generative adversarial networks; data fusion; dataset augmentation

1. Introduction

In recent years, data fusion has become a hot topic in the field of remote sensing, specifically the fusion of heterogeneous image data. This increased interest has largely been driven by the improved availability of remote sensing imagery acquired by different sensors [1].

As with any image based data fusion endeavour, a key first step is the determination of corresponding image parts. While considered a somewhat solved problem in traditional computer vision, image matching remains a challenging task when dealing with heterogeneous remote sensing data. One prominent example of this is matching synthetic aperture radar (SAR) and optical satellite imagery, where the sensors have vastly different geometric and radiometric properties making image matching a deeply complex problem [2].

In order to deal with these challenges, several sophisticated approaches have been proposed. Ye et al. [3] propose exploiting phase congruency as a generalization of the gradient information in order to match multimodal images. The approach presented in [4] extends this use of phase congruency to create a radiation-invariant feature transform, which is less susceptible to nonlinear radiation distortions. Using an epipolar-like search strategy and template matching, Qiu et al. [5]

proposed a strategy for simultaneous tie-point matching and 3D reconstruction relying on classical signal- and descriptor-based similarity measures.

While these approaches perform well in some circumstances, most still rely on hand-crafted features and template matching which are difficult to adapt and often suffer from poor discriminability in very high resolution (VHR) imagery. An example of such a failure case can be found when matching very high resolution (VHR) heterogeneous imagery of urban environments, which—in the SAR case—is often difficult even for trained experts to interpret and match [6].

More recently, deep learning has been applied to numerous optical image matching problems with great success. Initial approaches replaced handcrafted feature descriptors with descriptors learned using convolutional neural networks (CNN) [7,8]. However, these were soon outperformed by learning an end-to-end similarity metric for image matching, directly from the data [9,10].

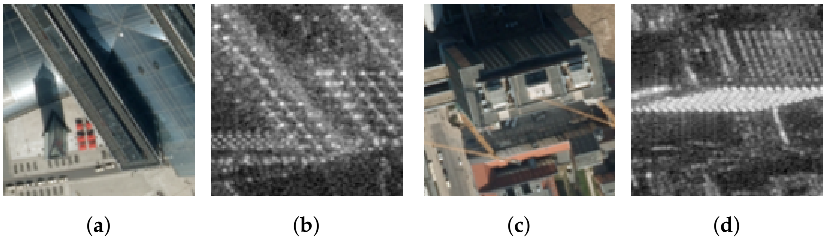
Based on the demonstrated successes in computer vision, deep learning approaches have been gaining interest in the remote sensing community [11]. One possible application is found in the matching of extremely multimodal Earth observation imagery. Merkle et al. [12] proposed the use of a Siamese CNN architecture to compute the relative shift between SAR and optical image patches, with the goal of improving the geo-localization accuracy of optical imagery. Taking inspiration from this success, Mou et al. [13] proposed the use of a pseudo-Siamese CNN in order to frame the SAR-optical correspondence problem as binary classification. With this approach, they provided a proof of concept towards the applicability of CNNs for matching heterogeneous remote sensing imagery. Hughes et al. [14] extended this initial investigation through a modified fusion layer and softmax loss function in order to compute a similarity probability score. Additionally, the investigation was extended to simulate a real-world feature matching scenario and was able to achieve around 86% accuracy with an 11% false positive rate (FPR). Taking a different approach, Merkle et al. [15] proposed the use of a generative adversarial network (GAN) to generate SAR like patches from medium resolution optical images. These generated SAR like patches were then used as the template in a template matching application. This hybrid approach was able to achieve an accuracy of 82% when the threshold for alignment was limited to an error of three pixels.

While these results show great promise for future applications, there has been little to no focus placed on the importance of matching within the scope of a low false positive rate (FPR)—which is arguably more important than achieving a high true positive rate. This requirement is largely driven by the need to reduce outliers in matching results in order to assist downstream applications subsequent to the matching step. This is especially true in multimodal data fusion tasks, such as SAR-optical stereogrammetry [5], where few other methods exist to detect and remove incorrect correspondences.

One common approach to improve the discriminability between classes in classification tasks, and thus reduce the FPR, is known as hard negative mining. This technique uses hard samples (samples which are statistically similar but belong to different classes) as negative examples during the training phase of the classifier [16]. Unlike the randomly assigned negative pairs used in [14], hard negative mining progressively increases the difficulty of the negative examples that the network is trained on. This is done by augmenting the data loading pipeline to replace or append the dataset with data samples which had the greatest misclassification in the previous training iteration. In other words, the samples which are classified as the incorrect class in the most certain manner are now included in the next training iterations as negative examples, thus reinforcing to the classifier that the result is incorrect.

While hard negative mining is simple to implement, it requires that the original dataset is large enough such that, even for low false positive rates, sufficient negative samples exist that can be used as hard negative samples for training. In conventional deep learning applications, this data constraint is often not an issue, as datasets are large enough or can easily be extended. However, for SAR-optical matching applications, this is not the case. While access to remote sensing imagery is becoming easier, and images are geo-coded, the vast differences in imaging geometry mean that geo-coded points cannot be trivially matched. This is particularly true for very high resolution data (see Figure 1).

Thus, expert knowledge or the use of computationally expensive procedures are often required in order to align and match the images such that an accurate patch pair dataset can be produced and labeled [17–19].



(**a**) (**b**) (**c**) (**d**)

Figure 1. An illustrative example of the vast differences between synthetic apature radar (SAR) and optical remote sensing imagery of the same scene. The corresponding image patch-pairs (**a**–**d**) would prove challenging to determine correspondence, even for experts in the field.

To overcome these issues related to data sparsity, researchers have turned to generative networks in order to generate artificial data which can be used to augment or pre-train deep architectures and thus reduce the requirement for large amounts of labeled data [20–23]. Zheng et al. [21] propose using a generative adversarial network (GAN) to generate unlabeled data which was used to improve the baseline in a person-re-identification task. They argued that the imperfect, generated samples act as a form of regularization and thus lead to a more discriminative classifier. In [22], the authors train a SAR to optical transcoding GAN in order to learn key features between various different land surfaces. The top layers of the generator are then used as the main feature extraction sub-network in a multi-modal land cover classifier. Their results show a significant improvement when compared to training the classifier from scratch. Ref. [23] utilizes a GAN to generate negative triplet embeddings in order to allow the discriminator to learn better embedding models.

Marmanis et al. [24] generated VHR SAR patches in order to increase their dataset size for training a SAR image classification network. While the quality of their generated images appears reasonable, they were unable to realize any conclusive results as to whether generated data improved their classification network. Ao et al. [25] proposed a Dialectical-GAN in order to generate VHR SAR imagery from a low resolution Sentinel-1 SAR image prior. However, their results were used as a proof of concept in image translation and were not applied to training of other tasks.

In this paper, we propose an alternative formulation of hard negative mining that can be applied to data sparse applications, such as SAR-optical matching, in order to improve the discriminability of the network and thus reduce the false positive rate. The main contributions are summarized as follows:

Firstly, a GAN architecture is proposed which is capable of generating realistic SAR images which look similar to an existing SAR image, but are modifiable via a continuous latent space. We validate that our generated SAR images are suitable for hard negative mining.

Secondly, we describe how these generated SAR images can be used as hard negative samples to train an existing SAR-optical matching network.

Finally, we demonstrate the effectiveness of our proposed approach by evaluating it on the matching network proposed in [14], and show how we are able to significantly decrease the false positive rate via hard negative mining for the first time.

2. Generative Framework for Hard Negative Mining

In this section, the main structure of our proposed approach to hard negative mining will be described, including our GAN based architecture and training procedure. We will further describe how this architecture can be incorporated into the SAR-optical matching network proposed in [14] in order to augment the training procedure with hard negative samples and thus reduce the false positive rate. An overview of our approach can be seen in Figure 2.

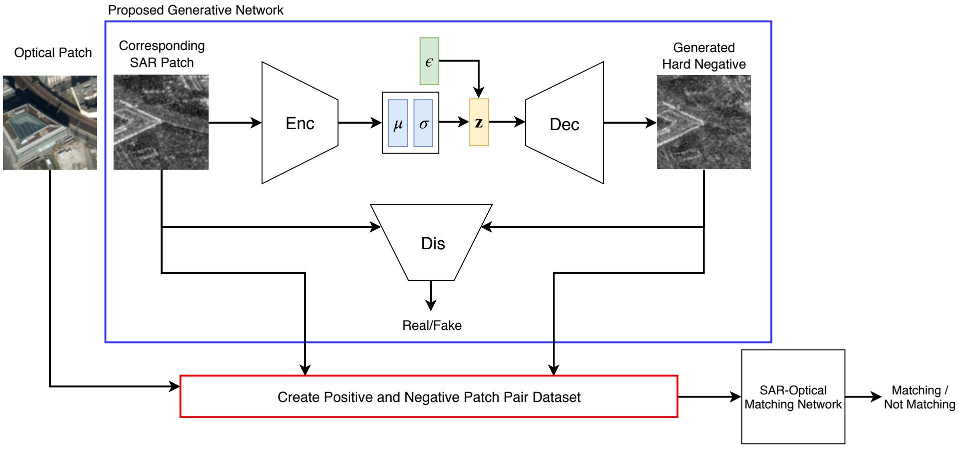


Figure 2. The proposed generative hard-negative mining framework. The GAN is trained to create hard-negative samples based on the input image patch. Together with the original corresponding optical patch, these samples are then used to train the SAR-optical matching network.

2.1. Proposed Generative Architecture

2.1.1. Generator

In order to generate hard negative SAR samples, which are suitable for training a VHR SAR-optical matching network, we make use of a generative model which is trained in an adversarial setting. More specifically, we extend the ProGAN architecture proposed by Karras et al. [26] to include an encoder network which learns a latent representation of our data. This latent representation in turn is used to generate new images. These modifications re-position the original generator network as the decoder network in an autoencoder (AE). We additionally impose a prior over our latent space, $p(\mathbf{z})$, to transform this AE into a variational autoencoder (VAE) which learns a distribution for our input data rather than a discrete latent code. Our proposed VAE consists of two sub-networks: an encoder network and a decoder network. The encoder network (Enc) learns to produce a latent representation, $\mathbf{z}$, from an input sample, $\mathbf{x}$, by

$$\mathbf{z} \sim \text{Enc}(\mathbf{x}) = q(\mathbf{z}|\mathbf{x}). \tag{1}$$

Analogously, the decoder network (Dec), which follows the structure of the generator in [26], learns the mapping from $\mathbf{z}$ back to the data space by

$$\tilde{\mathbf{x}} \sim \text{Dec}(\mathbf{z}) = p(\mathbf{x}|\mathbf{z}). \tag{2}$$

Additionally, we regularize the encoder network by imposing a unit Gaussian prior on the latent distribution $p(\mathbf{z})$, such that $\mathbf{z} \sim \mathcal{N}(1, 0)$.

The decoder network of our VAE follows the design of the ProGAN [26] generator network and is made up of a fully connected bottleneck layer followed by multiple convolutional modules, each of which consists of a nearest neighbor upsampling layer, followed by a convolutional layer, leaky rectified linear unit (LReLU) activation function and a pixel-wise feature vector normalization stage. The pixel-wise normalization layer was added by Karras et al. [26] in order to improve training stability as GANs are inherently unstable and suffer from mode collapse where the generated data collapses to a single sample and thus loses diversity. For full details of the workings of each of the layers in the decoder, we refer the reader to [26] for brevity.

Our encoder network is created by mirroring the structure of the decoder network, and replacing the upsampling operations with an average-pooling downsampling operation. Additionally, a fully connected layer with linear activation is added to the top convolutional layer in order to create the bottleneck required for mapping 2D features to a latent distribution. This mirrored structure has the benefit of simplifying the training procedure (as will become evident in Section 2.2). The structure of our generator VAE is shown in detail in Table 1.

Table 1. A detailed overview of the encoder and decoder network structure.

Encoder	**Act.**	**Output Shape**
Conv 1 × 1	LReLU	N × 1 × 128 × 128
Conv 3 × 3	LReLU	N × 128 × 128 × 128
Conv 3 × 3	LReLU	N × 256 × 128 × 128
Downsample	-	N × 256 × 64 × 64
Conv 3 × 3	LReLU	N × 256 × 64 × 64
Conv 3 × 3	LReLU	N × 512 × 64 × 64
Downsample	-	N × 512 × 32 × 32
Conv 3 × 3	LReLU	N × 512 × 32 × 32
Conv 3 × 3	LReLU	N × 512 × 32 × 32
Downsample	-	N × 512 × 16 × 16
Conv 3 × 3	LReLU	N × 512 × 16 × 16
Conv 3 × 3	LReLU	N × 512 × 16 × 16
Downsample	-	N × 512 × 8 × 8
Conv 3 × 3	LReLU	N × 512 × 8 × 8
Conv 3 × 3	LReLU	N × 512 × 8 × 8
Downsample	-	N × 512 × 4 × 4
Conv 3 × 3	LReLU	N × 512 × 4 × 4
Conv 4 × 4	LReLU	N × 512 × 1 × 1
Fully Connected	Linear	N × 1024 × 1 × 1
Mean	Split	N × 512 × 1 × 1
Std. Deviation		N × 512 × 1 × 1
Decoder	**Act.**	**Output Shape**
Latent Vector	-	N × 512 × 1 × 1
Conv 4 × 4	LReLU	N × 512 × 4 × 4
Conv 3 × 3	LReLU	N × 512 × 4 × 4
Upsample	-	N × 512 × 8 × 8
Conv 3 × 3	LReLU	N × 512 × 8 × 8
Conv 3 × 3	LReLU	N × 512 × 8 × 8
Upsample	-	N × 512 × 16 × 16
Conv 3 × 3	LReLU	N × 512 × 16 × 16
Conv 3 × 3	LReLU	N × 512 × 16 × 16
Upsample	-	N × 512 × 32 × 32
Conv 3 × 3	LReLU	N × 512 × 32 × 32
Conv 3 × 3	LReLU	N × 512 × 32 × 32
Upsample	-	N × 512 × 64 × 64
Conv 3 × 3	LReLU	N × 256 × 64 × 64
Conv 3 × 3	LReLU	N × 256 × 64 × 64
Upsample	-	N × 256 × 128 × 128
Conv 3 × 3	LReLU	N × 128 × 128 × 128
Conv 3 × 3	LReLU	N × 128 × 128 × 128
Conv 1 × 1	Linear	N × 1 × 128 × 128

Following the standard procedure for VAEs, we can define the loss for our proposed generator as the reconstruction error and a prior regularization term, such that $\mathcal{L}_{\text{VAE}} = \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{prior}}$. However, using pixel-wise reconstruction errors with images often leads to blurry and noisy results [27]. Thus, we follow the approach proposed by Larsen et al. [28]. By exploiting the fact that our decoder network can be viewed as the generator network of a standard GAN, we incorporate the standard GAN loss into our VAE loss [29]. In doing so, we combine the advantages of the high-quality generative nature of GANs with the VAEs ability to encode data into an inherently probabilistic latent space $\mathbf{z}$. Our loss terms can now be defined as $\mathcal{L}_{\text{VAE}} = \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{prior}} + \mathcal{L}_{\text{GAN}}$, with:

$$\mathcal{L}_{\text{recon}} = \|\mathbf{x} - \tilde{\mathbf{x}}\|\,, \tag{3}$$

$$\mathcal{L}_{\text{prior}} = D_{\text{KL}}\left(q(\mathbf{z}|\mathbf{x})\|p(\mathbf{z})\right), \tag{4}$$

$$\mathcal{L}_{\text{GAN}} = \log\left(\text{Dis}(\mathbf{x})\right) + \log\left(1 - \text{Dis}(\tilde{\mathbf{x}})\right) + \log\left(1 - \text{Dis}(\text{Dec}(\mathbf{z}_p))\right), \tag{5}$$

where $\mathbf{z}_p$ is a sample from our prior $p(\mathbf{z})$, Dis is our discriminator, and D_{KL} is the Kullback–Leibler divergence.

2.1.2. Discriminator

The discriminator network for our proposed hard negative GAN is designed to be able to distinguish between real SAR image patches and generated SAR-like image patches. The discriminator accepts grayscale images in the form of either the original SAR image patch $\mathbf{x}$ or the generated patch $\tilde{\mathbf{x}} = \text{Dec}(\text{Enc}(\mathbf{x}))$ as input and outputs a scalar score representing how real the images are. This approach is slightly different to standard GANs where the output of the discriminator is a probability [29].

Apart from the bottleneck layers, our discriminator follows the same structure as our encoder network described in Section 2.1.1. The top layers of the discriminator consist of two fully connected layers which reduce the output of the convolutional layers to a single scalar. A linear activation function is then applied to this value in order to obtain a scalar score of image 'realness'. An additional difference between the encoder and discriminator architecture is the inclusion of a mini-batch standard deviation layer which adds an additional feature map to one of the last layers of the discriminator. Karras et al. [26] added this layer in order to increase variation in the network. The full details of the discriminator are described in Table 2.

Table 2. A layer-wise overview of the discriminator network structure.

Discriminator	Act.	Output Shape
Conv 1 × 1	LReLU	N × 1 × 128 × 128
Conv 3 × 3	LReLU	N × 128 × 128 × 128
Conv 3 × 3	LReLU	N × 256 × 128 × 128
Downsample	-	N × 256 × 64 × 64
Conv 3 × 3	LReLU	N × 256 × 64 × 64
Conv 3 × 3	LReLU	N × 512 × 64 × 64
Downsample	-	N × 512 × 32 × 32
Conv 3 × 3	LReLU	N × 512 × 32 × 32
Conv 3 × 3	LReLU	N × 512 × 32 × 32
Downsample	-	N × 512 × 16 × 16
Conv 3 × 3	LReLU	N × 512 × 16 × 16
Conv 3 × 3	LReLU	N × 512 × 16 × 16
Downsample	-	N × 512 × 8 × 8
Conv 3 × 3	LReLU	N × 512 × 8 × 8
Conv 3 × 3	LReLU	N × 512 × 8 × 8
Downsample	-	N × 512 × 4 × 4
Mini-batch Std. Dev.	-	N × 513 × 4 × 4
Conv 3 × 3	LReLU	N × 512 × 4 × 4
Conv 4 × 4	LReLU	N × 512 × 1 × 1
Fully Connected	Linear	N × 1 × 1 × 1

2.2. Training Procedure

Our training procedure combines the training procedure of [26] with the dual GAN and VAE loss definitions of [28], as described in Section 2.1.1.

2.2.1. Progressive Growing

We initialize our networks to start the training process with an image resolution of 4 × 4 pixels. We then gradually increase this resolution by a factor of 2 after a specified number of training iterations. In order to prevent jolting the system when a new layer is added, we closely follow the process described in [26]. Adding new layers to the networks in a smooth manner consists of a two stage approach. During the *transition phase*, we treat layers which operate on the higher resolution as a residual block whose weight α increases linearly from 0 to 1 over a set number of training iterations. Additionally, we interpolate between two resolutions of the input image, in a similar manner to how the generator combines the new and old resolution. The second stage is the *stabilization phase*, whereby the networks are trained for a specific number of iterations before the resolution is doubled again. All of the networks are grown in this manner from a low resolution of 4 × 4 pixels to our final resolution of 128 × 128 pixels. Using networks that have a similar structure simplifies the process of managing multi-resolution data and eases the complexity involved in transitioning between layers. An example of the networks training progression is depicted in Figure 3.

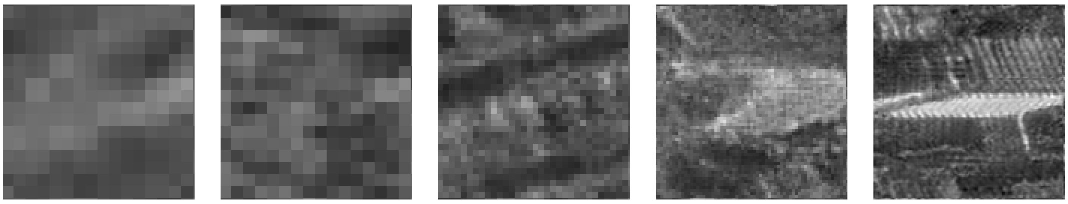

(**a**) (**b**) (**c**) (**d**) (**e**)

Figure 3. An example of progressively grown images taken at increasing image resolutions during the training processes. (**a**–**e**) shows the resolution growth from 8 × 8 pixels up to 128 × 128 pixels with the resolution doubling at each stage.

This progressive growing approach drastically speeds up training of the GAN and improves the overall training stability as the network only needs to learn small transformations between the previous and next layers.

2.2.2. WGAN-GP Loss

While the proposed training approach greatly improves stability and reduces the chances of mode collapse, it does not solve the issue of large gradients which occur when generating high resolution images. This gradient issue occurs due to the fact that fake images are significantly easier to distinguish at high resolutions and thus large gradients propagate from the discriminator.

In order to prevent this gradient problem, and further increase the stability of training, we make use of the improved Wasserstein GAN loss with gradient penalty (WGAN-GP) [30]. This loss function is used to train the discriminator network, as well as to replace the standard GAN loss $\mathcal{L}_{\text{GAN}}$ which is used to train our generator network (Equation (5)). Thus, our new loss functions can be defined as

$$\mathcal{L}_{\text{Dis}} = \underbrace{\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathbb{P}_{\text{real}}}[\text{Dis}(\mathbf{x})] - \mathop{\mathbb{E}}_{\tilde{\mathbf{x}}\sim p(\mathbf{x}|\mathbf{z})}[\text{Dis}(\tilde{\mathbf{x}})]}_{\text{Original Critic Loss}} + \underbrace{\lambda \mathop{\mathbb{E}}_{\hat{\mathbf{x}}\sim\mathbb{P}_{\hat{\mathbf{x}}}}[(\|\nabla_{\hat{\mathbf{x}}}\text{Dis}(\hat{\mathbf{x}})\| - 1)^2]}_{\text{Gradient Penalty}}, \tag{6}$$

$$\mathcal{L}_{\text{VAE}} = \underbrace{-\mathop{\mathbb{E}}_{\tilde{\mathbf{x}}\sim p(\mathbf{x}|\mathbf{z})}[\text{Dis}(\tilde{\mathbf{x}})] - \mathop{\mathbb{E}}_{\mathbf{z}_p\sim p(\mathbf{z})}[\text{Dis}(\text{Dec}(\mathbf{z}_p))]}_{\text{Original Generator Loss}} + \underbrace{\mathop{\mathbb{E}}_{q(\mathbf{z}|\mathbf{x})}[\log(\frac{q(\mathbf{z}|\mathbf{x})}{p(\mathbf{z})})]}_{\text{KL-Divergence}} + \underbrace{\gamma\sum_{i}^{N}\|\mathbf{x} - \text{Dec}(\text{Enc}(\mathbf{x}))\|}_{\text{Reconstruction Error}}, \tag{7}$$

where $\mathbb{P}_{\hat{\mathbf{x}}}$ is implicitly defined as sampling uniformly between pairs of points sampled from the data distribution $\mathbb{P}_{\text{real}}$ and the decoder distribution $p(\mathbf{x}|\mathbf{z})$ and λ and γ are weighting coefficients which are set as hyper-parameters.

2.2.3. Additional Training Details

Using the losses defined in Equations (6) and (7), we train our network using the *Adam* gradient descent with the momentum approach. The learning rate is initialized to 0.001 for the decoder and discriminator networks and 0.0005 for the encoder network. Additionally, the moving average filter parameters for the Adam optimizer are set to $\beta_1 = 0$, $\beta_2 = 0.99$ for all networks. Training data is fed to the network using an initial mini-batch size of 128 samples. However, this number is decreased to 16 samples as the resolution increases. All three sub-networks are grown simultaneously with a *transition rate* and *stabilization rate* of 60,000 images or approximately 10 epochs each.

Additionally, as per the findings of [28], we do not propagate the error signals from the $\mathcal{L}_{\text{GAN}}$ losses to the encoder network. Furthermore, as the decoder network receives error signals from both $\mathcal{L}_{\text{GAN}}$ and $\mathcal{L}_{\text{recon}}$, we set the weighting term $\delta = 0.6$ to add a slight preference to the network's ability to reconstruct the input over its ability to fool the discriminator. We also include reconstructed samples $\tilde{\mathbf{x}}$, as well as samples from our prior distribution $p(\mathbf{z})$ in our GAN objective, as this was found to produce better results than using only samples from the prior distribution. The inner training loop is detailed in Algorithm 1.

Algorithm 1: Training our Hard Negative GAN

```
$\Theta_{\text{Enc}}, \Theta_{\text{Dec}}, \Theta_{\text{Dis}} \leftarrow$ Glorot uniform initialization
repeat
    $\mathbf{X} \leftarrow$ random mini-batch from dataset
    $\mathbf{Z} \leftarrow \texttt{Enc}(\mathbf{X})$
    $\mathcal{L}_{\text{prior}} \leftarrow D_{\text{KL}}\left(q(\mathbf{Z},\mathbf{X})|p(\mathbf{Z})\right)$
    $\tilde{\mathbf{X}} \leftarrow \texttt{Dec}(\mathbf{Z})$
    $\mathcal{L}_{\text{recon}} \leftarrow \left\|\mathbf{X} - \tilde{\mathbf{X}}\right\|$
    $\mathbf{Z}_p \leftarrow$ samples from prior $\mathcal{N}(\mathbf{0},\mathbf{I})$
    $\mathbf{X}_p \leftarrow \texttt{Dec}(\mathbf{Z}_p)$
    $\mathcal{L}_{\text{WGAN}} \leftarrow \texttt{Dis}(\mathbf{X}) - \texttt{Dis}(\tilde{\mathbf{X}})$ See Equation (6)
    $\mathcal{L}_{\text{Dec}} \leftarrow \texttt{Dec}(\tilde{\mathbf{X}}) + \texttt{Dis}(\texttt{Dec}(\mathbf{Z}_p))$ See Equation (7)
    Update network according to gradients
    $\Theta_{\text{Dis}} \overset{+}{\leftarrow} -\nabla_{\Theta_{\text{Dis}}}(\mathcal{L}_{\text{WGAN}} + \lambda\mathcal{L}_{\text{GP}})$
    $\Theta_{\text{Enc}} \overset{+}{\leftarrow} -\nabla_{\Theta_{\text{Enc}}}(\mathcal{L}_{\text{prior}} + \mathcal{L}_{\text{recon}})$
    $\Theta_{\text{Dec}} \overset{+}{\leftarrow} -\nabla_{\Theta_{\text{Dec}}}(\gamma\mathcal{L}_{\text{recon}}\text{-}\ \mathcal{L}_{\text{Dec}})$
until convergence;
```

2.3. Generating Hard Negative Samples

In order to generate hard negative samples, we train our proposed GAN on 6629 SAR images from the training data which are used to train the SAR-optical matching network. In doing so, the encoder network learns the latent distribution of our training data and the decoder network learns to reconstruct the input data from this distribution. After training, the discriminator network is discarded and the VAE is used to generate hard negative SAR samples. As the latent space is continuous and follows a unit normal distribution, we can create novel, yet similar SAR patches by sampling the latent distribution near to the location of the encoded input image. This process is depicted in Figure 4.

These generated, SAR-like images are then used as hard negative samples in the SAR-optical matching training dataset. This is done by creating a non-corresponding patch-pair which consists of the generated SAR image, and the optical image which corresponds to the original SAR image, which was used to generate the hard-negative. Some examples of the appended dataset can be seen in Figure 5.

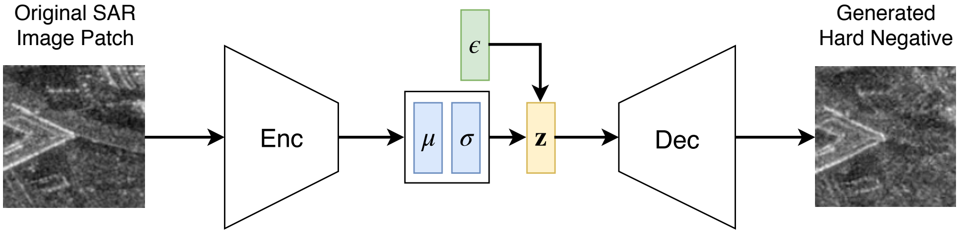


Figure 4. The inference network used to generate hard negative samples. The latent code **z** used to generate patches is created by sampling the latent distribution near to the original image. To keep the network end-to-end differentiable, this sampling is done via a re-parameterization trick using $\epsilon \sim \mathcal{N}(1, 0)$ to add randomness.

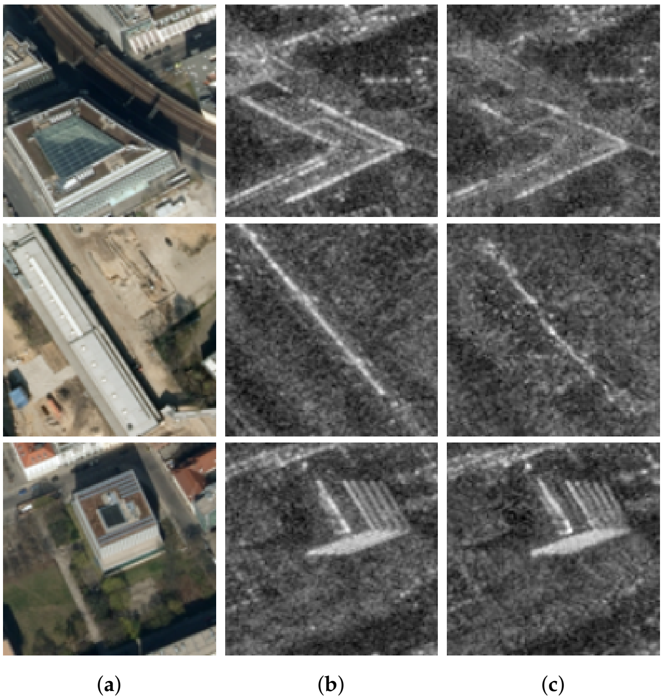


Figure 5. Using the proposed generative framework, we are able to generate SAR-like image patches (**c**) that can then be combined in conjunction with the original SAR patch (**b**), and a corresponding optical patch (**a**) in order to create a training dataset containing hard-negative samples.

3. Experiments and Results

In this section, we describe our experimental procedure and present results with respect to our network's ability to generate realistic SAR patches, and the suitability of these patches as hard negative samples for training the SAR-optical matching network of [14].

3.1. Dataset

We train our proposed hard negative GAN and the SAR-optical matching network of [14] on a dataset of corresponding unfiltered TerraSAR-X and UltraCam image patch pairs [17]. The patch pairs are generated from imagery taken of a study area in Berlin, Germany, which is depicted in Figure 6.

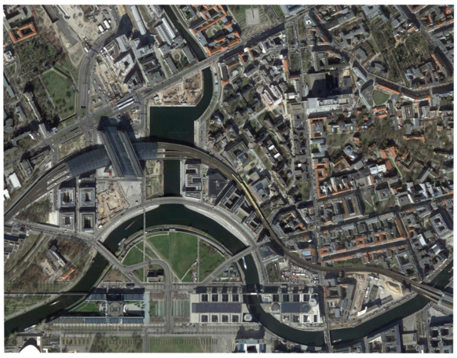


Figure 6. The common region of interest, in Berlin, Germany from which TerraSAR-X and UltraCam image patches were cut to generate the the SARptical dataset [17].

The dataset is deterministically split into a training, testing and validation set using the *cutting-cake* method proposed in [14]. Using this deterministically split dataset, we reduce the chances of the training and testing datasets having too similar distributions. Our datasets consist of 6629 (75%), 1327 (15%), and 885 (10%) corresponding image pairs for the training, testing and validation sets, respectively.

The non-corresponding pairs for the testing and validation datasets are created by assigning a randomly selected SAR image patch to each optical image. In doing so, we ensure that all experiments are subject to the same testing and validation datasets, and that our datasets are balanced in terms of corresponding and non-corresponding pairs.

The non-corresponding pairs for our training dataset are assigned according to the requirements of each experiment, in order to evaluate the success of our method.

The optical data was converted to gray-scale and all data were normalized to a radiometric range of $[0;1]$ and then standardized by subtraction of their means [14,19]. Furthermore, we make use of pair-wise data augmentation steps which include rotation, horizontal flipping, and translation.

3.2. Qualitative Evaluation of Generated Negative Samples

Measuring the quality of generated images is a challenging task, especially in the case of high resolution data [31]. Thus, we resort to a visual qualitative assessment of the generated hard negative SAR patches. These results can be seen in Figure 7.

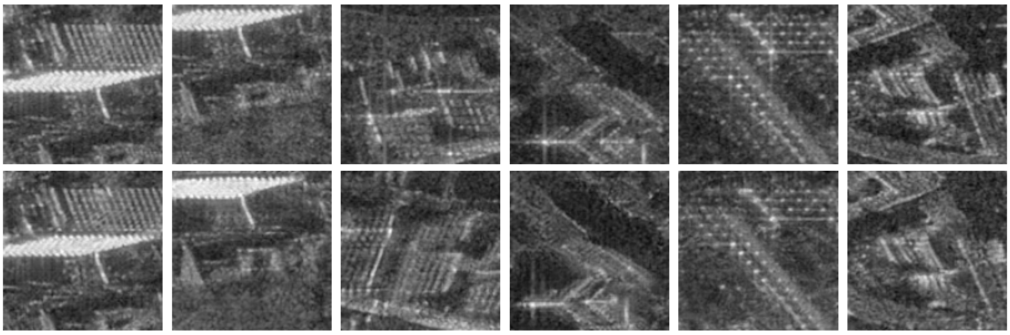


Figure 7. A selection of generated hard-negative samples (**bottom row**) and the corresponding training TerraSAR-X patch (**top row**). It can be seen that the generated patches have strong SAR-like features, which resemble those of the original patch, and are difficult to distinguish from the original patches.

3.3. Matching SAR and Optical Images

We apply our methods to the SAR-optical matching network proposed in [14]. This network has a pseudo-Siamese architecture which learns modal specific features for SAR and optical images in parallel. It then combines these features through a data fusion layer in order to obtain a prediction of whether the two patches match based on the content of the center pixel. The network architecture can be seen in Figure 8.

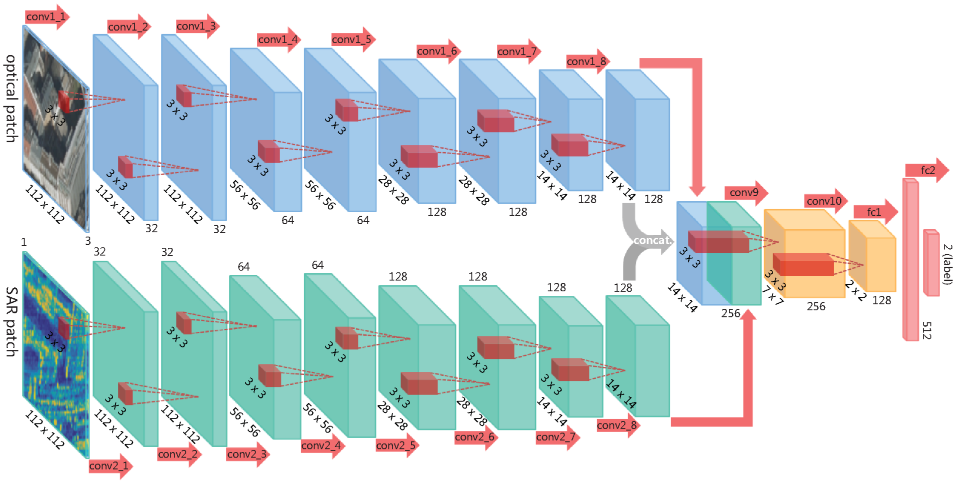


Figure 8. The pseudo-Siamese convolutional matching network proposed in [14]. The network attempts to predict the probability that the given input pair are corresponding in terms of the alignment of the center pixel in each patch. (© 2018 IEEE).

For all of our experiments, we train the network using the Adam optimizer with a learning rate of $\alpha_{\text{lr}} = 0.00005$. All of the networks are trained using early stopping based on the validation accuracy for a maximum period of 20 epochs.

3.4. Effect of a Hard Negative Inclusion Method

In order to evaluate our approach, we need to include our generated non-corresponding pairs into the existing training dataset. As we were unsure of the best approach for training the matching network with generated hard negatives, we evaluated two different training approaches with two dataset inclusion methods. These four approaches are defined below:

1. Fine-tuning with generated hard negatives,
2. Fine-tuning with concatenated dataset,
3. Training from scratch with generated hard negatives,
4. Training from scratch with concatenated dataset.

The generated hard negative dataset consists only of the original corresponding patch-pairs and their respective hard-negative patch-pairs, which were created as described in Section 2.3. In order to create the concatenated dataset, we combined the generated hard-negative dataset with the original training dataset in order to form a final dataset with both generated hard-negatives and randomly assigned hard-negatives for each of the corresponding patch-pairs. In order to keep the positive and negative classes balanced, we included each corresponding patch-pair twice.

To allow us to fine-tune the matching network, we first pre-trained it using the original training dataset which consists of randomly assigned negative patch-pairs for 30 epochs. This network was then fine-tuned using a lower learning rate of $\alpha = 0.000008$ and early stopping.

We evaluated the trained networks performance using the receiver–operator characteristic (ROC) in order to determine which approach leads to the most favourable results. The ROC curves for each of the four approaches are depicted in Figure 9.

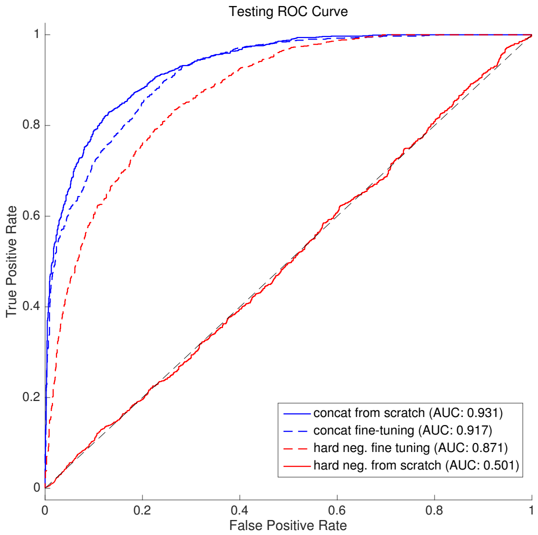


Figure 9. The receive operator characteristic (ROC) curves for various approaches for using generated hard-negatives in the training of a SAR-optical matching network. From these various experiments, it can be seen that including the generated hard-negatives into the original dataset leads to a matching network with better performance than using only generated negative samples for training.

A simple measure of the network's performance as a binary classifier is the area under the curve (AUC) of the ROC curve. From Figure 9, one can see that training the matching network used from scratch using a combined dataset provides the best performance. Thus, we select this approach as the proposed method of hard-negative inclusion, and will use it in further experiments.

3.5. Comparison to Existing Approaches

We compare the performance of the matching network trained using our proposed method to three alternative approaches, namely, random negative assignment, traditional hard-negative mining, and nearest neighbor assignment. These approaches are further detailed below:

Random negative assignment creates non-corresponding negative patch pairs by randomly selecting a SAR image patch from the patch pool and assigning it to a randomly selected optical patch. This random selection is done in a non-replacement manner such that every randomly created patch pair is unique and non-corresponding. This method is the most computationally efficient method for negative pair assignment, but it makes strong assumptions about the 'closeness' of the data.

Traditional hard-negative mining [32], starts training with random negative assignment and then iteratively updates the set of non-corresponding patch pairs at the end of each training epoch. The updates are performed by tracking the classification score of each patch-pair during training. The patch-pairs which were most severely mis-classified (non-corresponding pairs which were classified as corresponding with a high probability) are then explicitly labelled as negative pairs and added to the dataset, the remaining negatives pairs are reinitialized using random negative assignment. This process is computationally expensive and degrades to continuous random assignment when the false positive rate is low and/or the training dataset is small.

Nearest neighbor assignment [33], is a boostrapping method for hard-negative mining and is performed prior to training. For each positive patch pair, a non-corresponding pair is created by selecting the nearest neighbor SAR image from the training set. The nearest neighbor is defined as the image patch in the training dataset that has the greatest similarity to the positive image patch. In our case, we make use of the normalized cross correlation (NCC) score to determine which SAR images are most similar to each other. We then generate a non-corresponding patch pair using the SAR image with the greatest NCC score when compared to the positive = pair SAR image.

A detailed comparison of the results is presented in the form of an ROC comparison plot (see Figure 10). From the ROC plot, we can see that our approach provides a higher accuracy under the constraint of a low false positive rate. We further provide a detailed account of the precision and recall of the various approaches, as well as the respective accuracies when the decision boundary is tuned (on the validation set after training) to provide a maximum FPR of 5% or a maximum overall accuracy (see Table 3). From these results, our method is shown to boost the performance of the matching network on almost all fronts.

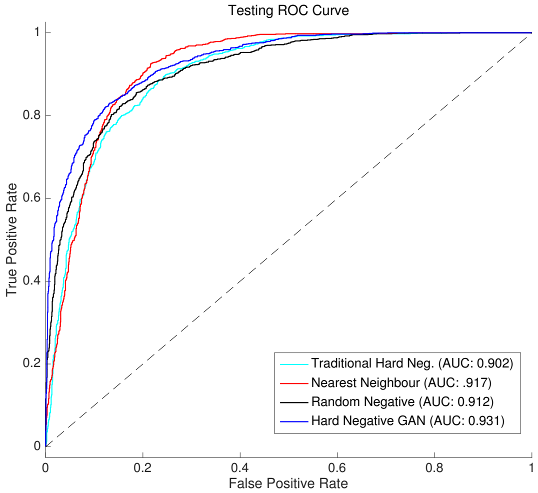


Figure 10. Comparison of training results, in the form of ROC curves, for the matching network performance on a test set when trained using randomly assigned negative pairs to three hard-negatives mining approaches. It can be seen that our proposed approach has a steeper onset than the other approaches, thus indicating better performance during matching.

Table 3. Details of SAR-optical matching results under the application of various hard-negative training strategies and at different false positive rates (FPR).

Method	Precision	Recall	Acc. (5% FPR)	Max Acc.	Max Acc. FPR
Random	**0.83**	0.84	0.76	0.83	0.16
Nearest Neighbour	0.77	**0.96**	0.70	0.85	0.21
Traditional Hard Neg.	0.79	0.89	0.72	0.83	0.19
Proposed Approach	**0.83**	0.87	**0.81**	**0.86**	**0.13**

4. Discussion

Generally, the results presented in Section 3 show that we are able to generate high quality hard-negative samples, and to use them to successfully train a SAR-optical matching network. The results further indicate that following this approach leads to a significant improvement in the overall performance and discriminability of the matching network, without the need for additional training data. In this section, we will further explore these results to gain a deeper understanding of the mechanisms at play.

4.1. Generative Ability

Considering the generated images presented in Figure 7, it is clear that our proposed generative framework is able to learn a diverse latent representation of the training dataset. By sampling this latent distribution, we are able to generate realistic VHR SAR-like images. The generated images are largely indistinguishable from the original TerraSAR-X patches and depict many SAR-like features such as layover, speckle, and radar shadow. Additionally, and arguably most importantly for our application, the images generated by sampling the posterior are visually similar to the original images but still contain novel components.

4.2. Effects of Data Inclusion Approach

As Figure 9 clearly indicates, the method of incorporating the generated hard negatives into the training procedure of the matching network plays a large role in the effectiveness of the approach. Training the network using both randomly assigned and generated non-corresponding pairs produced significantly better results than using only the generated samples. This is likely due to the generated samples adding sufficient variability to the dataset to act as independent datapoints, thus essentially increasing the size of the training dataset. This theory is backed up by the result of training from scratch using only the generated images as negative samples. In this case, it becomes clear that the validation and training dataset have a larger disparity in their distributions. Thus, apart from increasing the dataset size, training using both distributions likely has a regularization effect on the training of the network.

Additionally, Figure 9 shows that training from scratch is only a better approach when we include non-correspondences created using real data, even if these are just created using a random assignment approach. This is evident, from the case of training, that the matching network from scratch using only generated negative features where the network fails to learn any discriminative boundary that is suitable for matching real data. This is likely a consequence of the generated manifold being a Gaussian approximation to the original manifold and thus the distributions could have disjoint supports, and is subject to future investigation.

4.3. SAR-Optical Matching Performance

The comparison of our proposed approach to three alternative training methodologies shows promise for the use of generative hard-negative mining in improving matching performance in data sparse applications. As Figure 10 and Table 3 clearly indicate, training with generative hard-negative mining significantly improves the discriminative power and accuracy of the matching network when evaluated on an independent test dataset. Using this approach, we were able to train the matching network to achieve an accuracy exceeding 80% when the false positive rate is fixed to 5%. Additionally, the matching network was able to achieve an overall higher accuracy with a 3% point reduction in false positives.

The results of the traditional hard-negative mining agree with the literature, which states that the technique fails to add benefits if the dataset is significantly not large enough, as it effectively falls back to a random negative procedure [16,32]. Overall, this approach fails to improve any aspect of the original matching network, and in many ways preforms as a combination of the worst aspects of the other approaches, achieving an overall accuracy that matches that of the randomly assigned negatives, but with a worse false positive rate.

An interesting result is that of the nearest neighbor hard-negatives. These negative patches are assigned according to which image in the training dataset is the closest to the input image in a normalized cross-correlation (NCC) sense. As NCC is often used as a signal-based measure for multi-modal (including SAR-optical) image matching, it would appear to be a good choice for selecting hard-negatives. However, this approach produces the worst overall accuracy and false positive performance. It is suspected that the NCC hard-negatives cause the non-corresponding pairs to be too similar to the corresponding pairs, thus creating a matching problem which is too complex for the given network to resolve. This suspicion is further backed up by the high recall but low precision, which indicates that the network has become biased towards predicting patch pairs as corresponding (see Table 3).

4.4. Comments on Computational Overhead

Although our approach leads to the best performance of the SAR-optical matching network, this comes at the cost of a large computational overhead. This is caused by the fact that we need to train a relatively large and complex generative network. However, we can compute the dataset of negative samples prior to training the matching network, which allows for swapping out the online requirements of RAM and an additional GPU for additional storage capacity. In doing so, we reduce the computational burden of our approach to a once-off cost per dataset. Training this generative network for our small training dataset took 96 hours on a single NVidia GTX 1080 GPU. During training, the matching network using our offline approach took around 20 min on the same hardware.

On the other hand, traditional hard-negative mining directly impacts the computational cost of training the matching network. As it is performed on-the-fly, the computational burden persists across experiments and training operations. In the case of our investigation, the training time of the matching network increased to 25 min; however, this time increase grows along with the dataset size. Additionally, the training time memory requirements for the network increase as we need to keep a history of predicted labels for each item in the dataset so that items can be replaced by better hard-negative samples.

Thus, the added upfront computational expense of our approach may work better in environments with limited computational resources but sufficient storage capacity.

5. Conclusions

With this paper, we have proposed a generative framework for hard-negative mining that can be used in data sparse image matching applications to improve the discriminability and accuracy of the matching network. By combining the strong latent space encoding features of a variational autoencoder with the high quality generative capabilities of generative adversarial networks, we are able to produce realistic SAR-like image patches in a conditional manner. In doing so, we are able to produce a structurally similar, but novel SAR patch for each SAR image in our training dataset. We can then combine these SAR and SAR-like images with a corresponding optical image in order to create a balanced dataset of corresponding and non-corresponding patch pairs that can be used for training SAR-optical matching networks.

By applying this generative hard-negative approach to the existing SAR-optical matching network proposed in [14], we were able to confirm the capabilities of our approach in improving matching accuracy and reducing a false positive rate when tested on an independent dataset. Within the scope of sparse training data, our proposed method shows a significant improvement in matching accuracy at low FPRs and a small improvement in overall accuracy (but with a significant improvement in FPR) when compared to two commonly applied hard negative mining techniques.

Our generative hard-negative mining framework has applicability outside of the realm of SAR-optical matching. It is believed that this approach to hard-negative mining can be applied to many other problems that suffer from similar data constraints, both within and outside of remote sensing.

Author Contributions: Conceptualization, L.H.H. and M.S.; Methodology, L.H.H.; Software, L.H.H.; Data Curation and Investigation, L.H.H. and M.S.; Writing, L.H.H. and M.S.; Supervision, M.S. and X.Z.; Project Administration, M.S.; Funding Acquisition, M.S.; Resources: M.S. and X.Z.

Funding: This work was supported by the German Research Foundation (DFG), grant SCHM 3322/1-1, the Helmholtz Association under the framework of the Young Investigators Group SiPEO (VH-NG-1018, www.sipeo.bgu.tum.de), and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. ERC-2016-StG-714087, Acronym: *So2Sat*).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Schmitt, M.; Zhu, X.X. Data Fusion and Remote Sensing—An Ever-Growing Relationship. *IEEE Geosci. Remote Sens. Mag.* **2016**, *4*, 6–23. [CrossRef]
2. Schmitt, M.; Tupin, F.; Zhu, X.X. Fusion of SAR and optical remote sensing data—Challenges and recent trends. In Proceedings of the 2017 IEEE International Geoscience and Remote Sensing Symposium, Fort Worth, TX, USA, 23–28 July 2017; pp. 5458–5461.
3. Ye, Y.; Shan, J.; Bruzzone, L.; Shen, L. Robust Registration of Multimodal Remote Sensing Images Based on Structural Similarity. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 2941–2958. [CrossRef]
4. Li, J.; Hu, Q.; Ai, M. RIFT: Multi-modal Image Matching Based on Radiation-invariant Feature Transform. *arXiv* **2018**, arXiv:1804.09493.
5. Qiu, C.; Schmitt, M.; Zhu, X.X. Towards automatic SAR-optical stereogrammetry over urban areas using very high resolution imagery. *ISPRS J. Photogramm. Remote Sens.* **2018**, *138*, 218–231. [CrossRef] [PubMed]
6. Palubinskas, G.; Reinartz, P.; Bamler, R. Image acquisition geometry analysis for the fusion of optical and radar remote sensing data. *Int. J. Image Data Fusion* **2010**, *1*, 271–282. [CrossRef]
7. Balntas, V.; Johns, E.; Tang, L.; Mikolajczyk, K. PN-Net: Conjoined triple deep network for learning local image descriptors. *arXiv* **2016**, arXiv:1601.05030.
8. Yi, K.M.; Trulls, E.; Lepetit, V.; Fua, P. Lift: Learned invariant feature transform. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016; pp. 467–483.
9. Zagoruyko, S.; Komodakis, N. Deep compare: A study on using convolutional neural networks to compare image patches. *Comput. Vis. Image Underst.* **2017**, *164*, 38–55. [CrossRef]
10. Han, X.; Leung, T.; Jia, Y.; Sukthankar, R.; Berg, A.C. Matchnet: Unifying feature and metric learning for patch-based matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3279–3286.
11. Zhu, X.X.; Tuia, D.; Mou, L.; Xia, G.S.; Zhang, L.; Xu, F.; Fraundorfer, F. Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 8–36. [CrossRef]
12. Merkle, N.; Luo, W.; Auer, S.; Müller, R.; Urtasun, R. Exploiting Deep Matching and SAR Data for the Geo-Localization Accuracy Improvement of Optical Satellite Images. *Remote Sens.* **2017**, *9*, 586. [CrossRef]
13. Mou, L.; Schmitt, M.; Wang, Y.; Zhu, X. A CNN for the Identification of Corresponding Patches in SAR and Optical Imagery of Urban Scenes. In Proceedings of the 2017 Joint Urban Remote Sensing Event (JURSE), Dubai, UAE, 6–8 March 2017.
14. Hughes, L.H.; Schmitt, M.; Mou, L.; Wang, Y.; Zhu, X.X. Identifying Corresponding Patches in SAR and Optical Images With a Pseudo-Siamese CNN. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 784–788. [CrossRef]
15. Merkle, N.; Auer, S.; Müller, R.; Reinartz, P. Exploring the Potential of Conditional Adversarial Networks for Optical and SAR Image Matching. *IEEE J-STARS* **2018**, *11*, 1811–1820. [CrossRef]
16. Sung, K.K.; Poggio, T. Example-based learning for view-based human face detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **1998**, *20*, 39–51. [CrossRef]
17. Wang, Y.; Zhu, X.; Zeisl, B.; Pollefeys, M. Fusing meter-resolution 4-D InSAR point clouds and optical images for semantic urban infrastructure monitoring. *IEEE Trans. Geosci. Remote Sens.* **2016**, *55*, 14–26. [CrossRef]
18. Auer, S.; Hornig, I.; Schmitt, M.; Reinartz, P. Simulation-based Interpretation and Alignment of High-Resolution Optical and SAR Images. *IEEE J-STARS* **2017**, *10*, 4779–4793. [CrossRef]
19. Wang, Y.; Zhu, X.X. The SARptical Dataset for Joint Analysis of SAR and Optical Image in Dense Urban Area. *arXiv* **2018**, arXiv:1801.07532.
20. Ding, J.; Chen, B.; Liu, H.; Huang, M. Convolutional neural network with data augmentation for SAR target recognition. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 364–368. [CrossRef]
21. Zheng, Z.; Zheng, L.; Yang, Y. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. *arXiv* **2017**, arXiv:1701.07717.
22. Ley, A.; d'Hondt, O.; Valade, S.; Hänsch, R.; Hellwich, O. Exploiting GAN-based SAR to optical image transcoding for improved classification via deep learning. In Proceedings of the 12th European Conference on Synthetic Aperture Radar, Aachen, Germany, 4–7 June 2018; pp. 396–401.
23. Wang, P.; Li, S.; Pan, R. Incorporating GAN for Negative Sampling in Knowledge Representation Learning. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.

24. Marmanis, D.; Yao, W.; Adam, F.; Datcu, M.; Reinartz, P.; Schindler, K.; Wegner, J.D.; Stilla, U. Artificial generation of big data for improving image classification: A generative adversarial network approach on SAR data. *arXiv* **2017**, arXiv:1711.02010.
25. Ao, D.; Dumitru, C.O.; Schwarz, G.; Datcu, M. Dialectical GAN for SAR Image Translation: From Sentinel-1 to TerraSAR-X. *arXiv* **2018**, arXiv:1807.07778.
26. Karras, T.; Aila, T.; Laine, S.; Lehtinen, J. Progressive Growing of GANs for Improved Quality, Stability, and Variation. *arXiv* **2018**, arXiv:1710.10196.
27. Khan, S.H.; Hayat, M.; Barnes, N. Adversarial Training of Variational Auto-encoders for High Fidelity Image Generation. *arXiv* **2018**, arXiv:1804.10323.
28. Larsen, A.B.L.; Sønderby, S.K.; Larochelle, H.; Winther, O. Autoencoding beyond pixels using a learned similarity metric. *arXiv* **2015**, arXiv:1512.09300.
29. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
30. Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; Courville, A.C. Improved training of Wasserstein GANs. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5769–5779.
31. Theis, L.; van den Oord, A.; Bethge, M. A note on the evaluation of generative models. In Proceedings of the 2016 International Conference on Learning Representations, San Juan, Puerto Rico, 2–4 May 2016.
32. Sung, K.K. *Learning and Example Selection for Object and Pattern Detection*; Computer Science and Artificial Intelligence Lab: Cambridge, MA, USA, 1996.
33. Jiang, F. SVM-Based Negative Data Mining to Binary Classification. Ph.D. Thesis, Georgia State University, Atlanta, GA, USA, 2006.

4.5 A Semi-Supervised Approach to SAR-Optical Image Matching

ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume IV-2/W7, 2019
PIA19+MRSS19 – Photogrammetric Image Analysis & Munich Remote Sensing Symposium, 18–20 September 2019, Munich, Germany

A SEMI-SUPERVISED APPROACH TO SAR-OPTICAL IMAGE MATCHING

L. H. Hughes, M. Schmitt

Signal Processing in Earth Observation, Technical University of Munich, Munich, Germany
- (lloyd.hughes, m.schmitt)@tum.de

KEY WORDS: Image Matching, Synthetic Aperture Radar (SAR), Optical Remote Sensing, Deep Learning, Deep Matching, Semi-supervised Learning

ABSTRACT:

Matching synthetic aperture radar (SAR) and optical remote sensing imagery is a key first step towards exploiting the complementary nature of these data in data fusion frameworks. While numerous signal-based approaches to matching have been proposed, they often fail to perform well in multi-sensor situations. In recent years deep learning has become the go-to approach for solving image matching in computer vision applications, and has also been adapted to the case of SAR-optical image matching. However, the hitherto proposed techniques still fail to match SAR and optical imagery in a generalizable manner. These limitations are largely due to the complexities in creating large-scale datasets of corresponding SAR and optical image patches. In this paper we frame the matching problem within semi-supervised learning, and use this as a proxy for investigating the effects of data scarcity on matching. In doing so we make an initial contribution towards the use of semi-supervised learning for matching SAR and optical imagery. We further gain insight into the non-complementary nature of commonly used supervised and unsupervised loss functions, as well as dataset size requirements for semi-supervised matching.

1. INTRODUCTION

The collection and exploitation of complementary information from multi-modal data sources enables a deeper understanding of the world and is critical in many applications across multiple domains. A key first step in any data fusion process is determining correspondences among these data sources in order to align and further exploit the complementary information in each modality (Schmitt and Zhu, 2016). In the case of image-based data fusion this relates to determining corresponding image regions across images which may have been acquired by different sensors, at different viewpoints and at various resolutions.

While the task of determining correspondences in conventional computer vision applications, such as structure from motion and pose estimation, has seen great progress and is solved to the degree of being usable operationally, it is still an open and relevant problem in the field of remote sensing. This is especially true when considering the case of determining correspondences in highly complementary, but vastly different image sources such as between synthetic aperture radar (SAR) and optical imagery (Schmitt et al., 2017).

As can be seen in Figure 1, the vastly different image acquisition schemes of SAR and optical sensors lead to imagery that not only depicts different properties of a scene, but also contains significantly different geometric distortions and imaging artifacts. Synthetic aperture radar imagery captures the physical characteristics of a scene, such as surface roughness or water content, while optical imagery provides details as to the chemical composition of the target area. Furthermore, SAR imagery suffers from imaging artifacts such as speckle, layover and radar shadow - none of which are present in optical imagery. These vast differences make determining correspondences between the data a challenging task.

Although many traditional feature matching methods have been proposed for matching SAR and optical data, e.g. (Ye and Shen, 2016, Ye et al., 2017, Dellinger et al., 2015), many of them still exhibit sub-optimal performance especially in high and very high resolution imagery. The advent and success of deep learning in developing robust solutions to the correspondence problem in traditional computer vision settings, e.g. by (Han et al., 2015, Zagoruyko and Komodakis, 2017), has lead to its application to multi-modal matching within remote sensing, e.g. in (Mou et al., 2017, Merkle et al., 2017a, Hughes et al., 2018b). Despite deep networks being universal function approximators, the results of their application to the SAR and optical matching problem have been mixed and with varying degrees of robustness and generalizability. These effects can be attributed to three main challenges:

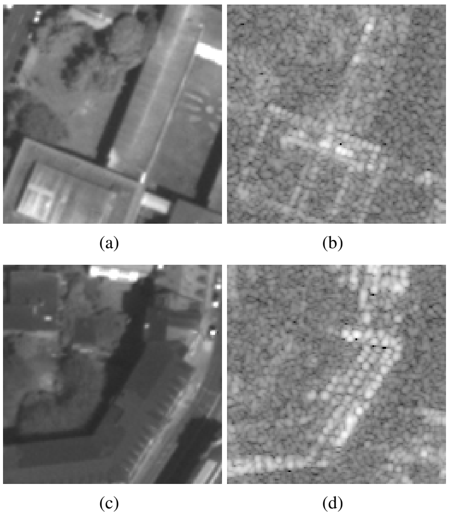


Figure 1: An example of corresponding SAR and optical patch pairs. Matching the image pairs in (**a**,**b**) and (**c**,**d**) proves to be a challenging task, even for domain experts.

https://doi.org/10.5194/isprs-annals-IV-2-W7-71-2019 | © Authors 2019. CC BY 4.0 License.

firstly the intractability of creating large-scale annotated datasets due to SAR imagery being difficult, even for experts, to interpret; secondly the complex nature of SAR image formation which prevents the creation of realistic, synthetic datasets and finally the natural ineffectiveness of transfer learning techniques to extract meaningful feature representations from SAR, and lesser so, from optical space-borne imagery. These factors are all directly impacting on the feasibility of training the complex deep networks required to accurately determine correspondences between complex multi-modal data sources such as SAR and optical data.

To this end, we propose the use of semi-supervised learning to relax the requirements for large-scale labeled data in order to learn a well-generalizing SAR-optical image matching network. As semi-supervised learning has not yet been applied within this domain, the question still remains as to how much labeled data is required, and how well features learned in an unsupervised manner generalize to support supervised tasks. Additionally, we strive to understand the effects of data scarcity on the accuracy of learned SAR-optical descriptors, and the interplay between the unsupervised and supervised objectives. The main contributions of this paper can be summarized as follows: We formulate a semi-supervised approach to SAR-optical image matching and use this approach as a framework to assess the relative effect of data scarcity on the network's ability to learn meaningful descriptors for SAR-optical image matching.

2. RELATED WORK

2.1 Deep Learning for SAR-Optical Matching

Deep learning is becoming an increasingly important method in the toolbox of remote sensing practitioners, especially in the area of data fusion, and thus also SAR-optical matching (Zhu et al., 2017).

The first notable examples of this were provided in short succession by (Merkle et al., 2017b) and (Mou et al., 2017) who both proposed variants of a 2-stream architecture. (Merkle et al., 2017b) trained a siamese network to predict the relative shift between SAR and optical patches in order to improve the geolocalization accuracy of the optical data, while (Mou et al., 2017) trained a pseudo-siamese variant as a binary correspondence classifier. Taking inspiration from these seminal works, we extended the network proposed by (Mou et al., 2017) by enhancing the feature fusion stage and converting the output to a similarity score based on the soft-max probability (Hughes et al., 2018b).

Taking a different approach to the problem, (Merkle et al., 2018) proposed the use of a generative adversarial network (GAN) to generate SAR-like templates from optical image patches. These templates were then used as input to standard template matching approaches such as mutual information (MI) or normalized cross correlation (NCC).

These works all make use of supervised learning, which require large-scale labeled datasets – in this case, corresponding SAR-optical patch pairs. As such many of them lack robustness and generalizability, due to the intractability of creating large datasets of pixel-wisely matched VHR imagery of urban scenes.

In an attempt improve on this, we proposed a novel hard-negative mining strategy which does not increase the requirements for training data in previous work (Hughes et al., 2018a). To do this, we trained a conditional GAN to generate SAR patches which could be used directly, along with a corresponding optical image, for hard-negative mining. However, this approach is computationally expensive and does not completely resolve the problems caused by the scarcity of labeled data in SAR-optical matching problems.

2.2 Semi-supervised Learning

Semi-supervised learning constitutes a set of techniques for exploiting large-scale unlabeled datasets in order to support the learning in environments where labeled data is scarce (Chapelle et al., 2009). While many such methods exist, they all are centered around the same basic principles. Namely, to exploit unlabeled data in an unsupervised, or self-supervised manner to learn generalizable features, and to use small amounts of labeled data to steer learning towards a specific task.

(Zhang et al., 2016) and (Rasmus et al., 2015) proposed combining supervised classification with an unsupervised autoencoder-based reconstruction loss for image recognition. (Lai et al., 2017) trained a deep network using an adversarial loss to predict the flow field between a pair of images. This method used sparse depth information from LiDAR for supervision, while using an image consistency loss for unsupervised training. (Mukherjee et al., 2017) proposed the use of deep matching autoencoders to learn a common latent space between multi-modal data. This was achieved using a statistical dependency measure to pair unlabeled data during training and supervised with corresponding training pairs. Using a multi-phase training approach (Bui et al., 2018) pretrained a classifier for each domain in a supervised manner and then used a second training phase to learn a transformation between the learned embeddings for cross-domain image retrieval.

Autoencoders and reconstruction losses form a fundamental part of many semi-supervised learning approaches. However, they are still most often used as an auxiliary loss in supervised learning for matching multi-modal data (Ngiam et al., 2011, Liu et al., 2018). This is largely due to increased complexity of semi-supervised learning and the fact that these techniques lend themselves best to well conditioned problems (Cholaquidis et al., 2018). While the image matching problem is known to be ill-conditioned, autoencoders have still shown success in the domain of supervised learning for multi-modal matching. Thus in this paper, we will propose extensions to supervised autoencoder based matching techniques to allow for semi-supervised learning in within this domain.

3. SEMI-SUPERVISED SAR OPTICAL MATCHING

In this section, we describe our proposed SAR-optical matching network, including the use of autoencoders for semi-supervised learning of descriptors from labeled and unlabeled data, and the use of an adversarial loss for aligning these descriptor latent spaces. Further, we describe the training procedure and how matching can be achieved using the final trained network. An overview of the proposed architecture can be seen in Figure 2.

3.1 Network Architecture

In a similar vein to the matching networks proposed by (Liu et al., 2018) and (Mukherjee et al., 2017), we propose a dual autoencoder network in order to learn SAR and optical descriptors which can later be matched in a computationally efficient manner. In doing so we are able to exploit the self-supervised nature

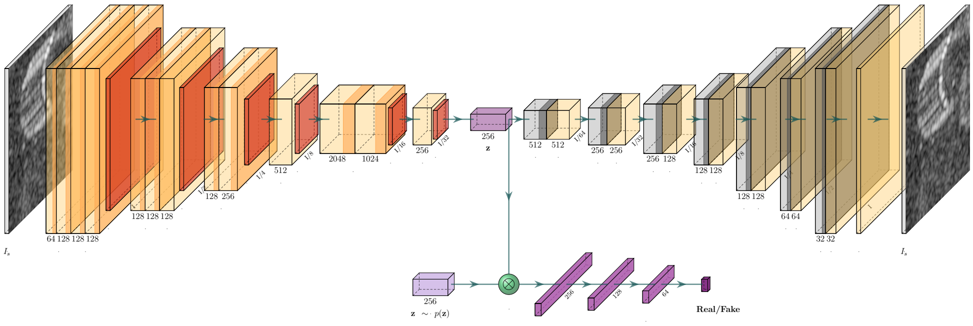


Figure 2: A single branch of the proposed network architecture. The autoencoder learns a meaningful latent code space $\mathbf{z}$ by learning to reconstruct the input image, while the discriminator network conditions the distribution of the latent codes using adversarial training and an arbitrary prior distribution. The optical branch is an exact mirror of the SAR branch and the discriminator network is shared between the branches.

of autoencoders to learn useful features from unpaired SAR and optical imagery. Furthermore, we use the latent code generated in the bottleneck as a natural descriptor and jointly train each domain specific autoencoder to align these latent codes. This alignment is achieved through the incorporation of a supervised loss function which is optimized using a small dataset of corresponding SAR-optical patch pairs.

Autoencoders typically consist of two networks, namely, an encoder and a decoder. Our proposed encoder network is based on the VGG11 (Simonyan and Zisserman, 2015) architecture. This architecture was chosen as a base due to its relative simplicity and low number of parameters. Furthermore, it has been used as a base to achieved state-of-the-art results in a variety of tasks (Iglovikov and Shvets, 2018), and is thus considered to be a good starting point for the exploration of semi-supervised learning for SAR-optical matching. The decoder network is based on a combination of convolutional and transposed convolution layers which are used to upsample the latent code in order to reconstruct the original image. The autoencoders for each modality (i.e. SAR and optical) have identical architectures and do not share any layers or weights. This allows for the learning of modality-specific features. As shown in Figure 2, the encoder network consists of blocks of 3×3 convolutions, batch normalization and activation with a Leaky ReLU function with a negative slope of 0.2. Similarly, the decoder network is made up of blocks of 3×3 transposed convolutions with a stride of 2 and ReLU activation, followed by a 3×3 convolutional layer and a ReLU activation. The depth of the feature maps are detailed in Figure 2.

For a given a SAR-optical image pair I_s, I_o we train the encoders $\mathrm{Enc}_s, \mathrm{Enc}_o$ to generate a descriptive latent code, $\mathbf{z}_s$ or $\mathbf{z}_o$ respectively, such that the decoder networks, $\mathrm{Dec}_s, \mathrm{Dec}_o$, can create an approximate reconstruction of the original inputs from the latent code. For a non-corresponding SAR-optical patch pair we seek to minimize the reconstruction loss such that,

$$\mathcal{L}_{recon} = \|I_s - \tilde{I}_s\|_2 + \|I_o - \tilde{I}_o\|_2, \tag{1}$$

where $\tilde{I}_s$ and $\tilde{I}_o$ are the reconstructed images generated by

$$\mathbf{z} \sim \mathrm{Enc}(I), \tag{2}$$

$$\tilde{I} \sim \mathrm{Dec}(\mathbf{z}) \tag{3}$$

using the appropriate, domain specific encoder and decoder networks.

For a pair of images labeled as either corresponding or non-corresponding, we augment the reconstruction loss, $\mathcal{L}_{recon}$, with a contrastive matching loss,

$$\mathcal{L}_{match} = y(\|\mathbf{z}_o - \mathbf{z}_s\|_2^2) + (1-y)\{\max\left(0, m - \|\mathbf{z}_o - \mathbf{z}_s\|_2^2\right)\}, \tag{4}$$

where y is the target label (zero for non-corresponding and one for corresponding), and m is the margin. The contrastive loss encourages the network in learning a latent space where corresponding pairs are near to each other, while non-corresponding pairs have a squared norm distance of at least margin m (Chopra et al., 2005). To ease the tuning of the margin hyperparameter, we took the L_2 norm of the each of the descriptor vectors $\mathbf{z}_o$ and $\mathbf{z}_s$ prior to the calculation of the contrastive loss. This ensures that both descriptors are on the hypersphere before matching and allows the use of normalized measures such as the cosine distance for matching the descriptors. This is significantly more efficient than descriptor-specific matching networks as the descriptors can be precomputed for each image patch.

In the end, the semi-supervised matching network is trained by minimizing the respective reconstruction losses $\mathcal{L}_{recon}$ for all SAR and optical data (paired and unpaired), while additionally minimizing the matching loss $\mathcal{L}_{match}$ for labeled, i.e. paired, data:

$$\mathcal{L}_{semisuper} = \sum_{i \in D_a} \left[\mathcal{L}_{recon}\left(I_s^i, \tilde{I}_s^i\right) + \mathcal{L}_{recon}\left(I_o^i, \tilde{I}_o^i\right)\right] + \sum_{j \in D_l} \mathcal{L}_{match}\left(\mathrm{Enc}(I_s^j), \mathrm{Enc}(I_o^j)\right), \tag{5}$$

where D_a and D_l represent the datasets of all, and labeled (corresponding and non-corresponding) SAR-optical patch pairs, respectively. Optimizing both the modality-specific reconstruction loss as well as the joint matching loss enables the network to learn to extract important features and generate descriptive latent codes from unlabeled data, while learning to align these latent spaces using a smaller labeled dataset.

While autoencoders are capable of learning complex data manifolds, these manifolds are often poorly conditioned with weak supports. Thus they often do not extend well to unseen data, such as imagery with a slightly different data distribution or from a different spatial region. This is due to the fact that the manifold is only smooth near to existing samples, i.e. the training samples. To reduce these effects, and simplify the alignment between the modality specific latent distributions we propose to impose a continuous prior distribution $p(\mathbf{z})$ on the respective latent codes. This is realized through the reformulation of our modality specific autoencoders as adversarial autoencoders with a joint adversary, and is described in the following.

3.2 Adversarial Training

An adversarial autoencoder is an autoencoder which is regularized by matching the generated posterior $q(\mathbf{z})$ to an arbitrary prior $p(\mathbf{z})$. This is achieved through a min-max game in which the generator network, the encoder (Enc) of the autoencoder, learns to maximize the error of a discriminator network (Dis), while the discriminator learns to minimize the classification error of samples coming from the prior and the posterior (Makhzani et al., 2016). This objective function can be expressed as:

$$\min_{Enc}\max_{Dis} \underset{\mathbf{z}\sim p(\mathbf{z})}{E}[\log(\text{Dis}(\mathbf{z}))] + \underset{I\sim D_a}{E}[\log(1-\text{Dis}(\text{Enc}(I)))]. \tag{6}$$

In order to prevent the discriminator being able to learn the prior and posterior distributions too easily, the discriminator network is kept relatively shallow and simplistic. In our case, the discriminator is comprised of three fully connected layers of decreasing size, each of which is followed by a Leaky ReLU activation with a negative gradient of 0.2. The last layer of the discriminator uses a sigmoid activation to classify the input vector as either coming from the prior or posterior distribution. This network structure can be seen in Figure 2.

As we wish for the SAR and optical latent spaces to be aligned, such that corresponding pairs appear nearby in the code space, we impose the same prior on both latent distributions and solve the min-max problem over both encoders and the shared discriminator. This is done by alternating between updating the discriminator network and the generator (encoder) network using samples from the full dataset of labeled and unlabeled SAR-optical pairs.

Due to instabilities which can arise during the optimization of the min-max game (Equation 6), we replace the traditional generative adversarial loss with a Wasserstein-distance-based loss (Gulrajani et al., 2017). The Wasserstein loss strives to optimize the min-max game in terms of distributions rather than directly as a classification problem, and is thus more robust against gradient explosion and problems of mode collapse. Thus our final semi-supervised matching network is trained by minimizing the discriminator and autoencoder objective functions,

$$\mathcal{L}_{dis} = \sum_{i\in D_a}\Big(\text{Dis}(\text{Enc}(I_s^i)) + \text{Dis}(\text{Enc}(I_o^i)) - 2\left(\text{Dis}(\mathbf{z}_p^i)\right)\Big), \tag{7}$$

$$\mathcal{L}_{ae} = \mathcal{L}_{hnet} - \sum_{i\in D_a}\Big(\text{Dis}(\text{Enc}(I_s^i)) + \text{Dis}(\text{Enc}(I_o^i))\Big), \tag{8}$$

where $\mathbf{z_p^i}$ is a sample from an arbitrary prior distribution $p(\mathbf{z})$. In our case, we define $p(\mathbf{z})$ as a normal distribution such that $p(\mathbf{z}) \sim \mathcal{N}(0, 5)$.

3.3 Implementation Details

We implement the proposed approach using the PyTorch deep learning framework (Paszke et al., 2017). The optimization of the autoencoders is performed using the Adam solver with $\beta_1 = 0.9$ and $\beta_2 = 0.99$ and a weight decay of 10^{-4}. The discriminator network is optimized using stochastic gradient descent (SGD) with a momentum of 0.9, weight decay equal to $3 \cdot 10^{-4}$ and a learning rate of $4 \cdot 10^{-3}$.

The learning rate for the Adam optimizer was determined using the search method proposed by (Smith and Topin, 2017), whereby the learning rate is rapidly increased from a small value, 10^{-7}, over consecutive batches while the loss is recorded. The learning rate is then selected to be in the region where the loss decreased in a smooth and constant manner (region of highest gradient). Using this approach we found the optimum learning rate for the Adam optimizer to be in the range of $5 \cdot 10^{-5}$ and $5 \cdot 10^{-4}$. This learning rate range was then used to initialize a one-cycle policy learning rate scheduler to dynamically vary the learning rate during training (Smith and Topin, 2017). The full network was then trained in an end-to-end manner for 100 epochs with a batch size of 32.

To improve the stability of the adversarial training the discriminator was trained using an update schedule with five times the frequency of that of the generator. Furthermore, the discriminator weights were clipped to the range of $[-0.1, 0.1]$ in order to preserve the *1-Lipschitz* constraints required for the Wasserstein loss (Petzka et al., 2017, Gulrajani et al., 2017).

Data augmentation was used to improve generalization and prevent overfitting due to the relatively small supervised dataset which we used. The data augmentation scheme included 1) horizontal and vertical flipping with a probability of 0.5 for each corresponding image pair, 2) the addition of Gaussian white noise with a standard deviation of $\sigma = 0.02$ to the optical image, and 3) scaling of image intensities by a randomly selected factor of $[0.95, 1.05]$, with a probability of 0.2. In order to preserve the accuracy of the labeled dataset, the same flipping and scaling transformations were applied to each image in the image pair. For the unlabelled dataset, these transformations are applied independently to each image.

4. EXPERIMENTS

4.1 Experimental Setup

As large-scale SAR-optical correspondence datasets are difficult to produce for very high resolution imagery, especially in urban areas, we make use of the UrbanAtlas dataset and reduce the region of interest for matching to areas which are mainly comprised of rural and semi-urban areas. In doing so we can limit the geometric differences between the SAR and optical imagery, and thus can derive corresponding points using the geo-localization information. While this approach may contain inaccuracies, these are assumed to be small at the spatial resolution of the dataset.

The UrbanAtlas dataset is comprised of high resolution ($2.5m$ GSD) TerraSAR-X and PRISM imagery of 23 cities across Europe. In order to increase the probability of salient features being present in both images we applied a Harris corner detector to the optical domain and applied a non-maximal suppression filter with a spatial constraint to ensure a minimum distance of 128

ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume IV-2/W7, 2019
PIA19+MRSS19 – Photogrammetric Image Analysis & Munich Remote Sensing Symposium, 18–20 September 2019, Munich, Germany

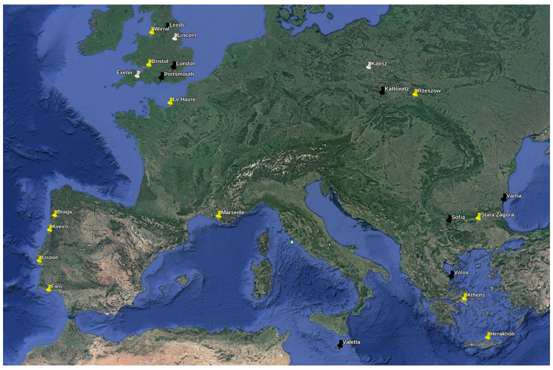


Figure 3: The distribution of the cities which were used for training (yellow), testing (black) and validation (white).

pixels between feature points. These feature points were then used as the center point when cutting SAR and optical patches, of 256 × 256 pixels, from the scenes. For training we extracted 50,000 patch pairs from 12 cities; 10,000 patch pairs from 3 cities for validation and 10,000 patch pairs from 8 cities for testing. The distribution of the cities into training, testing and validation sets is depicted in Figure 3.

In order to optimize the supervised loss we require both positive and negative training pairs. In order to achieve this we utilized a center crop of 128×128 pixels as the positive training pair, and an off center random crop of 128×128 to form a non-corresponding negative pair. The motivation for cropping both the positive and negative pair from the same patch was that nearby regions are likely to be more similar, giving the negative pair a similar distribution to the positive pair. This is expected to provide harder negative examples than purely random patch selection.

During pre-processing, all image patches were scaled to the range $[0, 1]$ and then standardized to zero mean and one standard deviation using the normal distributions as calculated from the SAR and optical images of the training set, i.e. $\mathcal{N}_{SAR}(0.5, 0.2)$ and $\mathcal{N}_{Opt}(0.45, 0.15)$. All other hyper-parameters were kept fixed for each scenario, such that the only variable was the degree of supervision.

For prediction at test time, we make use of a sliding window search procedure with a fixed optical template patch and a 256 × 256 SAR image search region. Matching is performed by calculating a descriptor for the central optical patch, and comparing this to the descriptors generated from a 128 × 128 sliding window over the SAR image. Thus, we obtain a $256 \times 128 \times 128$ descriptor volume for the SAR search region. The final correspondence map is then computed by calculating the cosine similarity between the descriptor volume and the descriptor of the optical template patch.

4.2 Matching under Data Scarcity

In order to assess our proposed network's ability to learn robust and discriminative features under conditions of data scarcity, we train the network with varying degrees of supervision. This further allows us to assess the effects of data scarcity on training the network, as well as the dynamics between the supervised and unsupervised loss functions.

We split the training dataset into supervised and unsupervised subsets with ratios of 100%, 75%, 50%, 25% and 5% supervised data to unsupervised data. The supervised subset is then oversampled to ensure that the distribution remains balanced. The network is then trained using alternating batches of unsupervised and supervised data.

The results of matching for these various scenarios are depicted as histograms/density functions of the pixel distance between the detected matching point and the ground truth location, as seen in Figure 4.

From Figure 4 it is clear that there is a non-linear relationship between the level of supervision and the number of well matched pairs. This relationship is particularly evident when observing the 1-percentile for each of the scenarios. The overall shape of the distribution should be noted too as it provides important insights into the network's matching abilities.

Due to the complexities of matching SAR and optical imagery it is expected that matching efforts will only yield a few correspondences. Thus it is often easier to obtain an intuition for the performance of a matching algorithm through a qualitative investigation of the correspondence maps for successful and unsuccessful matches. To this end Figure 5 depicts a few such examples for test scenes of varying building density and difficulty.

In an ideal matching scenario we would expect the correspondence maps, as shown in Figure 5, to have a single point of correspondence (red pixel) at the center, with the values at other offsets being relatively low in comparison (blue). However, in reality it is much more common to see a Gaussian like spread around the point of correspondence, with the peak value indicating the correct shift for maximal correspondence. From Figure 5 we can clearly see these point spread functions which depict the point of correspondence.

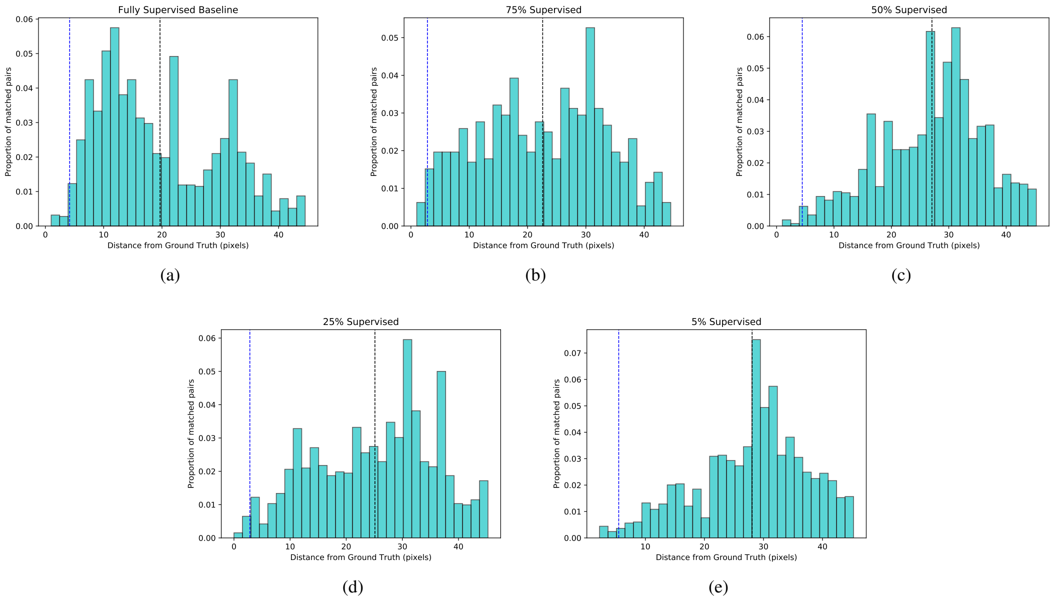
5. DISCUSSION

5.1 Semi-Supervised Matching

The examples in Figure 5 were selected as a fair depiction of the range of results which were obtained. From these examples, and in a qualitative manner, it is clear that the network is able to achieve SAR-optical matching, specifically in rural and semi-urban areas, across many levels of supervision.

On the other hand, the number of accurately matched points remains low, as evident from Figure 4. However, a large majority of data fusion tasks (such as stereogrammetry or image registration) require only a few reliable matches, i.e. they rely on a low false positive rate instead of only a high true positive rate. In conjunction, a high number of false negatives does not negatively impact follow-on applications.

The low number of detected correspondences is related to the vast differences in geometry between SAR and optical imagery which leads to salient points in the optical domain not always being visible in the SAR domain. Thus, the matching of these specific points becomes intractable even in the case of a fully supervised approach – which by nature of having more examples to learn from – should perform better than a semi-supervised approach. This outcome is also depicted in results corresponding to the SAR scene in Figure 5c, whereby the sharp edges and corners of the



(a)

(b)

(c)

(d)

(e)

Figure 4: Histograms reflecting the precision of the determined matched point when compared to the ground truth location for varying degrees of supervision. The dashed black line represents the mean matching distance while the dashed blue line represents the 1-percentile matching distance.

building in the optical domain is not clearly visible in the SAR domain.

The relative consistency of these correspondence maps, across multiple levels of data scarcity, support the hypothesis that using a shared adversary and supervised objective function, we are able to align these latent spaces in a meaningful way for cross domain matching; even with very little data.

Furthermore, we note from Figure 5 that the spread of the correspondence peak appears to grow as we decrease the amount of supervision. This is providing insight into the increased uncertainty in the matching process as the latent distributions are only aligned at a small number of locations. Furthermore, and perhaps more importantly, we note that in the case of failed correspondences the correspondence map no longer represents a Gaussian like distribution and instead becomes multi-modal or somewhat random – as depicted in the results corresponding to Figure 5c. This observation could perhaps be exploited in future work to filter out failed correspondences, or to design more sophisticated correspondence point selection schemes; as selecting the point of correspondence based on a single value rather than based on the distribution of values is susceptible to noise.

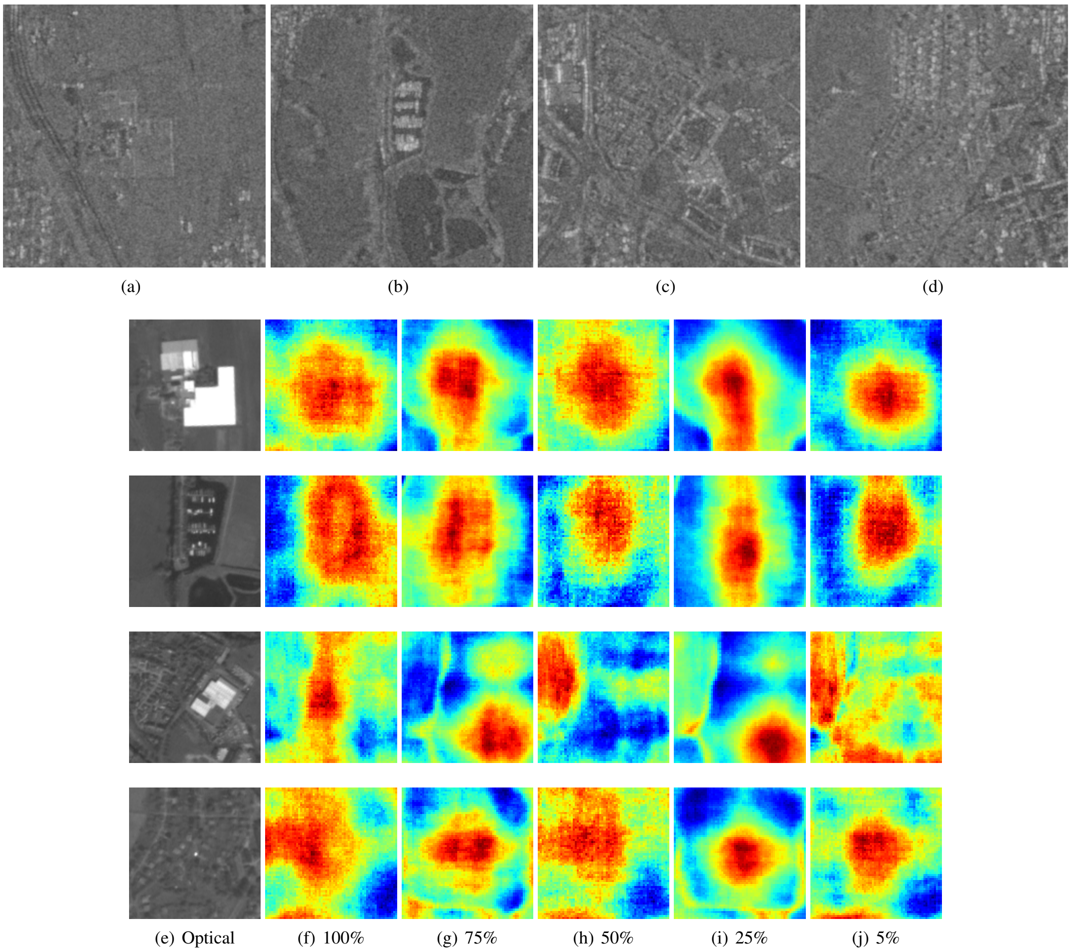
5.2 Effects of Data Scarcity

From the examples depicted in Figure 5 the impression arises that the proposed network performs best in semi-urban scenes (cf. Figure 5b), while it also shows reasonable performance in rural scenes (cf. Figure 5a). In urban scenes (Figure 5c-d), however, the matching accuracy varies significantly at different levels of supervision with the corresponding point shifting to a variety of locations. The reason for the better performance in semi-urban environments is likely due to the well distributed nature of objects in the scene, which allows the network to observe enough diversity in a patch that the descriptor can accurately capture the inherent details. In rural scenes, more often than not, there are fewer visual features and the scene has a relatively high self-similarity index, and thus the descriptors at multiple locations are similar. In urban scenes, the dense spacing of buildings, and thus the increased layover effects coupled with the $2.5m$ resolution obfuscate features and degrade the lower level structure of the scene, thus creating regions which have similar visual appearance, and in turn similar descriptors and multiple peaks in the correspondence map.

From Figure 4 the effects of data scarcity are visible in the overall distribution of the matching errors. As the amount of supervision is decreased the histogram becomes more skewed towards the right, and the number of successful matches for lower threshold values decreases significantly. This can be clearly observed when comparing the histograms of the fully supervised baseline (cf. Figure 4a) network to that of the scenario where only 5% supervision (cf. Figure 4d) was employed, where the former has a tighter distribution with a lower mean matching error, while the latter has a long tail and a very right-skewed distribution.

From further evaluation of Figure 4 it is clear that there isn't a linear trend between the number of accurately matched pairs and the amount of supervision used during training. This is evident in the accumulation of the number of matches which fall in the 1-percentile. Through this observation it is clear that 75% supervision and 25% supervision both have a higher number of low-error matches than the baseline approach.

At first glance this outcome can seem counter intuitive, however, an analysis of the literature (Dai et al., 2017) leads to the hypothesis that the unsupervised reconstruction loss and supervised matching loss are orthogonal to some degree. Thus, by optimizing for both losses in the baseline method the network ends up in a local minimum which is not necessarily best suited to either task. The reduction in the amount of supervision in the net-


(a) (b) (c) (d)

(e) Optical (f) 100% (g) 75% (h) 50% (i) 25% (j) 5%

Figure 5: Correspondence maps produced under varying conditions of data scarcity, on example scenes of differing density. (a-d) exemplary SAR test scenes, corresponding rows depicting (e) optical image patch, and (f - j) correspondence maps when trained with supervision percentage of 100%, 75%, 50%, 25% and 5% respectively.

work can be likened to applying some weighting function to the loss functions, and thus prioritizing the one objective over the other. In doing so the network is able to find a better optimum for the latent space generation task (reconstruction and adversarial losses) and the alignment of these spaces becomes an auxiliary task. While we would prefer to improve matching over reconstruction, it appears from the results that the prioritization of the adversarial task (by decreasing the supervision level) does in turn improve the matching task in some situations. This, however, would need to be subject to further investigation to fully understand the dynamics at play.

6. SUMMARY AND OUTLOOK

In this work, we proposed a semi-supervised approach to learn modality-specific features which are matchable via a simple distance-based metric, in our case cosine similarity. The approach consists of modality-specific autoencoders, which learn feature representations from unlabeled data, and are trained in an adversarial manner to enforce smoothness on the latent space. These learned representations (descriptors) are then aligned, using a supervised matching loss such that matching can be performed.

We further evaluated the effects of data scarcity on learning meaningful feature descriptors for SAR-optical matching by training our proposed network at varying levels of supervision and analysing the matching results in the form of correspondence maps, as well as the precision achieved for matching on our test set.

Overall we showed that even under very low data conditions, i.e. only 5% of supervision, we were able to obtain accurate correspondences in rural and semi-urban areas. While the overall number of accurate (1-percentile) correspondences was low, the strong structure of their correspondence maps leads us to believe that they could be filtered out during a post-processing step. This will be subject to further investigation in future work.

Furthermore, we found that the unsupervised and supervised objective functions are not fully complementary. That leads to a stunted baseline approach due to the strong trade-offs in the feature space required for each task. However, it was found that decreasing the amount of supervision can be sufficient to enable the network to learn a better latent distribution, and thus achieve higher accuracy in matching. This paper provides an initial contribution to the use of semi-supervised learning to exploit unlabelled training data in order to support SAR-optical matching, where training data is usually scarce and difficult to obtain. In future work we will investigate post-processing methods for extracting high accuracy correspondences based on the structure of their correspondence maps. We will further investigate the hypothesis that lowering supervision signals is equivalent to applying a weighting between the loss functions, and then will investigate ways of automatically learning an inverse weighting to reprioritize the matching/alignment task objective over the unsupervised objectives.

ACKNOWLEDGEMENTS

This work is supported by the German Research Foundation (DFG) as grant SCHM 3322/1-1

REFERENCES

Bui, T., Ribeiro, L., Ponti, M., Collomosse, J., 2018. Sketching out the details: Sketch-based image retrieval using convolutional neural networks with multi-stage regression. *Computers & Graphics*, 71, 77–87.

Chapelle, O., Scholkopf, B., Zien, A., 2009. Semi-supervised learning. *IEEE Trans. Neural Netw.*, 20(3), 542–542.

Cholaquidis, A., Fraimand, R. and Sued, M., 2018. Semi-supervised learning: When and why it works. *arXiv preprint arXiv:1805.09180.*

Chopra, S., Hadsell, R., LeCun, Y., 2005. Learning a similarity metric discriminatively, with application to face verification. In: *Proc. CVPR*, 539–546.

Dai, Z., Yang, Z., Yang, F., Cohen, W.W., Salakhutdinov, R. R., 2017. Good semi-supervised learning that requires a bad GAN. In: *Proc. NeurIPS*, 6510–6520.

Dellinger, F., Delon, J., Gousseau, Y., Michel, J., Tupin, F., 2015. SAR-SIFT: a SIFT-like algorithm for SAR images. *IEEE Trans. Geosci. Remote Sens.*, 53(1), 453–466.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A. C., 2017. Improved training of Wasserstein GANs. In: *Proc. NeurIPS*, Long Beach, USA, 5769–5779.

Han, X., Leung, T., Jia, Y., Sukthankar, R., Berg, A.C., 2015. MatchNet: Unifying feature and metric learning for patch-based matching. In: *Proc. CVPR*, 3279–3286.

Hughes, L.H., Schmitt, M., Zhu, X.X., 2018a. Mining hard negative samples for SAR-optical image matching using generative adversarial networks. *Remote Sensing*.

Hughes, L.H., Schmitt, M., Mou, L., Wang, Y., Zhu, X.X., 2018b. Identifying corresponding patches in SAR and optical images with a pseudo-siamese CNN. *IEEE Geosci. Remote Sens. Lett.*, 15(5), 784–788.

Iglovikov, V., Shvets, A., 2018. Ternausnet: U-net with VGG11 encoder pre-trained on imagenet for image segmentation. *CoRR*.

Lai, W.-S., Huang, J.-B., Yang, M.-H., 2017. Semi-supervised learning for optical flow with generative adversarial networks. In: *Proc. NeurIPS*, 354–364.

Liu, W., Shen, X., Wang, C., Zhang, Z., Wen, C., Li, J., 2018. H-net: Neural network for cross-domain image patch matching. In: *Proc. IJCAI*, 856–863.

Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., 2016. Adversarial autoencoders. In: *Proc. ICLR*.

Merkle, N., Auer, S., Müller, R., Reinartz, P., 2018. Exploring the potential of conditional adversarial networks for optical and SAR image matching. *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, 11(6), 1811–1820.

Merkle, N., Fischer, P., Auer, S., Müller, R., 2017a. On the possibility of conditional adversarial networks for multi-sensor image matching. In: *Proc. IGARSS*, Fort Worth, USA, 2633–2636.

Merkle, N., Luo, W., Auer, S., Müller, R., Urtasun, R., 2017b. Exploiting deep matching and SAR data for the geo-localization accuracy improvement of optical satellite images. *Remote Sensing*, 9(6), 586.

Mou, L., Schmitt, M., Wang, Y., Zhu, X., 2017. A CNN for the identification of corresponding patches in SAR and optical imagery of urban scenes. In: *Proc. JURSE*, Dubai, U.A.E.

Mukherjee, T., Yamada, M., Hospedales, T.M., 2017. Deep matching autoencoders. *CoRR*.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y., 2011. Multimodal deep learning. In: *Proc. ICML*, 689–696.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A., 2017. Automatic differentiation in PyTorch. In: *Proc. NeurIPS*.

Petzka, H., Fischer, A., Lukovnicov, D., 2017. On the regularization of Wasserstein GANs. *arXiv preprint arXiv:1709.08894.*

Rasmus, A., Valpola, H., Honkala, M., Berglund, M., Raiko, T., 2015. Semi-supervised learning with ladder network. *CoRR*.

Schmitt, M., Zhu, X.X., 2016. Data fusion and remote sensing – an ever-growing relationship. *IEEE Geosci. Remote Sens. Mag.*, 4, 6–23.

Schmitt, M., Tupin, F., Zhu, X., 2017. Fusion of SAR and optical remote sensing data - challenges and recent trends. In: *Proc. IGARSS*, Fort Worth, TX, USA, 5458–5461.

Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. In: *Proc. ICLR*.

Smith, L.N., Topin, N., 2017. Super-convergence: Very fast training of residual networks using large learning rates. *CoRR*.

Ye, Y., Shen, L., 2016. HOPC: A novel similarity metric based on geometric structural properties for multi-modal remote sensing image matching. *ISPRS Annals*, 3, 9.

Ye, Y., Shan, J., Bruzzone, L., Shen, L., 2017. Robust registration of multimodal remote sensing images based on structural similarity. *IEEE Trans. Geosci. Remote Sens.*, 55(5), 2941–2958.

Zagoruyko, S., Komodakis, N., 2017. Deep compare: A study on using convolutional neural networks to compare image patches. *CVIU* 164, 38–55.

Zhang, Y., Lee, K., Lee, H., 2016. Augmenting supervised neural networks with unsupervised objectives for large-scale image classification. In: *Proc. ICML*, 612–621.

Zhu, X.X., Tuia, D., Mou, L., Xia, G.S., Zhang, L., Xu, F., Fraundorfer, F., 2017. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geosci. Remote Sens. Mag.*, 5(4), 8–36.

4.6 A Deep Learning Framework for Sparse Matching of SAR and Optical Imagery

A Deep Learning Framework for Sparse Matching of SAR and Optical Imagery

Lloyd Haydn Hughes[a], Diego Marcos[b], Sylvain Lobry[b], Devis Tuia[b,*] and Michael Schmitt[a,*]

[a]Signal Processing in Earth Observation, Technical University of Munich (TUM), Arcisstr. 21, 80333 Munich, Germany
[b]Laboratory of Geo-Information Science and Remote Sensing, Wageningen University, The Netherlands

ARTICLE INFO

Keywords:
Multi-modal image matching
Image registration
Feature detection
Deep learning
Synthetic aperture radar (SAR)
Optical imagery

ABSTRACT

SAR and optical imagery provide highly complementary information about observed scenes. A combined use of these two modalities is thus desirable in many data fusion scenarios. However, any data fusion task requires measurements to be accurately aigned. While for both datas sources images are usually provided in a georeferenced manner, the geo-localization of optical images is often inaccurate due to propagation of angular measurement errors. Many methods for the matching of homologous image regions exist for both SAR and optical imagery, however, these methods are unsuitable for SAR-optical image matching due to significant geometric and radiometric differences between the two modalities. In this paper, we present a three-step framework for sparse image matching of SAR and optical imagery, whereby each step is encoded by a deep neural network. We first predict regions in each image which are deemed most suitable for matching. A correspondence heatmap is then generated through a multi-scale, feature-space cross-correlation operator. Finally, outliers are removed by classifying the correspondence surface as a positive or negative match. Our experiments show that the proposed approach provides a substantial improvement over previous methods for SAR-optical image matching and can be used to register even large-scale scenes. This opens up the possibility of using both types of data jointly, for example for the improvement of the geo-localization of optical satellite imagery or multi-sensor stereogrammetry.

1. Introduction

Two of the most used modalities for space-borne remote sensing are Synthetic Aperture Radar (SAR) and optical imagery, since the information they provide about observed scenes is highly complementary. Thus SAR-optical data fusion has become a relevant area of research within the field of remote sensing (Schmitt et al., 2017).

As with any data fusion task, a fundamental first step is the alignment of the various data sources. In the case of image-based data fusion this alignment usually takes place through the process of image matching. More specifically this relates to the determination of corresponding points or regions across images which have different viewpoints, resolutions and may have been acquired by different sensors.

In classical computer vision, where problems are often restricted to a single modality or sensor, the task of image matching is largely considered to be solved to the degree of being usable. However, this is not true when dealing with highly heterogeneous datasets and multiple modalities such as in the case of SAR-optical image matching. Although remote sensing imagery often contains geographical coordinates for each pixel, we cannot rely on this geocoding to provide accurate correspondences as optical imagery often contains significant geolocalization errors (Merkle et al., 2017; Müller et al., 2012). Thus we need to rely on an image matching process which is subject to many complexities related to the large geometric and radiometric differences between the SAR and optical modalities (Schmitt et al., 2017; Hughes et al., 2019). For instance, the geometric distortions present in SAR imagery, such as layover, foreshortening and radar shadow, have no direct analogues in the optical domain. Optical imagery, on the other hand, suffers from illumination effects, related to clouds, object shadows, and the global scene illumination.

To tackle these challenges, researchers took inspiration from classical computer vision and developed a number of approaches for SAR-optical matching. Suri and Reinartz (2010) used mutual information to create a histogram-based method of registering SAR and optical imagery. Later a multitude of hand-crafted approaches were developed which were aimed at improving the performance of the scale-invariant feature transform (SIFT) detection and description algorithm (Lowe, 2004), by adapting the gradient operator and scale-space to be more suited to the properties of SAR imagery (Dellinger et al., 2015; Gong et al., 2014; Suri et al., 2010). These approaches were relatively successful in matching images in the SAR domain, however, they failed to match across modalities as the detected and described features were independent of those features detected in the optical domain (Ma et al., 2017). This is partially due to the vast radiometric differences between SAR and optical imagery. To address this, (Ye and Shen, 2016) proposed the histogram of oriented phase congruency (HOPC) descriptor whereby phase congruency was used as a proxy for gradient information. This ensures a commonality between features and descriptors in both modalities. Xiang et al. (2018) argued for the use of modality-specific gradient operators with a Harris scale-space to better handle the large radiometric differences while still allowing for repeatable features to be detected

[*]Corresponding authors

✉ devis.tuia@wur.nl (D. Tuia); m.schmitt@tum.de (M. Schmitt)
ORCID(s): 0000-0003-0293-4491 (L.H. Hughes); 0000-0001-5607-4445 (D. Marcos); 0000-0003-4738-2416 (S. Lobry); 0000-0003-0374-2459 (D. Tuia); 0000-0002-0575-2362 (M. Schmitt)

across modalities. Li et al. (2020) combined these previous approaches and the use of phase congruency to create Radiation-variation Insensitive Feature Transform (RIFT), which was shown to be less sensitive to rotational and radiometric differences across modalities while still providing repeatable features.

While feature based methods are able to find correspondences between SAR and optical modalities, their success is limited to imagery which obeys specific geometric and radiometric constraints. These constraints often include the limitation to flat, sub-urban or rural environments where the SAR and optical geometry is similar and the radiometric properties are more strongly correlated (Li et al., 2020; Xiang et al., 2018; Ye and Shen, 2016).

At a higher level, the constraints on geometric and radiometric differences are a consequence of the hand-crafted nature of the feature detectors and descriptors and thus also exist in single domain matching problems. For instance, in classical computer vision many handcrafted approaches break down under large baselines, or strong radiometric differences. For this reason, and with the advent of modern deep learning techniques, there has been a strong movement towards deep matching – or learning to solve the image matching problem directly from data (Kuppala et al., 2020).

Fischer et al. (2014) demonstrated that features extracted from the last layer of a Convolutional Neural Network (CNN), pretrained on ImageNet, can outperform the SIFT descriptor in image matching tasks. This lead to the development of a number of CNN-based descriptors, which learned similarity metrics directly from corresponding image patch pairs. Simo-Serra et al. (2015) proposed the use of a siamese network trained with pairs of corresponding and non-corresponding patches, and a Euclidean distance metric to learn a 128-dimensional descriptor for image matching. A similar approach was proposed in (Zagoruyko and Komodakis, 2015), however, an additional network was added to focus the matching around the center of the image patch pair. Building on these approaches Han et al. (2015) proposed MatchNet, which made use of a triplet loss and hard negative mining to better discriminate between corresponding and non-corresponding patch pairs. In (Balntas et al., 2016a) and (Balntas et al., 2016b), a triplet-based approach was proposed, which used a simple shallow network and thus lead to a drastic improvement in computational and training performance without sacrificing accuracy. Taking a different approach Yi et al. (2016) proposed a learned variant of SIFT, in which each component of the SIFT matching pipeline was implemented as an independent CNN trained using SIFT as the ground truth.

Driven by these successes remote sensing practitioners turned to deep learning to address the various shortfalls of handcrafted approaches for matching SAR and optical imagery (Hughes et al., 2019). To this end a number of approaches have been developed which specifically account for the multi-modal and inherently heterogenous nature of the imagery. The first notable examples of deep SAR-optical matching made use of (pseudo-)siamese networks: Merkle et al. (2017) proposed a siamese network to directly predict the relative shift between a larger SAR search patch and a smaller optical template patch. Similarly, Mou et al. (2017) framed the matching as a binary classification problem and trained a pseudo-siamese network to predict the correspondence of the center pixel between SAR and optical patches. Taking inspiration from these initial works we extended the pseudo-siamese network proposed in (Mou et al., 2017) to include a more robust fusion network and modified the binary classification problem to output a similarity index based on a soft-max activation (Hughes et al., 2018). Citak and Bilgin (2019) proposed the use of SAR and optical visual saliency maps as an attention mechanism in the feature extraction arms of a siamese matching network. Wang et al. (2018) use a self-learned deep neural network to directly learn the mapping between a source and reference image with the goal of applying this mapping remote sensing image registration. Bürgmann et al. (2019) proposed modifications to HardNet (Mishchuk et al., 2017) and applied it to matching SAR Ground Control Points (GCPs) in optical imagery. Hoffmann et al. (2019) trained a Fully Convolutional Network (FCN) to learn a similarity metric which was invariant to small affine transformations between SAR and optical patch pairs. Ma et al. (2019) proposed a two-step, coarse-to-fine registration method based on features extracted from fine-tuned VGG16 model (Simonyan and Zisserman, 2015).

Although we have seen significant progress in the matching of SAR and optical imagery, these approaches rely on the selection of good feature points for the extraction of matchable candidate search and template patches. Given the large differences between SAR and optical imagery it is often the case that salient features are not visible in both domains. Thus the selection of candidate patches in previous works has largely relied on features extracted from a single modality (Bürgmann et al., 2019; Merkle et al., 2017; Hughes and Schmitt, 2019) or assumed correspondence based on geolocalization (Citak and Bilgin, 2019; Hoffmann et al., 2019; Ma et al., 2019). For instance in (Merkle et al., 2017) the locations of road intersections extracted from OpenStreetMap (OSM) data were used as features for extracting candidate regions for matching. While this showed reasonable results, OSM data is known to have varying accuracy and is not globally consistent (Vargas-Muñoz et al., 2019). Furthermore, the approach also required significant preprocessing and manual intervention. Bürgmann et al. (2019) made use of GCPs derived from a geodetic stereo SAR approach as features for the extraction template patches from the SAR image. The generation of these GCPs is computationally complex and requires multiple SAR acquisitions of the same scene with specific acquisition geometry. Furthermore, these GCPs are not generic features and often do not exist in rural areas.

Even in the best case scenario, where the proposed candidate patches meet all the requirements for increasing the likelihood of matching, outliers and incorrect matches will still exist. This is both due to the ambiguity and the complexity of the task of matching under extreme heterogeneity. The

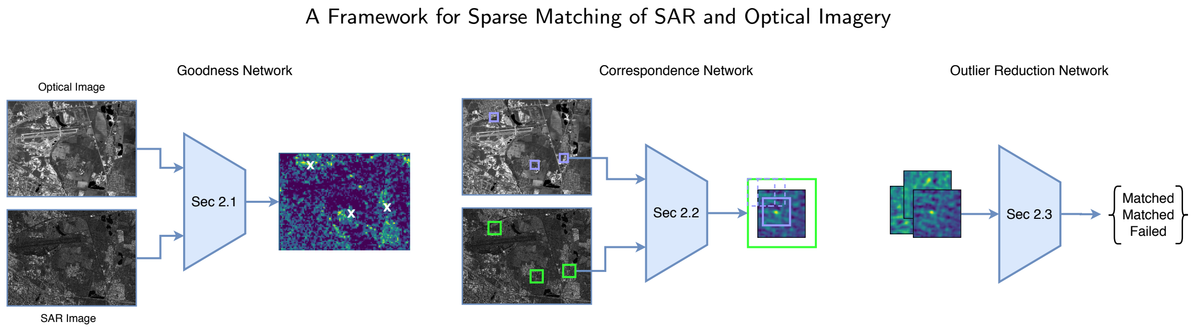


Figure 1: The proposed SAR-optical matching framework. First the SAR and optical images of the scene are processed by the *goodness network* to create a scene-wise map of suitability of regions for matching. Candidate search (green boxes) and template patches (blue boxes) are then extracted from the local maxima and the *correspondence network* is used to determine the point of correspondence via feature-space cross correlation, then producing a correspondence heatmap. The quality of the match is then assessed by the *outlier reduction network* to filter out incorrect or ambiguous correspondences.

task of identifying and removing outliers in classical computer vision usually falls on statistical approaches such as the RANdom Sampling and Consensus (RANSAC) algorithm (Fischler and Bolles, 1981). These approaches, however, have not seen use in SAR-optical matching due to the complexity of modelling the feature transfer between domains in the presence of large geometric differences. Therefore, the removal of outliers in SAR-optical matching approaches has largely relied on filtering matches based purely on the similarity score. Thus many of the previously mentioned approaches suffer from high false positive rates, which degrade the performance of downstream tasks.

In this paper we propose a fully-automated, multi-scale SAR-optical matching framework to address some of the shortfalls and constraints of previous approaches. This framework is comprised of three neural networks used in sequence: first is a *goodness network*, made of domain-specific subnetworks. This first network highlights regions with a high likelihood of containing salient features which are matchable across modalities. Second is a *multi-scale matching network*, architected around a feature space correlation function, which produces correspondence heatmaps for the matching of candidate patches. Finally, an *outlier reduction network* is used to directly estimate the quality of the matching result and allow for the removal of incorrect matching results. We evaluate the effectiveness of the individual subcomponents, as well as the complete SAR-optical matching pipeline on a large and diverse dataset of high resolution SAR and optical imagery.

2. Multi-modal Feature Proposal and Matching Framework

In this section we detail the architecture and design of the three components which make up the proposed end-to-end SAR-optical matching framework. An overview of the framework and definition of these main components is depicted in Figure 1.

2.1. Goodness Network

The first stage of our framework aims at extracting the candidate patches which are used, by the correspondence network, for matching SAR and optical imagery. To extract these patches, we assess the *goodness* of regions for matching, *i.e.* the suitability of a region for matching.

This assessment is made using two independent domain-specific CNNs, each one producing a map indicating the likelihood of a region being matchable. With each network being trained on a single modality, but supervised by the matching loss generated by the correspondence network (see Section 3.2 for details), we expect the domain specific CNNs to learn which features are likely to be discernible in the other modality. These two maps are then merged into a *cross-modality scene goodness map*.

To cope with the geo-coding errors which exist in optical remote sensing imagery and the large differences in geometry between SAR and optical imagery, we generate the *goodness maps* at a reduced resolution. This allows for the coarse alignment of the SAR and optical goodness maps, and thus the extraction of jointly good regions, i.e. regions which have a high goodness in both domains. While this alone solves the correspondence problem, it only does so at a significantly reduced resolution and thus the identified regions are used to extract candidate patches for higher resolution matching with the correspondence network presented in the next section. Furthermore, in identifying regions for matching in this manner we reduce the overall number of candidate points. However, many downstream applications of the determined correspondences only require a few, well distributed and accurately matched feature points.

The modality specific networks are based on the VGG11 architecture (Simonyan and Zisserman, 2015) and are described in Table 1. This base was chosen due to its simplicity, relatively low number of parameters and proven performance in a variety of tasks (Ma et al., 2019; Iglovikov and Shvets, 2018; Hughes and Schmitt, 2019). The backbone architecture consists of four blocks of two 3 × 3 convolutional layers, with each convolutional layer being a sequence of convolution, activation by a rectified linear unit (ReLU)

Table 1
Overview of the domain specific goodness networks. Conv(k, s, p) and MaxPool(k, s) represent a convolutional layer and a pooling layer, with a kernel of size k, a stride of s, and a padding of p, respectively

Block	Layer	# Filters
1	Conv(3,1,1) -> ReLU -> BN	32
	Conv(3,1,1) -> ReLU -> BN	32
	MaxPool(2,2)	
2	Conv(3,1,1) -> ReLU -> BN	64
	Conv(3,1,1) -> ReLU -> BN	64
	MaxPool(2,2)	
3	Conv(3,1,1) -> ReLU -> BN	128
	Conv(3,1,1) -> ReLU -> BN	128
	MaxPool(2,2)	
4	Conv(3,1,1) -> ReLU -> BN	128
	Conv(3,1,1) -> ReLU -> BN	128
Head	Conv(3,2,1) -> ReLU -> BN	128
	Conv(3,2,1) -> ReLU -> BN	64
	Conv(1,1,0) -> ReLU	64
	Conv(1,1,0) -> ReLU	1
	AvgPool(N_p,N_k) -> Sigmoid	

and batch normalization (BN). The first three convolutional blocks are downsampled by a factor of 2 using max-pooling. The head of the network consists of two convolutional layers with a stride of 2 (thus downsampling the spatial dimension of the tensor by a factor of 2), followed by fully connected layers implemented using a 1×1 convolutional block. Thus creating a network with an effective downsampling scale of 32, which is slightly larger than the maximum expected offset as reported by Merkle et al. (2017). Finally, an average pooling layer, with a kernel size N_p and stride N_k ensures a receptive field that accounts the maximum expected offset between the domains, as well as the size of the desired template patch.

The cross domain goodness networks are trained using co-registered SAR-optical patch pairs, $\mathbf{I}_s, \mathbf{I}_o$, as well as a binary label, y_m, which represents if the pair is matchable.

The networks are trained using a Binary Cross Entropy (BCE) loss function:

$$\mathcal{L}_g = -\frac{1}{N}\sum_{i}^{N} y_i \log(\tilde{y}_i) + (1 - y_i)\log(1 - \tilde{y}_i), \qquad (1)$$

where y_i is a binary label indicating if the pair can be matched, $\tilde{y}_i$ is the predicted label, and N is the total number of samples.

To produce the cross-modality scene goodness map $\mathbf{G}$, we select either the minimum or maximum goodness response for each pixel across the modalities and then apply the spatial non-local-maximum suppression proposed in (Dusmanu et al., 2019):

$$\hat{G}_{ij} = \frac{\exp(G_{ij})}{\sum_{kl \in \mathcal{N}_{ij}} \exp(G_{kl})}, \qquad (2)$$

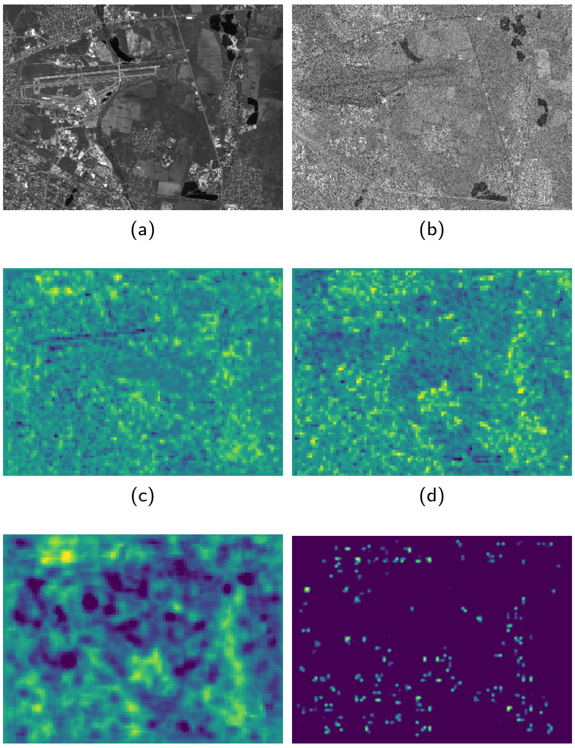


Figure 2: Example of scene goodness maps produced by the domain specific networks, and the final, fused goodness map. (a) and (b) are the optical and SAR images of the scene. (c) and (d) are the respective domain-specific goodness maps. (e) is the minimal response cross-modality goodness map $\mathbf{G}$ and (f) the final cross-modality scene goodness map $\hat{\mathbf{G}}$, where points of high goodness are clearly visible.

where G_{ij} is the value of $\mathbf{G}$ at pixel (i, j), and $\mathcal{N}_{ij}$ is a 3×3 window centered on (i,j). An example of the domain specific, and cross-modality scene goodness maps is depicted in Figure 2.

Finally, candidate patches are extracted around the points of high goodness by transforming these point locations into the original image space. This is done by undoing the poolings and strides to find the point in the original image space.

2.2. Correspondence Network

The goodness network informs about regions of the two images that seem to be interesting to find matching keypoints, but does that only at a coarse resolution. The next step is to find a fine grained matching keypoint between the two. To do so, a second correspondence network slides a small subpatch of the optical image (*template patch*, $\mathbf{I}_t$, of size $\mathcal{N}_t \times \mathcal{N}_t$) over the wider SAR image (*search patch*, $\mathbf{I}_s$, of size $\mathcal{N}_s \times \mathcal{N}_s$) in search of a match. In other words, the correspondence network aims to determine the most likely point of correspondence for the center pixel of the template patch within the search region. This can be seen as equivalent to

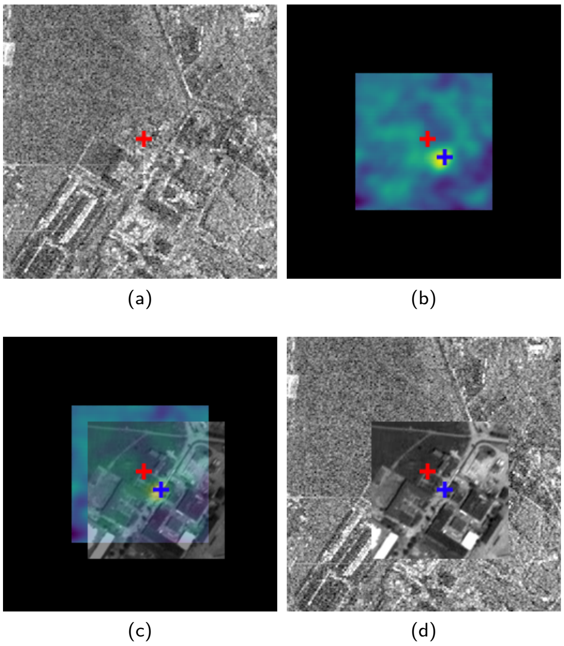


Figure 3: Example of the process by which the correspondence heatmap can be used to determine the corresponding point for the center pixel of the optical template patch. (a) The search window with its center pixel marked by a red plus, (b) the resultant heatmap from the correspondence network with its center pixel aligned to that of the search window, and the peak point of correspondence depicted by a blue plus. (c) The center of the optical template patch is aligned to the peak point of correspondence, (d) the final alignment of the optical template patch, with the located point of correspondence marked by the blue plus.

finding the offset which leads to the best overall alignment of the template within the search patch. An example of the matching process using a candidate patch pair, and the output correspondence map is depicted in Figure 3.

Existing approaches to SAR-optical matching largely rely on features extracted from the final layers of deep CNNs. While these features contain global semantic information they are low resolution and invariant to disturbances such as translation. Thus it can be argued that they lack the fine detailed features required to accurately determine correspondence between images. For this reason we architected our correspondence network around the concept of convolutional hypercolumns (Hariharan et al., 2015) which are constructed by stacking feature maps extracted from multiple levels of a shallow CNN.

The correspondence network consists of two four-layer CNNs, one for each modality, from which feature maps are extracted to form the modality specific hypercolumns. The number of channels in each hypercolumn is then reduced by a modality specific feature reduction network, before being matched using a feature space correlation operator.

The hypercolumn is constructed by extracting feature maps at each of the four layers of the feature extraction network. These feature maps are then upsampled, using bi-linear interpolation and stacked into a hypercolumn. The depth is then reduced to the desired number of features, $\mathcal{N}_d$, using a series of 1×1 convolutional layers. To improve response of salient features in each modality, a spatial attention map, as proposed in (Woo et al., 2018), is created and applied to each hypercolumn. The reduced hypercolumn is then normalized along the channel dimension using $L2$ normalization.

The search and template hypercolumns are then matched in feature space using a correlation operation and valid padding. Finally, the result is upsampled and padded to match the extent of the search window. The output of which is a heatmap containing the matching scores for each offset of the template window within the search window. The full architecture of the correspondence network, as well as the input and output datum, is depicted in Figure 4.

We can train the network using a 2D Kronecker delta function as the ground truth, whereby the position of the unit impulse is parameterized as the true point of correspondence of the template patch within the search patch. The network is then trained via backpropagation using a modified mean-squared error (MSE) loss,

$$\mathcal{L}_{mse} = \frac{1}{\mathcal{N}_1 + \mathcal{N}_0} \sum_i \mathbf{w}_i \left(\mathbf{y}_i - f_{ss}(\tilde{\mathbf{y}}_i)\right)^2, \tag{3}$$

$$\mathbf{w}_i = \mathbf{y}_i \frac{\mathcal{N}_0}{\mathcal{N}_1} + (1 - \mathbf{y}_i), \tag{4}$$

where, $\mathbf{y}_i$ and $\tilde{\mathbf{y}}_i$ represent the target labels and the predicted heatmap of the i^{th} sample. The function f_{ss} is a spatial softmax operation which is applied to the predicted heatmap in order to convert the matching scores into a probability distribution with the peak at the point of correspondence. The softmax activation relates all points in the heatmap and thus to obtain a strong peak it encourages the suppression of the matching score in other regions. As the ground truth map contains only a single non-zero value we make use of a weighting vector, $\mathbf{w}_i$, to ensure the loss at the peak is given the same importance as the loss created by all non-corresponding points in the heatmap. This further exaggerating the requirement for a strong peak in the heatmap. Thus $\mathcal{N}_1$ and $\mathcal{N}_0$ represent the count of the number of zero- and non-zero pixels in $\mathbf{y}_i$.

Due to the spatial softmax operation f_{ss}, which normalizes $\sum_{x,y} \tilde{\mathbf{y}} = 1$, and the loss function which prioritizes peakiness, the network tends to overfit the training dataset. It achieves this by exploiting the peak-to-peak range of the pre-activated heatmaps, $\hat{\mathbf{y}}_i$. To reduce overfitting, encourage sparsity, and limit the dynamic range of $\hat{\mathbf{y}}_i$, we augment our $\mathcal{L}_{mse}$ loss with an L_1 regularization term. Thus the overall loss function can be expressed as

$$\mathcal{L}_{cor} = \mathcal{L}_{mse} + \lambda \sum_i |\hat{\mathbf{y}}_i|, \tag{5}$$

where λ is a hyperparameter to adjust the strength of the regularization.

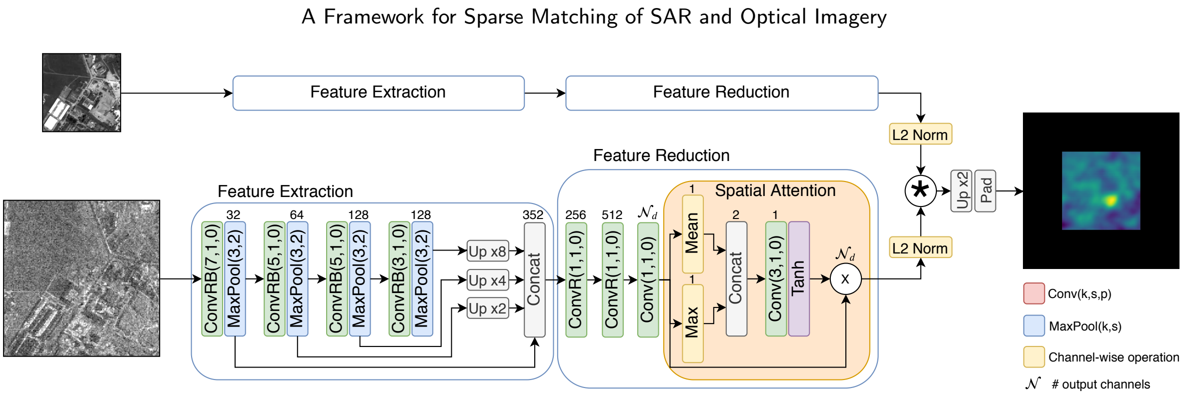


Figure 4: The correspondence network architecture showing the layer details for the SAR branch with Conv(k, s, p) and MaxPool(k, s), representing a convolutional layer, and pooling layer, with a kernel of size k, stride of s, and padding of p, respectively. Convolution followed by ReLU is represented as ConvR(k, s, p), and the addition of batch normalization as ConvRB(k, s, p).

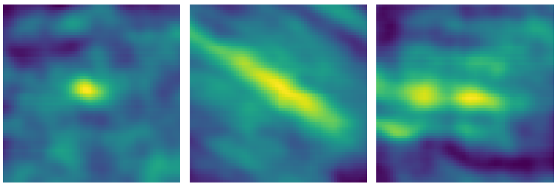

(a) (b) (c)

Figure 5: Examples of common patterns seen in the correspondence heatmaps. For brevity only the *valid* region of the heatmap is depicted. (a) High likelihood of an accurate match as the heatmap contains only a single, strong response with a low spread. (b) A matching ambiguity exists along a single axis, which leads to a lower likelihood of the correct point of correspondence being identified. (c) A strongly multimodal response, with a wide spread which leads to multiple ambiguities in matching and thus a lower confidence.

2.3. Outlier Reduction Network

Due to the nature of the spatial softmax operation, which is applied to the correspondence heatmaps, $\tilde{\mathbf{y}}$ will likely always contain a small cluster of pixels which exhibits a strong response. However, the magnitude and location of these pixels is insufficient to discern the quality of the matching result. Thus we hypothesise that a better approach in determining the matching quality is to analyze the topology of the pre-softmax heatmap, $\hat{\mathbf{y}}$. We base this hypothesis on the observation that good matches tend to exhibit a single narrow peak, while incorrect matches are often multi-modal, or have a wide spread. Examples of various correspondence heatmaps are presented in Figure 5.

To this end we train an Outlier Reduction Network (ORN) on $\hat{\mathbf{y}}$ to classify good and bad matches. The overall goal of the ORN is to reduce the overall number of inaccurate matches of the correspondence network, as a low false positive rate is more important than a high number of matches for many downstream applications of SAR-optical matching such as, for example, image coregistration or stereogrammetry (Müller et al., 2012; Merkle et al., 2017; Qiu et al., 2018; Bagheri et al., 2018).

Table 2
An overview of the layers of the ORN architecture.

Layer	# Filters
Conv(7,1,0) -> ReLU -> IN	32
MaxPool(3,2)	
Conv(5,1,0) -> ReLU -> BN	64
MaxPool(3,2)	
Conv(5,1,0) -> ReLU -> BN	128
MaxPool(3,2)	
Conv(3,1,0) -> ReLU -> BN	128
Conv(1,1,0) -> ReLU	128
Conv(1,1,0)	1
AdaptAvgPool	
Sigmoid	

The ORN is based on the same architecture as the correspondence feature extraction network, with some minor modifications. As the heatmaps produced by the correspondence network are not normalized and have a variable dynamic range, they cannot be assumed to have been drawn from the same distribution. Thus we adapt the input layer to use instance normalization (IN), instead of BN, as it operates on each sample independently. We formulate the problem of determining outliers as binary classification, and thus we need to adapt the head of the network to be suitable for this task. This modification includes the addition of an adaptive average pooling layer (AdaptAvgPool), which pools the entire spatial extent to output a single value, and a sigmoid activation to allow for training using a BCE loss.

Training is then supervised using ground truth labels which are derived based on the accuracy of the matching result as reported by the correspondence network, this process is described in detail in Section 3.3. The problem can be summarized as: given a correspondence heatmap $\hat{\mathbf{y}}$, is it more likely to represent a successful or unsuccessful match. The full architecture is described in Table 2.

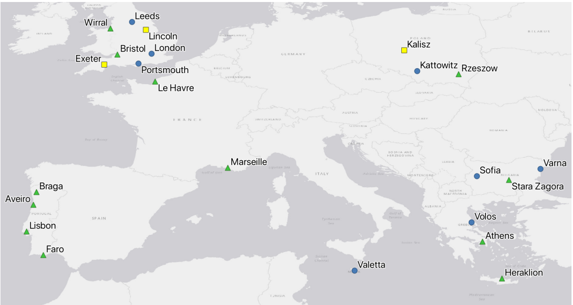
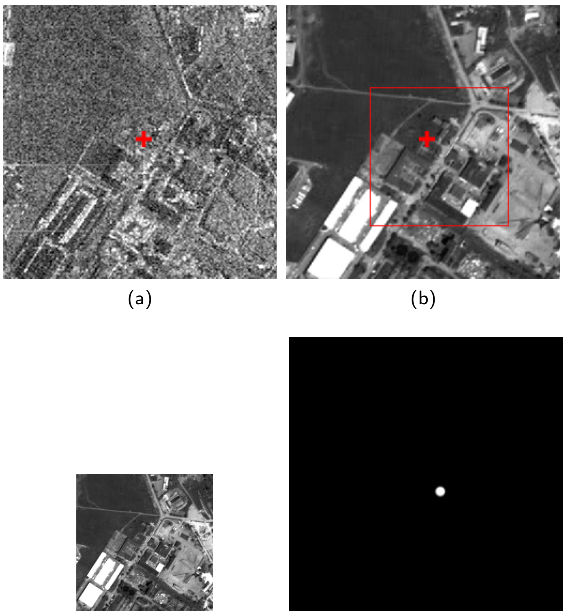


Figure 6: The distribution of cities in the Urban Atlas dataset. The cities used for training, validation and testing are depicted as green triangles, yellow squares and blue circles respectively.



(a) (b) (c) (d)

Figure 7: A single training sample from our correspondence dataset. (a) The SAR search patch cropped around the location of the optical Harris corner (represented by the red cross), (b) the optical patch from which we extract the template search patch with random offset during training (depicted by the red box), (c) The extracted template patch, and (d) the derived ground truth label representing the true point of correspondence.

3. Datasets and Workflow

While the logical structure of the framework follows from the goodness network via the correspondence network to the outlier reduction network, as depicted in Figure 1, this is not the case for training. As the training of goodness and outlier reduction networks rely on a trained correspondence network, we start by describing the dataset for the correspondence network followed by the description of the datasets, derived from the correspondence network outputs, which are used for training the goodness and outlier reduction networks. We further provide insights into the assumptions which were made and outline the way in which training, validation and testing samples were selected.

3.1. SAR and Optical Correspondence

To train the correspondence network we require a large dataset of salient candidate search and template patches with known points of correspondence. Due to the complexity of creating such a dataset, and the intractability of manually annotating correspondence across heterogenous domains, we rely on simplifying assumptions (such as the correspondence of points at ground level in co-registered imagery) and the Urban Atlas dataset (Schneider et al., 2010) to generate our training and validation data.

The Urban Atlas dataset consists of manually co-registered, high resolution TerraSAR-X and PRISM imagery over 23 European cities and their surrounding areas. The imagery has a ground sampling distance (GSD) of $1.5m$ and $2.5m$ respectively, and has been manually co-registered such that there is accurate correspondence for pixels at ground level. We downsample the TerraSAR-X imagery to match the GSD of the PRISM imagery in order to reduce the complexity of the problem. The 23 cities are then divided into three groups for training, validation and testing. This division and the distribution of the cities can be seen in Figure 6.

We then apply a Harris corner detector to the optical images to select points which are salient in at least one modality. Using these points, and the knowledge that the SAR and optical data in the Urban Atlas data set has been accurately co-registered, we select the corresponding points from the SAR imagery using the geo-reference information for each pixel. We then use OpenStreetMap data and non-maximal suppression to reduce the overall point set to contain points which are more likely to be at ground level, such as near roads, and away from buildings and forested areas. This step is performed as in the case of co-registered data, the assumption of correspondence, at the same geo-location, only holds for points with no height above the ground.

We then cut 256×256 pixel patches from the SAR and optical imagery, centred around the identified points of correspondence. Then during training we randomly crop a 128×128 template patch from the optical patch, with a maximum offset of 32 pixels around the center (accounting for the maximum shift (Merkle et al., 2017)). In doing so we ensure that the correspondence network learns to match the template image to the search window under realistic conditions, while allowing for the generation of ground truth data for the supervision and evaluation of the training process. An example of a candidate patch pair and the corresponding ground truth label is depicted in Figure 7.

The 128×128 pixel extent of the optical template patch was chosen such that it captures sufficient spatial context to enable matching under the assumed worst case scenario, while remaining small enough to allow for better selectivity and finer grained matching. The extent of the SAR search patch was then selected such that is allowed for a maximum matching offset of up to 32 pixels (Merkle et al., 2017), while ensuring that even under extreme cases there is sufficient spatial context for matching.

We then standardise the dynamic range of the SAR imagery, and convert the speckle into an approximate additive Gaussian noise model. This is done by converting the pixel values to Decibels (dB) and then clipping their range to the 3σ range of the training images, $I_{SAR} \in [10, 30]$dB. For the optical imagery we simply normalize the values to the range of $I_{opt} \in [0, 1]$ by dividing through by 255.

While the test scenes are processed in the same manner as the training and validation scenes, the candidate patches are only useful for the evaluation of the correspondence network. Thus to evaluate the entire pipeline in an end-to-end manner we also create larger test scenes which can be used for evaluation. The train, test and validation patches are extracted from spatially distinct regions with a maximum patch overlap of 50%, while the 8 larger test scenes are created from each testing city and are thus spatially diverse and contain no overlap. The final dataset consists of 40,314 training candidate patch pairs, 4,205 validations pairs and 6,353 testing pairs, as well as, 8 larger test scenes.

3.2. Goodness

As no goodness dataset exists, and the creation of such a dataset is non-obvious for manual annotation, we rely on the trained correspondence network to identify patches which can act as positive and negative samples for training and evaluating the goodness network.

To do this we make use of the SAR and optical patches from the correspondence datasets, as well as the recorded matching loss, $\mathcal{L}_{mse}$, and an L_2 correspondence point error for each sample $\mathcal{L}_e$. We then create binary goodness labels for each sample by thresholding $-\log(\mathcal{L}_{mse})$ and the L_2 error. The negative log loss is used to invert the loss and reduce the dynamic range, which makes the task of selecting thresholds easier. We label the patch pairs such that,

$$\mathbf{S}_i = \begin{cases} 1 \text{ if } & -\log(\mathcal{L}_{mse}) \geq 1.2 \text{ and } \mathcal{L}_e \leq 1 \\ 0 \text{ if } & -\log(\mathcal{L}_{mse}) \leq 1 \text{ or } \mathcal{L}_e \geq 2.5 \end{cases} \tag{6}$$

where $\mathbf{S}_i$ is the i^{th} patch pair. The values for the thresholds were chosen based on the training dataset such that we avoid possibly ambiguous samples, this process is depicted in Figure 8. The negative log loss allows for easier selection of patches which produce correspondence heatmaps with desirable properties (low matching loss), such as a single peak with a narrow spread and small values everywhere else, while the $L2$ threshold ensures that these heatmaps actually correspond to positive matches.

As the correspondence dataset was created with no guarantees of mutually visible features, there is a large imbalance in the final goodness dataset with many more negative examples being present. To correct this we reduce the number of negative samples, by random selection, to be equal to the number of positive samples.

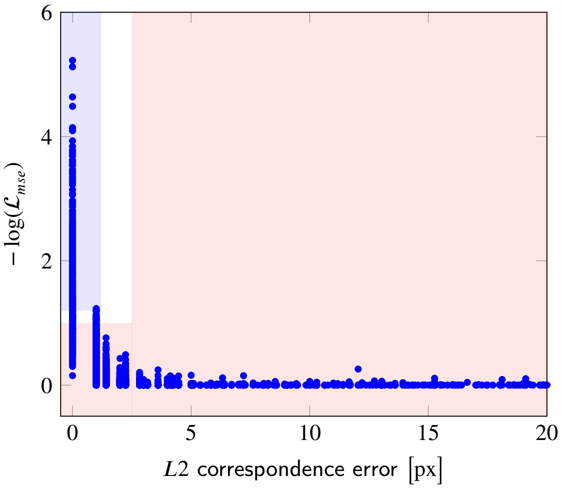


Figure 8: A plot of the negative log matching loss versus the $L2$ pixel error for the training dataset. The region from which positive samples are drawn is highlighted in blue, and the negative samples are drawn from the area in red.

The final step in creating the goodness dataset is to crop the SAR search patches to the same extent as the corresponding optical template patch. This is done as the goodness score is derived only from the maximum point of correspondence, thus regions beyond the extent of the template patch do not contribute to whether the patch was good for matching or not.

3.3. Outlier Reduction

To train the outlier removal network we make use of the *valid* region of the heatmaps generated from the correspondence network. These heatmaps are used as inputs to the outlier reduction network and the binary training labels indicate whether they were the result of a successful or unsuccessful matching result.

The generation of the heatmap labels follows the same approach to labelling as the previously described goodness dataset. However, we only apply the L_2 threshold as the label relies solely on whether the patch was accurately matched. Some labelled examples from the training dataset are shown in Figure 9.

4. Implementation Details

Due to the data requirements discussed in Section 3, we first train the correspondence network and then use the results of this training to generate the data needed to train the goodness and outlier reduction networks.

The average pooling parameters of the goodness network were set as $\mathcal{N}_p = 4$ and $\mathcal{N}_k = 1$. This corresponds to creating a receptive field of 128 × 128 pixels, which is large enough to account for co-registration errors of up to 160 meters between domains, while exhibiting a 75% overlap be-

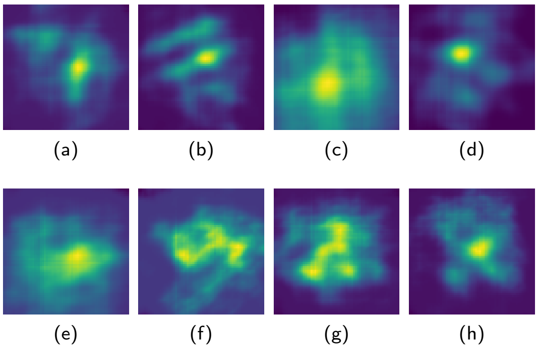


Figure 9: Examples of the positive (a-d) and negative (e-h) correspondence heatmaps used to train the outlier reduction network. Only the *valid* region of the heatmap is used as the padded area contains no additional information.

Table 3
The hyperparameters used for training each of the sub-networks, where lr is the learning rate, β_1 and β_2 control the momentum.

Network	lr	β_1	β_2	weight decay
SAR Goodness	5×10^{-5}	0.9	0.999	0
OPT Goodness	9×10^{-4}	0.9	0.999	0
Correspondence	1×10^{-4}	0.9	0.999	1×10^{-6}
Outlier Reduction	1×10^{-4}	0.9	0.999	1×10^{-6}

tween the evaluated regions. Furthermore, the hypercolumn depth $\mathcal{N}_d$ of the correspondence network was set to 256.

We make use of the PyTorch deep learning framework (Paszke et al., 2019) to implement all aspects of our proposed pipeline. The various sub-networks were randomly initialized using the method proposed by (He et al., 2015), and are trained using the Adam solver (Kingma and Ba, 2014). The hyperparameters used for the solver are specified in Table 3. For each of the sub-networks the optimal learning rate was determined using the search method proposed by (Smith, 2017).

We make use of a fixed batch size of 16 samples, which constitutes the maximum batch size that could be used to train the correspondence network on a Nvidia GTX1080Ti GPU. This batch size further allowed for both goodness networks and the outlier reduction network to be trained simultaneously on the same GPU. The correspondence network was trained for 50 epochs, while the remaining sub-networks were trained for 1000 epochs due to the relatively small dataset size in comparison to the correspondence dataset size.

Data augmentation was used to improve generalization and reduce the risks of overfitting. This step was found to be of increased importance when training the goodness and outlier reduction networks due to the reduced dataset size. The data augmentation pipeline consisted of horizontal (HF) and vertical flipping (VF), image scaling (IS) by a factor of ± 0.1, intensity scaling (CS) by a random value between $(0.7, 1.3)$, and coarse image dropout (CD) of between $(1\%, 5\%)$ of the image pixels, taken on a version of the image which is downsampled to between $(80\%, 98\%)$ of the original size. Each of these augmentations is applied is applied with a certain probability in. The probabilities used for augmentation during the training of each network can be found in Table 4.

Table 4
The probabilities used in the data augmentation pipeline while training each sub-network.

Network	HF	VF	IS	CS	CD
SAR Goodness	0.5	0.5	0	0.7	0.8
OPT Goodness	0.5	0.5	0	0.7	0.8
Correspondence	0.5	0.5	0	0	0
Outlier Reduction	0.5	0.5	0.1	0	0.8

Table 5
Correspondence network configurations used in the ablation study. The use of a specific layer or inclusion of regularization is indicated by a yes (Y) or no (N).

Network	Attention	Spatial Softmax	L1 Reg.
CorrBase	N	N	N
CorrA	Y	N	N
CorrAS	Y	Y	N
CorrASL	Y	Y	Y

To aid future development and in the interest of openness in science, a full implementation of the framework has been released[1].

5. Experiments and Results

In this section we first motivate our architectural choices by performing ablation studies. We further evaluate the performance of the individual sub-networks in comparison to existing methods, as well as their effects on the accuracy of the final set of correspondences. Finally, we evaluate the overall performance of the matching framework over a larger test scene.

5.1. Ablation Study

To aid the design of the correspondence network described in Section 2.2 we performed an ablation study to compare the performance of the network as various architectural and regularization elements were added. We tested four variants of the correspondence network which are detailed in Table 5.

The networks were trained as previously described, and the random elements in the training process were made deterministic such that all networks were trained on the same data and augmentations. Finally, we evaluated the performance of the various networks using the validation dataset to prevent biasing our architecture selection to the test data.

We evaluate performance in terms of matching accuracy and precision. Whereby, matching accuracy is defined by the percentage of matches which have an $L2$ distance to the

[1] https://github.com/system123/SOMatch

Table 6
Influence of attention (A), spatial-softmax activation (S) and $L1$ regularization (L) on the matching performance (evaluated on the validation dataset) of the correspondence network.

Network	Matching Accuracy		Matching Precision
	≤ 1px [%]	avg. $L2$ [px]	mAP [px]
CorrBase	28.44	2.34	1.27
CorrA	28.13	2.36	**1.25**
CorrAS	44.42	3.0	1.99
CorrASL	**54.46**	**2.32**	1.53

ground truth point of correspondence of at most one pixel, as well as the mean $L2$ error. Matching precision is defined as the mean average precision (mAP), where the standard deviation is used as a measure of precision. The results of the ablation study are described in Table 6.

From Table 6, it can be see that the addition of the spatial softmax operator leads to a significant improvement in terms of matching accuracy, however, this comes with a reduction in precision. The addition of the $L1$ regularization term further improves the matching accuracy while simultaneously only having slightly reduced precision over the baseline network with attention. Thus the *CorrASL* network was selected as the preferred architecture for our SAR-optical matching framework, and all further experiments are conducted with reference to this result.

5.2. Matching Results

As the correspondence network plays a vital role in training the goodness and outlier reduction networks, it is imperative that we evaluate its performance relative to existing methods. To do so we make use of two relevant and available methods: Normalized Cross Correlation (NCC) (Burger and Burge, 2009), as well as the pseudo-Siamese matching approach (PSiam) presented in (Hughes et al., 2018).

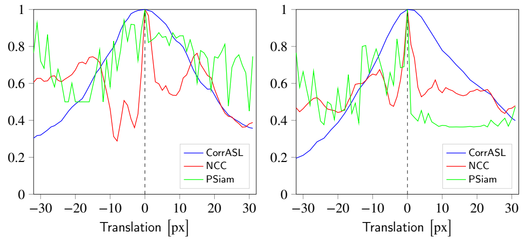
To ensure a fair comparison we retrained the pseudo-siamese approach on the same dataset, and under the same data augmentations and pre-processing as our correspondence network. As the pseudo-siamese network requires corresponding and non-corresponding SAR-optical patch pairs, for training, we applied random offsets for the creation of the non-corresponding pairs. Furthermore, both the SAR and optical pairs were cropped to an extent of 128 × 128 pixels. During the evaluation phase we apply the pseudo-siamese network over the full extent of the SAR search patch, using a sliding window approach, to generate a correspondence heatmap.

Table 7 shows the matching accuracy and precision for the baseline methods compared to the proposed method when assessed on our ground-level Harris corner derived test dataset.

From Table 7 it is clear that our proposed matching architecture provides a significant improvement in matching accuracy as well as precision over the selected baseline methods. The discrepancy between the test precision in and the validation precision reported in Table 6 is most likely due to a wider diversity of scenes being used for testing.

Table 7
A comparison of the matching accuracy and precision (evaluated on the testing dataset) of NCC (Burger and Burge, 2009), PSiam (Hughes et al., 2018) and our proposed correspondence network.

Network	Matching Accuracy		Matching Precision
	≤ 1px [%]	avg. $L2$ [px]	mAP [px]
NCC	8.2	7.85	6.81
PSiam	18.4	5.22	5.93
CorrASL	**46.9**	**2.1**	**2.62**



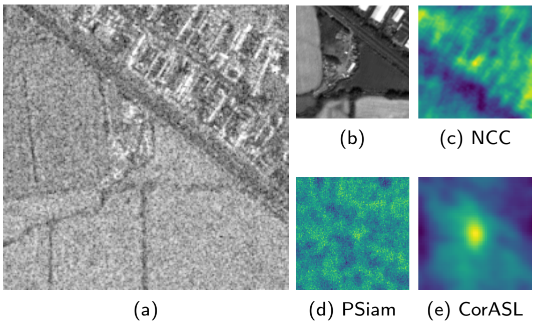
(a) (b)

Figure 10: The median correspondence heatmap peak shape along the (a) x-axis and (b) y-axis for each of the evaluated approaches.

In Figure 10 we evaluate the peakiness and smoothness of correspondence heatmaps generated by the various methods. Both of these are desirable properties as they lead to better selectivity and interpretability, while reducing ambiguity in the resultant heatmaps. To perform this evaluation we compare the shape of the heatmaps at locations surrounding the point of correspondence. We normalize the heatmaps of the successful matches, for each method, such that their dynamic range is comparable, and their peaks are aligned. We then generate the median heatmaps and analyze the row and column cross-sections, relative to the global maximum peak.

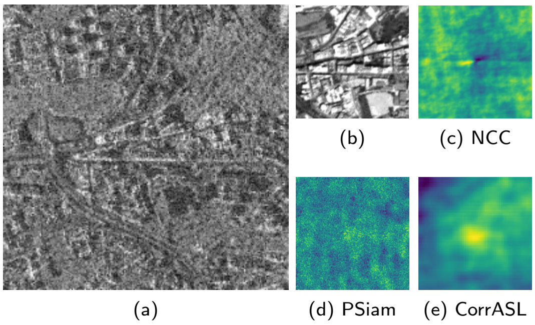
From Figure 10, it is evident that both NCC and PSiam approaches suffer from a high number of local maxima which leads to a lower dynamic range in the heatmaps, and a less interpretable result. Our proposed solution on the other hand has a tendency to produce smooth heatmaps with a single global maximum for accurately matched results.

We further investigate the quality of the produced correspondence heatmaps through a qualitative process by evaluating a subset of example heatmaps. This subset was selected based on scenes where all three methods obtained a similar matching accuracy. We thus evaluated the correspondence heatmaps in three categories, namely, positive matches (less than 1 pixel error), inaccurate matches (between 3 and 5 pixels error) and unsuccessful matches where by the $L2$ error is larger than 7 pixels. An example result for each category can be seen in Figure 11, Figure 12 and Figure 13, respectively. For each heatmap the true point of correspondence is



(a) (b) (c) NCC (d) PSiam (e) CorASL

Figure 11: A positive matching result where (c) NCC, (d) Pseudo-Siamese(Hughes et al., 2018), and the proposed approach (e), could all find the correspondence of the template patch (b) within the search region (a) with an accuracy of ≤ 1 pixel. The true point of correspondence is located at center point of (a),(c),(d) and (e). For brevity only the *valid* region of the heatmaps in (c-e) is depicted.

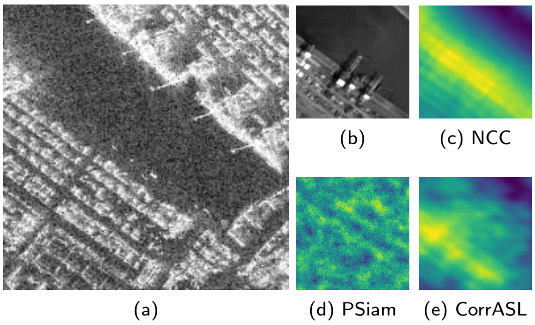


(a) (b) (c) NCC (d) PSiam (e) CorrASL

Figure 12: An inaccurate match, where (c) NCC, (d) Pseudo-Siamese (Hughes et al., 2018), and the proposed approach (e), all had a matching error of between 3 and 5 pixels when matching the template patch (b) within the search region (a). The expected point of correspondence is in the center of (a), (c), (d) and (e), however, we can see it is slightly offset from center in (c),(d) and (e). For brevity only the *valid* region of the heatmaps in (c-e) is depicted.

the center of the search window, and all heatmaps are computed with valid padding.

Figure 11 shows the single global peak produced using the correspondence network, compared to the reasonable NCC result, and the very noisy PSiam heatmap. The same trends continue when observing matches with slight inaccuracies, in Figure 12, although in this case the result achieved with the proposed method looses smoothness and local maxima begin to develop. Finally, in the case of unsuccessful matching, Figure 13, the heatmap shape for all methods deteriorates to have multiple local maxima, although these all occur along the direction of ambiguity. Figure 12 and Figure 13 indicate that our method fails in a predictable manner, and thus the hypothesis, that correspondence heatmaps can be used directly for the detection of outliers, holds true.



(a) (b) (c) NCC (d) PSiam (e) CorrASL

Figure 13: An unsuccessful match, where (c) NCC, (d) Pseudo-Siamese (Hughes et al., 2018), and the proposed approach (e), all had a matching error larger than 7 pixels when matching the template patch (b) within the search region (a). The true point of correspondence is located at center point of (a),(c),(d) and (e). For brevity only the *valid* region of the heatmaps in (c-e) is depicted.

Table 8
Performance of the goodness networks with respect to the test dataset. The first two rows are the results when selecting candidate patches based on the domain specific goodness only. While the last two rows represent the accuracy for the cross domain goodness results when fusing the SAR and optical goodness maps using the minimum and maximum operators, respectively.

Modality	Accuracy	Precision	Recall
SAR	63.6	68.9	69.0
Optical	65.1	69.8	71.3
Cross-Min	62.1	**75.1**	61.6
Cross-Max	**67.0**	66.4	**88.7**

5.3. Goodness Results

To gain an understanding for the performance of the domain specific goodness networks, as well as the effects of minimum or maximum fusion on the cross-domain goodness, we assess the binary classification accuracy with respect to the test dataset. The results for this investigation are described in Table 8.

The overall, relatively low accuracy of the goodness network, see Table 8, highlights the complexity of determining matchable regions across vastly heterogeneous domains, such as SAR and optical. However, by comparing the cross-domain goodness results we see an improvement in the precision of the goodness network when using minimum fusion, and a large improvement in recall when using maximum fusion. These results show how the selection of the fusion algorithm, for producing the cross-domain goodness, has a significant effect on the trade off between the number of identified regions and the confidence of those regions containing good features for matching.

Figure 14 depicts examples of regions with high and low goodness, as well as regions which were incorrectly classified. These example regions were drawn from the cross-domain goodness results generated using minimum fusion.

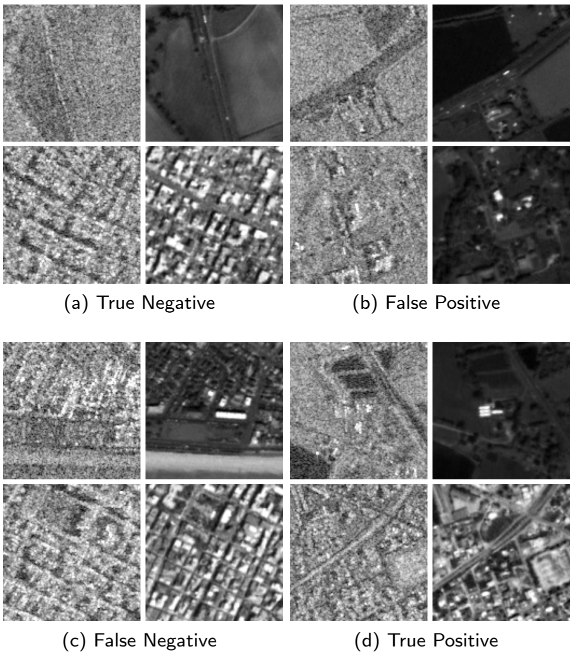


Figure 14: Examples of regions of low and high goodness (a) and (d) respectively, along with misclassified regions (b) and (c). The SAR patch is shown on the left, and optical on the right for each of the patch pairs.

From Figure 14 we can see that the identified regions of high goodness contain strong, unambiguous and discriminable features in both modalities, while the low goodness regions lack these properties. In the case of the false positive regions, strong features do exist in both domains, however, these features are potentially ambiguous or lack discriminability. The same properties can be seen in the false negative results.

As the purpose of the goodness network is to improve the matching accuracy of the correspondence network, by preselecting regions which have a higher probability of being correctly matched, we further evaluate the goodness network through the process of matching. Table 9, shows the matching performance when we match against the test patches which have been identified to have either a high domain specific, or cross-modality goodness. The proportion of the original test dataset which was identified to have high goodness is described as the number of regions (# Regions).

From the results presented in Table 9 it is evident that the pre-filtering of regions, based on goodness, leads to improved matching accuracy and precision over the baseline (Table 7). Furthermore, the low percentage of good regions found in the evaluation dataset hints to the non-optimal choice of using optical domain Harris corners to create the dataset.

Another interesting observation is that the use of minimum fusion, provides a large boost in accuracy and precision with respect to the single modality, and Cross-Max goodness. While Cross-Max goodness leads to the identification of the most regions, although this comes at the cost of decrease in accuracy. Thus these results agree with the observations, made with respect to Table 8, about the selection of fusion approach and the trade-off between precision and recall.

Table 9
The percentage of the dataset used for matching (# Regions), and matching performance obtained when pre-selecting search and template patches based on their domain specific and cross-domain goodness scores. The first row represents the matching results when matching without the pre-selection of good patches.

Goodness	# Regions [%]	≤ 1px [%]	avg. L_2 [px]	mAP [px]
CorrASL	100	46.9	2.1	2.62
SAR+CorrASL	55	54.7	1.97	1.69
Opt+CorrASL	62	54.9	1.94	1.58
Cross-Min+CorrASL	48	**59.8**	**1.62**	**1.24**
Cross-Max+CorrASL	75	53.7	2.01	1.87

While the use of high goodness regions leads to improved matching performance, it comes at the cost of having fewer overall correspondences as the regions are significantly larger than those used to compute point features. This, however, is deemed to be an acceptable trade-off as many downstream tasks such as co-registration (Müller et al., 2012; Suri and Reinartz, 2010; Merkle et al., 2017) and SAR-optical stereogrammetry (Qiu et al., 2018; Bagheri et al., 2018) favour accuracy and spatial diversity over the number of correspondences.

5.4. Outlier Reduction

The final component of the proposed matching pipeline is the outlier reduction network. We evaluate its performance in classifying the correspondence heatmaps of the test dataset. We further investigate the effects the inclusion of the ORN has on matching accuracy and finally we evaluate the full matching framework in an end-to-end manner on the test dataset.

A binary classification accuracy of 81%, with a precision of 76.1% and a recall of 89.5%, was achieved when evaluating the ORN on the test dataset. This shows that the classification of successful matches can be achieved based on the correspondence heatmap alone. Figure 15, provides visual examples of both positive and negative classification results.

The correspondence surface shapes, as shown in Figure 15, highlight that the network relies on more than just the local characteristics of the peak for classification although these do appear to have a relatively strong effect.

In Table 10 we investigate the effect of the outlier reduction network on matching performance. To do so we apply the ORN to the matching heatmaps of both the test dataset matching results, as shown in Table 7, as well as the minimum fusion (Cross-Min) goodness results, Table 9. The latter resulting in an equivilant end-to-end evaluation of the network.

From Table 10 it is clear that the addition of the outlier

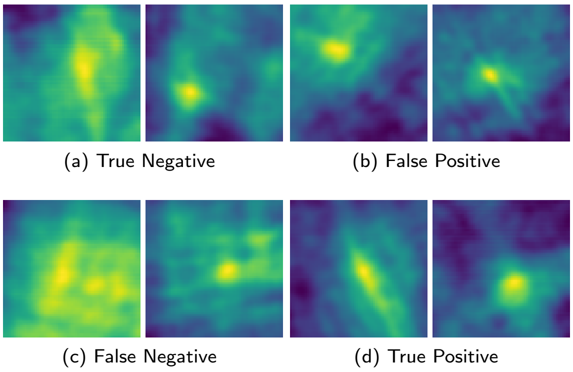

(a) True Negative
(b) False Positive
(c) False Negative
(d) True Positive

Figure 15: Examples of heatmaps corresponding to incorrectly (a) and correctly (d) matched regions, along with mis-classified correspondence heatmaps (b) and (c). The ORN only makes use of the *valid* region of the heatmap for classification.

Table 10
Final matching results after removing outliers as classified by the ORN.

Dataset	≤ 1px [%]	avg. L_2 [px]	mAP [px]
CorrASL+ORN	54.1	1.30	1.09
Cross-Min+CorrASL+ORN	65.2	1.71	1.13

reduction network substantially increases the accuracy of the resultant set of correspondences, irrespective of the features or regions used for matching. However, the matching performance using the the full framework in an end-to-end manner achieves an overall better result with higher accuracy and improved precision.

5.5. Large-Scale Scene Matching

While we have evaluated the performance the individual sub-components of our framework, as well as the framework as a whole, these investigations have remained limited to the patch-based test dataset. Thus to fully evaluate the end-to-end performance and applicability of our proposed framework, we apply it to the task of determining correspondence on a large-scale test scene (approximately 0.8km × 1.8km) which has not undergone manual co-registration. The example scene is taken from the city of Portsmouth, England and is depicted in Figure 16 with the final set of correspondences overlaid.

To examine the improvement in co-registration we take the mean shift derived from the final set of correspondences and apply this to the optical scene in order to align it with the SAR image. The checkerboard overlays in Figure 17a,c depict subsets of the original, non-coregistered scene. While Figure 17b,d show the same subsets after the alignment has been adjusted using the mean shift of the predicted set of correspondences. The mean shift was found to be $(11.03, -12.74)$ pixels with a standard deviation of $(1.99, 2.20)$ pixels in the x and y dimensions, respectively.

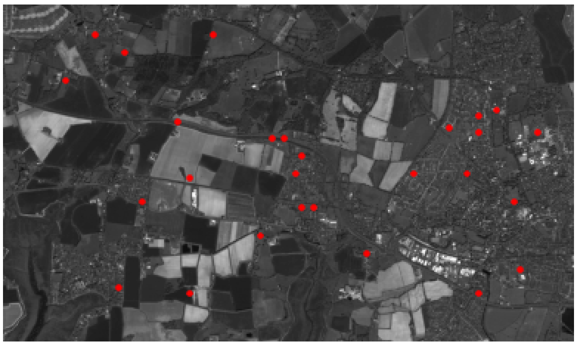


Figure 16: The final set of correspondences superimposed on the PRISM optical image, taken near the city of Portsmouth, England.

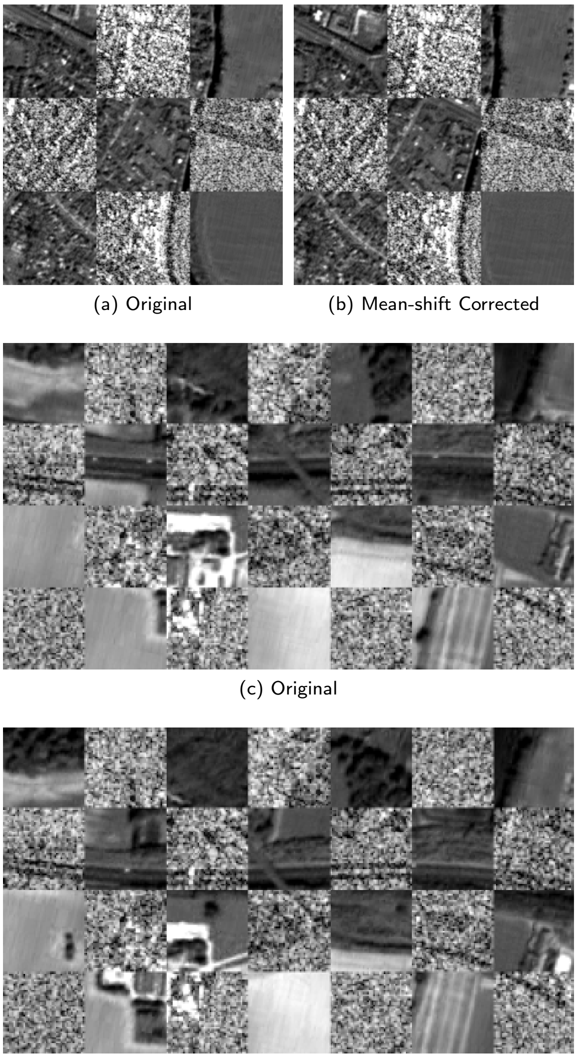
From Figure 16 it can be seen that our proposed framework does not produce a large set of correspondences. However, Figure 17 highlights the accuracy and utility of these correspondences in being able to, seemingly, accurately coregister SAR and optical imagery. While our method for correcting co-registration can be improved by using the correspondences as GCPs to correct the overall optical sensor model (Müller et al., 2012), in the case of our relatively small and flat test scene, such an approach is unlikely to provide a large increase in accuracy over the mean shift method which we followed.

6. Conclusion

In this paper we proposed an end-to-end framework for the sparse matching of SAR and optical imagery. The framework consists of three sub-components, each of which were trained to perform a specific task within the standard proposal, matching, outlier detection pipeline. The goodness network proposes candidate patches with a high chance of being matchable in both domains. The correspondence network performs cross correlation on a multi-scale, feature space to produce a correspondence heatmap, which is finally filtered by the outlier reduction network in order to reduce the number of false positive correspondences.

We demonstrated that, individually, each of these sub-components improves the matching accuracy and precision achieved on a test dataset in comparison to existing SAR-optical matching approaches, namely NCC (Burger and Burge, 2009) and Pseudo-Siamese (Hughes et al., 2018). We further evaluated the pipeline in and end-to-end manner and showed that it was able to achieve and average L_2 (distance to ground truth correspondence) of 1.71 pixels with a precision of 1.13 pixels. Finally, we demonstrated the effectiveness of our framework in producing an accurate set of correspondences which can be applied to the task of improving the overall geo-localization accuracy of optical imagery.

Still, there is room for improvement: the size of the final correspondence set is mostly limited by the goodness and outlier reduction networks. Thus in future work we will investigate alternative architectures for the goodness network which can operate on the full scale image while still account-



(a) Original

(b) Mean-shift Corrected

(c) Original

(d) Mean-shift Corrected

Figure 17: Checkerboard overlays comparing the alignment of a TerraSAR-X image to the original (non-coregisterd), and mean-shifted optical imagery for two subsets of the Portsmouth, England test scene. The original imagery is depicted in (a) and (c), while the mean-shift, correct imagery is shown in (b) and (d). All images have a pixel spacing of 2.5 meters.

ing for the offsets between domains. Furthermore, recent research has shown success in progressive training strategies, whereby multiple sub-components are trained in an iterative and alternating manner (Karras et al., 2017; Shaham et al., 2019). The application of such an approach to training the goodness and correspondence network could reduce the effects of the non-optimally selected training points, by allowing the network to iteratively refine these locations, and thus potentially lead to improved performance.

Acknowledgements

This work is supported by the German Research Foundation (DFG) as grant SCHM 3322/1-1

References

Bagheri, H., Schmitt, M., d'Angelo, P., Zhu, X.X., 2018. A framework for SAR-optical stereogrammetry over urban areas. ISPRS J. Photogramm. Remote Sens. 146, 389–408.

Balntas, V., Johns, E., Tang, L., Mikolajczyk, K., 2016a. PN-Net: Conjoined triple deep network for learning local image descriptors. arXiv preprint arXiv:1601.05030 .

Balntas, V., Riba, E., Ponsa, D., Mikolajczyk, K., 2016b. Learning local feature descriptors with triplets and shallow convolutional neural networks, in: Procedings of the British Machine Vision Conference 2016, British Machine Vision Association. p. 3.

Burger, W., Burge, M.J., 2009. Principles of Digital Image Processing. volume 54. Springer London.

Bürgmann, T., Koppe, W., Schmitt, M., 2019. Matching of TerraSAR-X derived ground control points to optical image patches using deep learning. ISPRS J. Photogramm. Remote Sens. 158, 241–248.

Citak, E., Bilgin, G., 2019. Visual saliency aided SAR and optical image matching, in: 2019 Innovations in Intelligent Systems and Applications Conference (ASYU), IEEE. IEEE. pp. 1–5.

Dellinger, F., Delon, J., Gousseau, Y., Michel, J., Tupin, F., 2015. SAR-SIFT: A SIFT-like algorithm for SAR images. IEEE Trans. Geosci. Remote Sensing 53, 453–466.

Dusmanu, M., Rocco, I., Pajdla, T., Pollefeys, M., Sivic, J., Torii, A., Sattler, T., 2019. D2-Net: A trainable CNN for joint description and detection of local features, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE. pp. 8092–8101.

Fischer, P., Dosovitskiy, A., Brox, T., 2014. Descriptor matching with convolutional neural networks: A comparison to SIFT. arXiv preprint arXiv:1405.5769 .

Fischler, M.A., Bolles, R.C., 1981. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM 24, 381–395.

Gong, M., Zhao, S., Jiao, L., Tian, D., Wang, S., 2014. A novel coarse-to-fine scheme for automatic image registration based on SIFT and mutual information. IEEE Trans. Geosci. Remote Sensing 52, 4328–4338.

Han, X., Leung, T., Jia, Y., Sukthankar, R., Berg, A.C., 2015. MatchNet: Unifying feature and metric learning for patch-based matching, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE. pp. 3279–3286.

Hariharan, B., Arbelaez, P., Girshick, R., Malik, J., 2015. Hypercolumns for object segmentation and fine-grained localization, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE. pp. 447–456.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification, in: 2015 IEEE International Conference on Computer Vision (ICCV), IEEE. pp. 1026–1034.

Hoffmann, S., Brust, C.A., Shadaydeh, M., Denzler, J., 2019. Registration of high resolution SAR and optical satellite imagery using fully convolutional networks, in: 2019 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), IEEE. IEEE. pp. 5152–5155.

Hughes, L.H., Merkle, N., Burgmann, T., Auer, S., Schmitt, M., 2019. Deep learning for SAR-optical image matching, in: 2019 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), IEEE. pp. 4877–4880.

Hughes, L.H., Schmitt, M., 2019. A semi-supervised approach to SAR-optical image matching. ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci. IV-2/W7, 71–78.

Hughes, L.H., Schmitt, M., Mou, L., Wang, Y., Zhu, X.X., 2018. Identifying corresponding patches in SAR and optical images with a pseudo-Siamese CNN. IEEE Geosci. Remote Sensing Lett. 15, 784–788.

Iglovikov, V., Shvets, A., 2018. Ternausnet: U-Net with VGG11 encoder pre-trained on imagenet for image segmentation. arXiv preprint abs/1801.05746. arXiv:1801.05746.

Karras, T., Aila, T., Laine, S., Lehtinen, J., 2017. Progressive growing of GANs for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196 .

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 .

Kuppala, K., Banda, S., Barige, T.R., 2020. An overview of deep learning methods for image registration with focus on feature-based approaches. International Journal of Image and Data Fusion 0, 1–23.

Li, J., Hu, Q., Ai, M., 2020. RIFT: Multi-modal image matching based on radiation-variation insensitive feature transform. IEEE Trans. on Image Process. 29, 3296–3310.

Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision 60, 91–110.

Ma, W., Wen, Z., Wu, Y., Jiao, L., Gong, M., Zheng, Y., Liu, L., 2017. Remote sensing image registration with modified SIFT and enhanced feature matching. IEEE Geosci. Remote Sensing Lett. 14, 3–7.

Ma, W., Zhang, J., Wu, Y., Jiao, L., Zhu, H., Zhao, W., 2019. A novel two-step registration method for remote sensing images based on deep and local features. IEEE Trans. Geosci. Remote Sensing 57, 4834–4843.

Merkle, N., Luo, W., Auer, S., Müller, R., Urtasun, R., 2017. Exploiting deep matching and SAR data for the Geo-localization accuracy improvement of optical satellite images. Remote Sensing 9, 586.

Mishchuk, A., Mishkin, D., Radenoviundefined, F., Matas, J., 2017. Working hard to know your neighbor's margins: Local descriptor learning loss, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 4829–4840.

Mou, L., Schmitt, M., Wang, Y., Zhu, X.X., 2017. A CNN for the identification of corresponding patches in SAR and optical imagery of urban scenes, in: Proc. JURSE, IEEE, Dubai. pp. 1–4.

Müller, R., Krauß, T., Schneider, M., Reinartz, P., 2012. Automated georeferencing of optical satellite data with integrated sensor model improvement. photogramm eng remote sensing 78, 61–74.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. PyTorch: An imperative style, high-performance deep learning library, in: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 32. Curran Associates, Inc., pp. 8024–8035.

Qiu, C., Schmitt, M., Zhu, X.X., 2018. Towards automatic SAR-optical stereogrammetry over urban areas using very high resolution imagery. ISPRS J. Photogramm. Remote Sens. 138, 218–231.

Schmitt, M., Tupin, F., Zhu, X.X., 2017. Fusion of SAR and optical remote sensing data – challenges and recent trends, in: 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), IEEE, Fort Worth, TX, USA. pp. 5458–5461.

Schneider, M., Müller, R., Krauss, T., Reinartz, P., Hörsch, B., Schmuck, S., 2010. Urban Atlas – DLR processing chain for orthorectification of PRISM and AVNIR-2 images and TerraSAR-X as possible GCP source. Internet Proceedings , 1–6.

Shaham, T.R., Dekel, T., Michaeli, T., 2019. SinGAN: Learning a generative model from a single natural image, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE. pp. 4570–4580.

Simo-Serra, E., Trulls, E., Ferraz, L., Kokkinos, I., Fua, P., Moreno-Noguer, F., 2015. Discriminative learning of deep convolutional feature point descriptors, in: 2015 IEEE International Conference on Computer Vision (ICCV), IEEE. pp. 118–126.

Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition, in: International Conference on Learning Representations.

Smith, L.N., 2017. Cyclical learning rates for training neural networks, in: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE. IEEE. pp. 464–472.

Suri, S., Reinartz, P., 2010. Mutual-information-based registration of TerraSAR-X and Ikonos imagery in urban areas. IEEE Trans. Geosci. Remote Sensing 48, 939–949.

Suri, S., Schwind, P., Uhl, J., Reinartz, P., 2010. Modifications in the SIFT operator for effective SAR image matching. International Journal of Image and Data Fusion 1, 243–256.

Vargas-Muñoz, J.E., Lobry, S., Falcão, A.X., Tuia, D., 2019. Correcting rural building annotations in OpenStreetMap using convolutional neural networks. ISPRS J. Photogramm. Remote Sens. 147, 283–293.

Wang, S., Quan, D., Liang, X., Ning, M., Guo, Y., Jiao, L., 2018. A deep learning framework for remote sensing image registration. ISPRS J. Photogramm. Remote Sens. 145, 148–164.

Woo, S., Park, J., Lee, J.Y., Kweon, I.S., 2018. CBAM: Convolutional block attention module, in: Proceedings European Conference on Computer Vision 2018, Springer International Publishing, Cham. pp. 3–19.

Xiang, Y., Wang, F., You, H., 2018. OS-SIFT: A robust SIFT-like algorithm for high-resolution optical-to-SAR image registration in suburban areas. IEEE Trans. Geosci. Remote Sensing 56, 3078–3090.

Ye, Y., Shen, L., 2016. HOPC: A novel similarity metric based on geometric structural properties for multi-modal remote sensing image matching. ISPRS Annals 3, 9.

Yi, K.M., Trulls, E., Lepetit, V., Fua, P., 2016. LIFT: Learned invariant feature transform, in: Proceedings European Conference on Computer Vision, Springer. pp. 467–483.

Zagoruyko, S., Komodakis, N., 2015. Learning to compare image patches via convolutional neural networks, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE. pp. 4353–4361.

4.7 Supplementary Publications

This section lists additional publications which were created during the period of the doctoral process. These publications are listed here as they do not directly fit within the scope of this thesis, or were not subject to peer-review. However, these publications are included to highlight the diversity and completeness in the body of work taken on, by the author of this dissertation. The author was directly involved with the ideation and implementation of various methodologies presented in these publications.

Hughes, L. H., Schmitt, M., & Zhu, X. X. (2018). Generative adversarial networks for hard negative mining in CNN-based SAR-optical image matching. *Proc. International Geoscience and Remote Sensing Symposium Conference*, 4391–4394.

Hughes, L. H., Merkle, N., Bürgmann, T., Auer, S., & Schmitt, M. (2019). Deep learning for SAR-optical image matching. *Proc. International Geoscience and Remote Sensing Symposium Conference*, 4877–4880.

Hughes, L. H., Streicher, S., Chuprikova, E., & Du Preez, J. (2019). A cluster graph approach to land cover classification boosting. *Data*, *4*(1), 10.

Schmitt, M., Hughes, L., Körner, M., & Zhu, X. (2018). Colorizing Sentinel-1 SAR images using a variational autoencoder conditioned on Sentinel-2 imagery. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *42*, 2.

Schmitt, M., Hughes, L., Qiu, C., & Zhu, X. (2019a). SEN12MS-A curated dataset of georeferenced multi-spectral Sentinel-1/2 imagery for deep learning and data fusion. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 153–160.

Schmitt, M., Hughes, L. H., Qiu, C., & Zhu, X. X. (2019b). Aggregating cloud free Sentinel-2 images with Google Earth Engine. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *4*.

Soal, K. I., Volkmar, R., Hughes, L., Govers, Y., & Böswald, M. (2019). Evolutionary based approach to modal parameter identification. *Proc. International Operational Modal Analysis Conference.*

Volkmar, R., Soal, K. I., Hughes, L. H., Govers, Y., & Böswald, M. (2019). Automated optimization of output only modal parameter identification. *Proc. International Operational Modal Analysis Conference.*

Zhu, X., Hu, J., Qiu, C., Shi, Y., Kang, J., Mou, L., Bagheri, H., Haberle, M., Hua, Y., Huang, R., Hughes, L. H., Li, H., Sun, Y., Zhang, G., Han, S., Schmitt, M., & Wang, Y. (2020). So2Sat LCZ42: A benchmark dataset for global local climate zones classification. *IEEE Geoscience and Remote Sensing Magazine.*

5. Experimental Evaluation

Since the various architectures proposed throughout Chapter 3 were all trained and evaluated under different data conditions and assumptions, it is difficult to compare their results directly. To this end, and in the name of completeness, this chapter aims to evaluate the various contributions of this thesis in a comparable manner.

The investigation is focussed on evaluating the performance of the various correspondence networks proposed throughout this thesis (see Section 3.1 and Section 3.2.2). Furthermore, the effects of using *good* regions for matching, and performing outlier removal using the ORN (see Section 3.3.3) are evaluated in terms of their influence on the accuracy of the resultant correspondence set. Finally, the comprehensive matching framework proposed in Section 3.3.4 is assessed within the frame of a large-scale scene co-registration problem.

5.1 Experimental Setup

In order to train and evaluate the previously proposed methods (see Chapter 3) in a comparable manner, the evaluations need to be conducted under similar data and experimental conditions. To this end common datasets with equivalent train/test splits need to be created, and evaluation mechanisms and metrics need to be defined.

This section details the setup and definitions required for comparable experiments to be conducted. Furthermore, to gain a more complete picture of the contributions of this thesis, within the scope of the broader work on SAR-optical matching, a few openly available baseline approaches are introduced.

5.1.1 Datasets

As previously mentioned, to produce comparable results the various architectures need to be trained and evaluated on similar data. As the focus of this thesis is on matching high-resolution SAR and optical imagery the Urban Atlas-based dataset, proposed in Section 3.1.1, was chosen to form the basis of the data used for these investigations.

The Urban Atlas dataset was utilized as described in Section 3.1.1, with the training, validation and testing datasets being created from imagery of different cities around Europe, as depicted in Figure 3.3. This lead to the creation of sub-datasets containing 40,314, 4,205 and 6,353 corresponding high-resolution SAR-optical patch pairs, for training, validation and testing respectively. Each patch pair consists of a SAR and optical patch of 256×256 pixels, with the point of correspondence between the patches being located at the center pixel.

5.1.2 Training and Evaluation

As each architecture is trained differently, with unique input data and target label criteria, the previously described datasets needed to be further adapted to each architecture and training mechanism.

The pseudo-siamese architecture (PSiam) proposed in Section 3.1.2 is trained using (non-)corresponding SAR-optical patch pairs, with binary labels for supervision. Thus the training and validation datasets are further modified to create suitable input data and labels for training the PSiam network. The corresponding patch pairs are created by cropping 128 × 128 pixel patches around the center pixel of the SAR and optical imagery. The non-corresponding pairs are then formed by cropping 128 pixel patches randomly from one of the four corners of the larger SAR patch and pairing this with the optical patch from the corresponding pair.

The semi-supervised network (SSNet), introduced in Section 3.2.2, requires a supervised and unsupervised dataset for training. As the supervised dataset requirements match that of the PSiam network, the same dataset is used. On the other hand, the unsupervised dataset has no strict criteria and requires no labelled data. Thus to ensure equivalency, in the amount of data used to train the networks, the unsupervised dataset is created by removing target labels from 75% of the training samples. The removal of these supervised samples is done such that overlapping patches do not create data leakage events (i.e. in a spatially diverse manner). Furthermore, to better understand the performance of the feature vector based matching network at the center of the SSNet design, the SSNet is also trained in a fully supervised manner using all the available labels.

In a different manner to the previous two approaches the feature-space correlation network (CorrASL), described in Section 3.1.3, is trained using a large 256 × 256 pixel SAR search patch, and a 128 × 128 optical template patch, and is supervised with a 2D Kronecker delta function. The procedure for creating and training the CorrASL network using the Urban Atlas dataset is described in Section 3.1.3.

Each architecture was then trained from scratch on the Urban Atlas dataset until convergence. The optimal hyperparameters and specific training algorithms are detailed in the publications from which this thesis is comprised, see Chapter 4.

Based on the expected inputs and outputs of the *goodness* and outlier reduction networks, which form the supporting tasks in the comprehensive matching framework described in Section 3.3.4, the correspondence networks were evaluated in terms of a template search problem. Thus the PSiam network and SSNet were applied in a rolling window manner in order to generate correspondence heatmaps. These could then be used to determine the best point of correspondence for the center pixel of the optical template patch. As the CorrASL architecture is inherently designed around solving the correspondence problem as a template search problem, no modifications to the inference process were required. Based on this formulation of the SAR-optical correspondence problem, the same test dataset was used for all three networks, thus ensuring comparable results across the different methodologies.

5.1.3 Evaluation Metrics

To evaluate the performance of the correspondence networks, as well as the effects of the goodness and outlier reduction networks on matching, various metrics need to be defined.

An accurate correspondence is defined as being within a 1 pixel radius of the ground truth location. Thus matching accuracy can be interpreted as the percentage of the test dataset which was accurately matched to within 1 pixel of the ground truth. Additionally, the matching accuracy of a network can be evaluated as the average L_2 distance, taken over the final correspondence set, between the determined point of correspondence and the ground truth location.

The matching precision of the final correspondence set is defined as the mean average precision (mAP), whereby the standard deviation is used as a measure of precision. The precision is calculated across the set of positive correspondences, as the threshold for accuracy is increased from 1 pixel up to the maximum allowable offset of 32 pixels. These individual precisions are then averaged to compute the mAP. Prior to outlier removal, the precision (standard deviation) is a biased metric as it cannot account for outliers, thus the mAP provides a better overall evaluation of the networks precision as it equally weights the precision under differing definitions of accuracy.

5.1.4 Baseline Methods

Three baseline methods are used to better evaluate the performance of the proposed SAR-optical matching methodologies in relation to the scope of existing work on the topic. Based on availability and reproducibility the normalized cross-correlation (NCC), mutual information (MI) (Suri & Reinartz, 2010) and deep matching architecture (DeepMatch) proposed by Merkle (2018) are used as baseline methods for comparison.

Although neither NCC nor MI are well suited to large geometric differences, both methods have still seen widespread use in SAR-optical matching due to their simplicity and ease of application. Both methods were described in Section 2.2.2, and are thus not described further.

The deep matching network (DeepMatch) proposed by Merkle, Luo, et al. (2017) is based on a 2-stream siamese architecture, and directly outputs correspondence heatmaps which are created using the dot product between the template feature vector and search space feature vectors. An overview of the DeepMatch architecture is depicted in Figure 5.1.

The DeepMatch architecture frames the matching problem as a multi-class classification problem, whereby the network tries to predict the offset of the template patch within the search window using predefined offset classes. Thus to match the data requirements of the network the datasets used for training the CorrASL network were used as the basis for training the DeepMatch architecture, with some minor modifications. The first modification was to crop template patches of 201 × 201 pixels and search regions of 221 × 221 pixels, this is based on the requirements of the DeepMatch architecture. Thus it should be noted the DeepMatch network can only estimate correspondence within

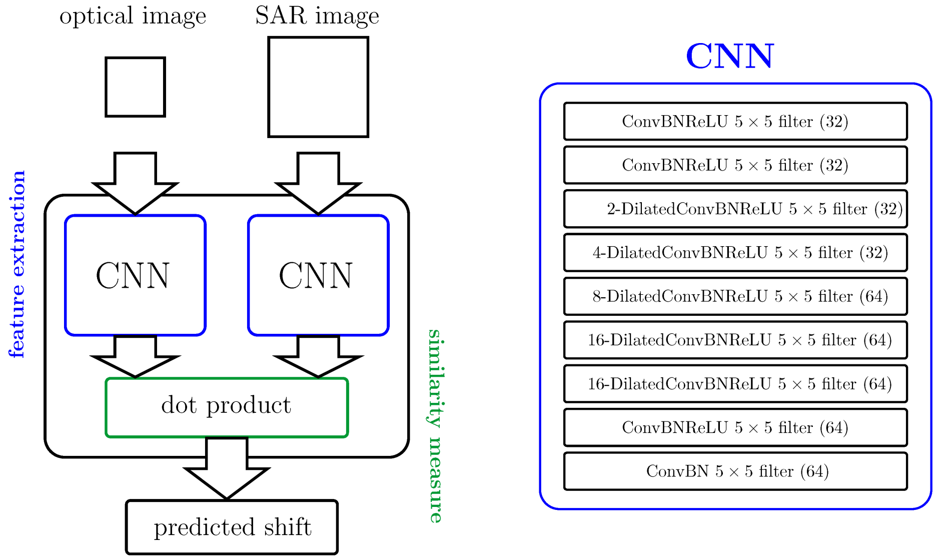

optical image
SAR image
feature extraction
CNN
CNN
dot product
similarity measure
predicted shift
CNN
ConvBNReLU 5 × 5 filter (32)
ConvBNReLU 5 × 5 filter (32)
2-DilatedConvBNReLU 5 × 5 filter (32)
4-DilatedConvBNReLU 5 × 5 filter (32)
8-DilatedConvBNReLU 5 × 5 filter (64)
16-DilatedConvBNReLU 5 × 5 filter (64)
16-DilatedConvBNReLU 5 × 5 filter (64)
ConvBNReLU 5 × 5 filter (64)
ConvBN 5 × 5 filter (64)

FIGURE 5.1: The DeepMatch network architecture with a detailed overview of the convolutional layers of each stream. As the network is based on a siamese architecture the weights of the template and search streams are shared. Image taken from (Merkle, Luo, et al., 2017).

a maximum radius of 10 pixels from the true point of correspondence. The second modification was to form a 441-dimensional one-hot encoded target vector whereby each position represents a specific offset of the template within the window. Using these modification the DeepMatch network was then trained in accordance with the description provided in (Merkle, Luo, et al., 2017).

5.2 Computational Performance

Given the diversity of available SAR and optical sensors, and the lack of a global-scale SAR-optical correspondence dataset it is likely that deep learning matching approaches will need to be retrained, or at least fine-tuned to perform optimally with different sensors, or in other geographical locations. Thus the training time of the various approaches is an important consideration in evaluating the usability of deep matching methods.

Furthermore, the inference time of various matching approaches is an equally important consideration, as the trade-off between computational efficiency and accuracy is largely dependant on the application. For instance, in matching across large spatial regions it is impractical to run slow correspondence algorithms in order to gain a slight improvement in accuracy.

Thus the computational performance of the various approaches is evaluated in terms of their training and inference times. The results of these evaluations are presented in Table 5.1. Since absolute timings hold little meaning, due to the dependence on the underlying hardware and the number of samples evaluated, the computational efficiency is reported in computational units (CUs) where 1 CU is defined as the time taken for the NCC baseline algorithm to evaluate a single correspondence.

Table 5.1: The computational performance of the various approaches in terms of training and inference time. Measurements are relative to one computational unit (CU) which is defined by the time taken for NCC to evaluate one correspondence. It should be noted that algorithms marked with (*) are largely restricted to CPU processing for inference, due to system and algorithm constraints.

Method	**Training** [CU/Epoch]	**Inference** [CU/Sample]
NCC	N.A.	1
MI (Suri & Reinartz, 2010)*	N.A.	417
DeepMatch (Merkle, Luo, et al., 2017)	86,466	4
PSiam (Section 3.1.2)*	8,915	404
CorrASL (Section 3.1.3)	54,665	9
SSNet 100% (Section 3.2.2)	154,516	19
SSNet 25% (Section 3.2.2)	134,240	19

From Table 5.1 it is clear that similarity metrics which are not inherently designed around matching using correlation or the dot product, such as MI and PSiam, are significantly less performant during inference. On the other hand, the remaining networks have a reasonably similar performance with the network depth and complexity, having the most significant effect on the final computational efficiency. In terms of training time, the SSNet, with its complex training algorithm, is by far the slowest approach. Although the PSiam architecture is the most performant network to train, the fixed fusion stage of the architecture severely limits its performance during inference. The CorrASL network provides a good trade-off between inference efficiency and training time.

5.3 Correspondence Results

The various proposed and baseline matching methodologies are evaluated in both a quantitative as well as a qualitative manner. The evaluation is performed by matching between the candidate patch pairs defined in the previously described Urban Atlas-based test dataset. Under these conditions the template patch is cropped around the center pixel of the optical patch, which corresponds to a ground-level Harris corner feature in the optical domain, as outlined in Section 3.1.1. The template patches used for evaluation are 128×128 pixels in size for all networks. With the exception of the DeepMatch baseline, which requires template patches of 201×201 pixels. The size of the search patches are 256×256 pixels in all cases.

Using these test data and the metrics defined in Section 5.1.3, the performance of the deep matching architectures developed within this thesis are compared against baseline matching approaches. The results of this evaluation are detailed in Table 5.2.

From the results presented in Table 5.2 it is clear that the CorrASL network significantly outperforms other SAR-optical matching methodologies. As the CorrASL architecture is somewhat comparable to the PSiam and DeepMatch architectures in

Table 5.2: A comparison of the matching accuracy and precision achieved by various SAR-optical matching approaches on the Urban Atlas-based dataset.

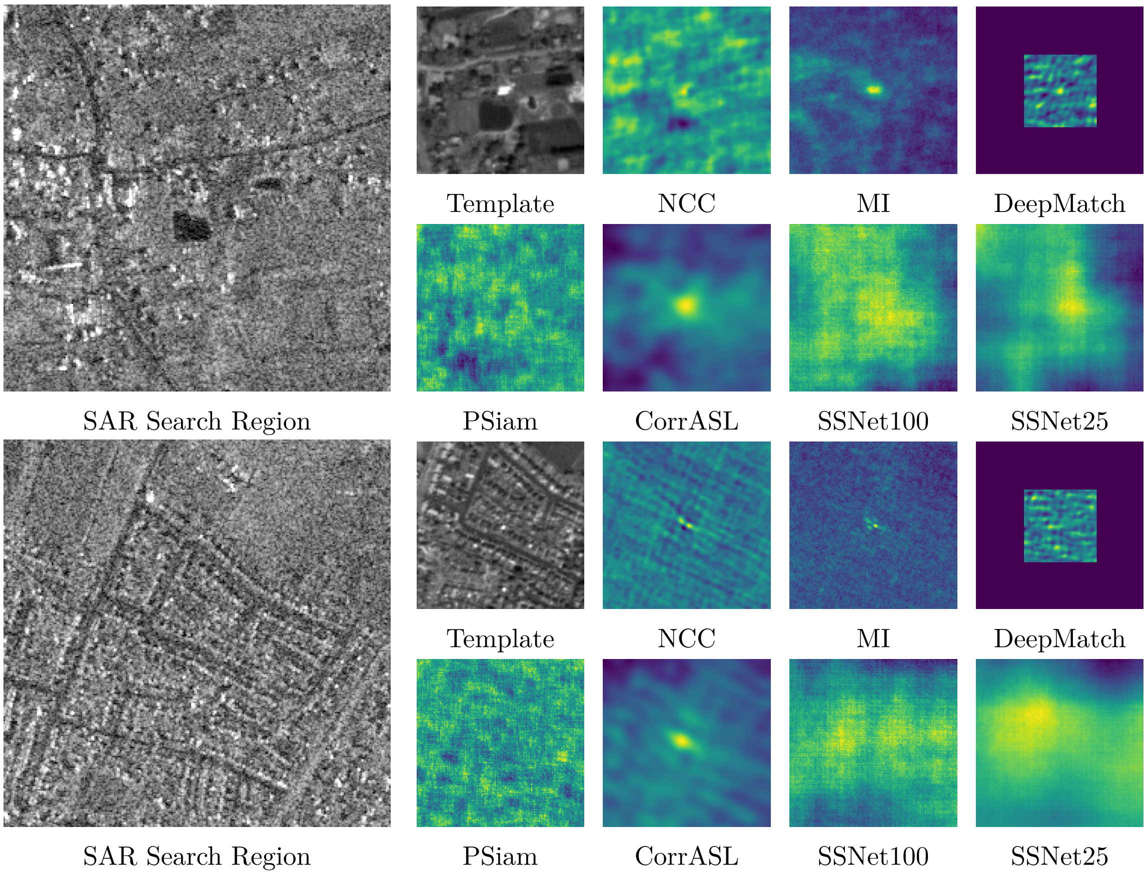
	Matching Accuracy		Matching Precision
Method	≤ 1 pixel [%]	μ [pixel]	mAP [pixel]
NCC	8.2	7.85	6.81
MI (Suri & Reinartz, 2010)	13.7	4.25	3.88
DeepMatch (Merkle, 2018)	14.3	4.37	3.32
PSiam (Section 3.1.2)	18.4	5.22	5.93
CorrASL (Section 3.1.3)	46.9	2.1	2.62
SSNet 100% (Section 3.2.2)	9.1	13.82	11.47
SSNet 25% (Section 3.2.2)	7.6	19.67	12.55

terms of depth and parameters, the improved performance can be attributed to the use of hypercolumns, which encode image features in a multi-scale feature space that can be efficiently matched using standard cross-correlation operators.

More generally, it can be seen that supervised deep matching methods provide improved matching accuracy almost across the board. With the exception of the SSNet variants, which fail to outperform the all the intensity-based baseline methods. As previously discussed in Section 3.2.2, the lower performance of the SSNet architecture is primarily due to the non-complementarity between the reconstruction loss and the matching loss used to train the network. This hypothesis is further corroborated by the fact that the SSNet frames SAR-optical matching in terms of feature vectors, in the same manner as the DeepMatch architecture, and relies on a VGG backend which is similar to the PSiam network. Therefore it should be able to achieve comparable performance to these two networks when trained in a fully-supervised manner. Thus more work is required to reformulate the unsupervised feature representation learning in a manner which better supports matching applications. However, even under these non-optimal conditions, both the fully supervised and partially supervised variants achieve similar matching accuracy as a standard NCC approach.

The design of the PSiam network was based around the idea that the network should have independent streams for each modality, such that modality-specific features can be learnt before fusion. Based on Table 5.2, it is clear that this design decision leads to improved accuracy over the siamese-based DeepMatch network. However, the same improvements are not reflected in the matching precision or average matching error. The design of the fusion network and the formulation of matching as a binary classification problem are the driving factors behind the networks reduced precision. Under this design and training formulation, the network does not learn the spatial structures associated with correspondence as a search problem, i.e. each evaluation in the search region is treated as an independent event and thus the network does not learn to encode spatial consistency as part of the result.

Apart from the matching accuracy and precision, the structure and smoothness of



Template NCC MI DeepMatch

SAR Search Region PSiam CorrASL SSNet100 SSNet25

Template NCC MI DeepMatch

SAR Search Region PSiam CorrASL SSNet100 SSNet25

FIGURE 5.2: Exemplary correspondence heatmaps generated by the various approaches when matching the optical template within the SAR search region, across scenes with unique and unambiguous structures. The expected point of correspondence is located in the center of the heatmap, and only the valid region of heatmaps are depicted for brevity.

the generated correspondence heatmaps are essential factors to consider if the correspondence network is to be incorporated into the matching framework proposed in Section 3.3.4. As previously mentioned, the reason for this is that the ORN relies on the structure of the correspondence heatmaps to identify inaccurate matches. Thus examples of the correspondence heatmaps produced by each of the various approaches, under increasingly difficult matching conditions, are illustrated in Figure 5.2, Figure 5.3, and Figure 5.4, respectively. It should be noted that only the valid region of the heatmaps are depicted for brevity, and the colormap used ranges from blue to yellow, with yellow representing the highest matching score.

The candidate patches presented in Figure 5.2 contain unique salient features in both the SAR and optical modalities. Thus the majority of matching approaches were able to determine the point of correspondence correctly. However, only the heatmaps produced by the CorrASL and MI methodologies are representative of optimal matching in both examples, i.e. contain a single, narrow peak at the point of correspondence. Even though successful matching was achieved by the NCC, DeepMatch and PSiam approaches, their respective heatmaps contain multiple peaks and no definite structure which could be used by the ORN network to classify the correspondence as a success.

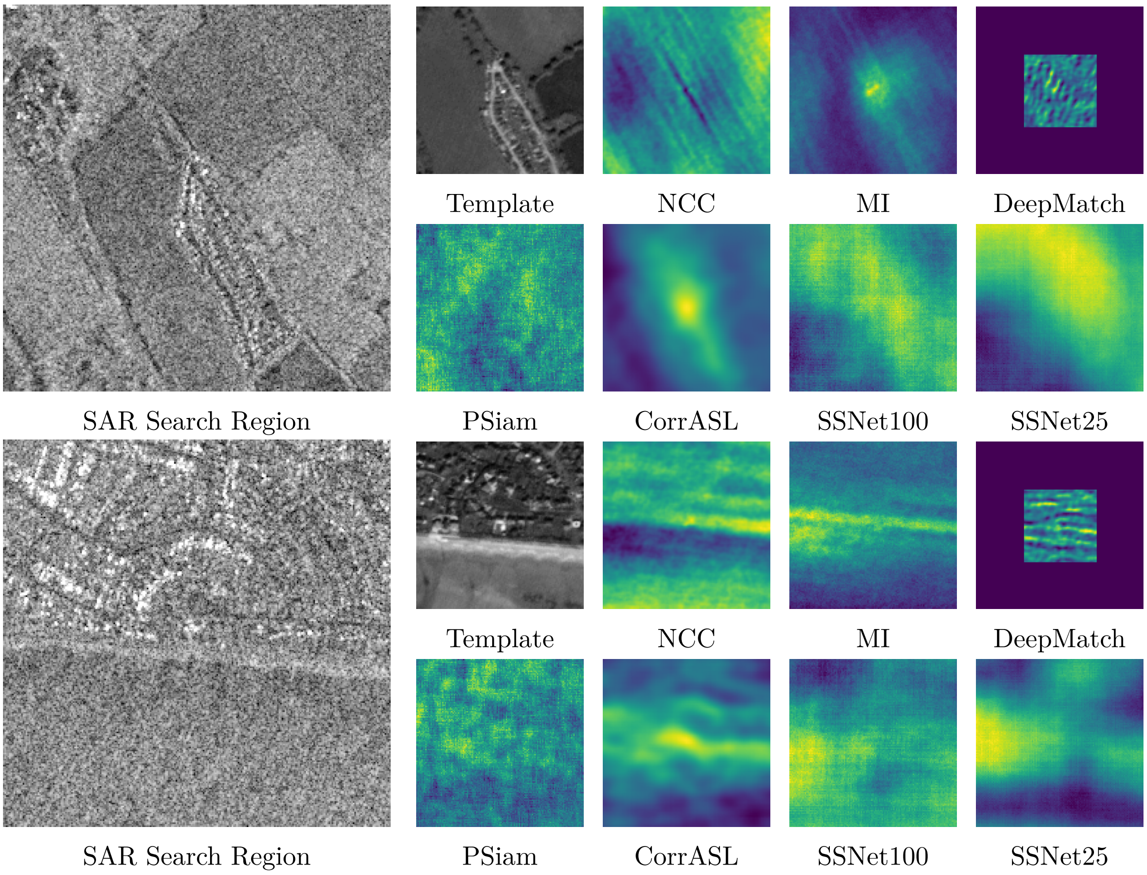


FIGURE 5.3: Exemplary correspondence heatmaps generated by the various approaches when matching the optical template within the SAR search region, across scenes with strong structure in one orientation which could lead to matching ambiguities. The expected point of correspondence is located in the center of the heatmap, and only the valid region of heatmaps are depicted for brevity.

On the other hand, SSNet variants produce smoother heatmaps with a more consistent structure, although the localization of the point of correspondence is less accurate than in other approaches.

The same trends can be seen in the case of the examples presented in Figure 5.3. However, the candidate patches in these examples contain features which could lead to ambiguous matching results in a specific direction. This phenomenon is observable in the spread of the correspondence peak being elongated along the ambiguous direction. Once again, even under more difficult matching conditions, the heatmaps produced by the CorrASL method provide a strong response at the point of correspondence and predictably capture the structure of the matching ambiguities.

In both examples depicted in Figure 5.4, the CorrASL network is the only methodology which is still able to determine the correct point of correspondence accurately. Furthermore, all the approaches, except for DeepMatch, manage to produce heatmaps which capture the matching ambiguities in a somewhat reasonable manner.

Based on these evaluations, it is clear that the CorrASL architecture provides robust

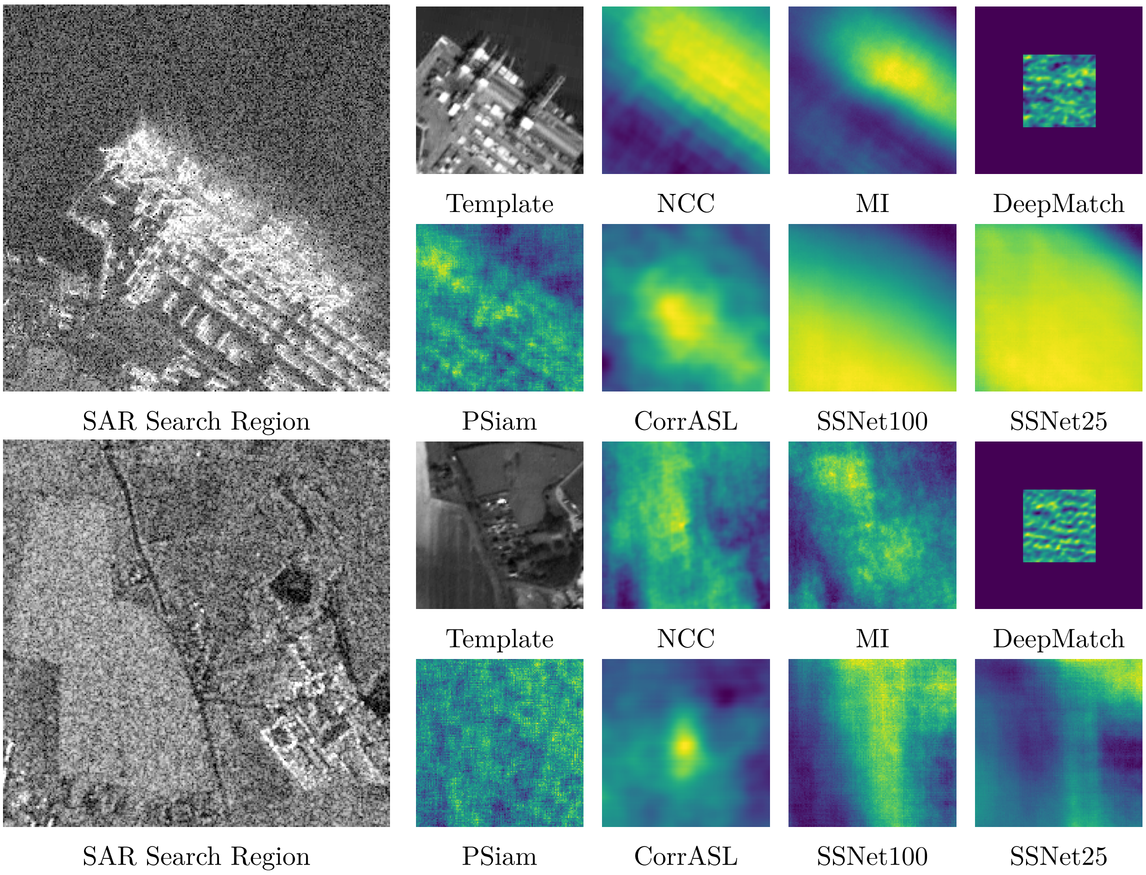


Figure 5.4: Exemplary correspondence heatmaps generated by the various approaches when matching the optical template within the SAR search region, across scenes which have a high likelihood to create ambiguities. The expected point of correspondence is located in the center of the heatmap, and only the valid region of heatmaps are depicted for brevity.

SAR-optical matching performance, and can do so while producing correspondence heatmaps with a predictable and descriptive structure.

Similarly, the heatmaps produced by the SSNet variants encode the ambiguities of the matching process. However, they fail to accurately and uniquely encode the location of correspondence correctly. One interesting observation is that the heatmaps produced by the partially-supervised variant (SSNet25) are very similar in structure to those produced when the network is trained in a fully supervised manner (SSNet100). Thus highlighting the networks ability to learn a consistent matching strategy even when supervised with significantly less data, and further backs up the hypothesis that the networks converged to a sub-optimal solution, rather than the architecture being ill-suited to the SAR-optical matching process.

5.4 Good Regions and Outlier Removal

As introduced in Section 3.3.2, the purpose of the *goodness* network is to improve the matching accuracy of the correspondence network by preselecting regions which have a higher likelihood of being correctly matched. On the other hand, the outlier reduction network, presented in Section 3.3.3, aims to improve the accuracy of the final set of corresponding points by identifying outliers based on the structure of the heatmaps generated by the correspondence network.

Thus to gain a better understanding of their individual and combined impact on the overall matching accuracy, a small scale ablation study was performed. Based on the evaluation of the proposed correspondence network architectures, the CorrASL model was selected as the matching approach to be used in this investigation.

Firstly, both the goodness and outlier reduction networks are trained on datasets derived from the training and validation dataset used to train the CorrASL network. The training and dataset creation details are as discussed in Section 3.3. Next, the goodness network is applied to the previously defined test dataset, and patch pairs which do not exhibit high goodness are removed from the dataset. The optical template patch is then extracted around the identified point of highest goodness, and the SAR-optical pair are then matched using the CorrASL network. Finally, the ORN is then applied to the resultant heatmap, and a threshold of 0.5 is applied to the output to classify the matching process as (un)successful. The effect of each of these processes on matching accuracy is then evaluated throughout the matching pipeline. The results of this investigation are presented in Table 5.3.

From Table 5.3 it can be observed that individually, and jointly, the application of the goodness network and ORN lead to improved matching performance. When both networks are added to the matching pipeline, there is a further improvement in the accuracy, although the average L_2 error increases slightly. This increase is due to the accumulation of errors which occurs when two non-perfect filters are applied sequentially.

Table 5.3: The effects of the goodness and outlier reduction networks on the accuracy of the final set of correspondences, in relation to the baseline approach which performs no pre or post filtering on the test dataset or correspondence set.

	Accuracy		**Precision**
Pipeline	$\leq$ 1 pixel [%]	μ [pixel]	mAP [pixel]
CorrASL	46.9	2.1	2.62
Goodness + CorrASL	59.8	1.62	1.24
CorrASL + ORN	54.1	1.30	1.09
Goodness + CorrASL + ORN	65.2	1.71	1.13

5.5 Large-Scale Scene Matching

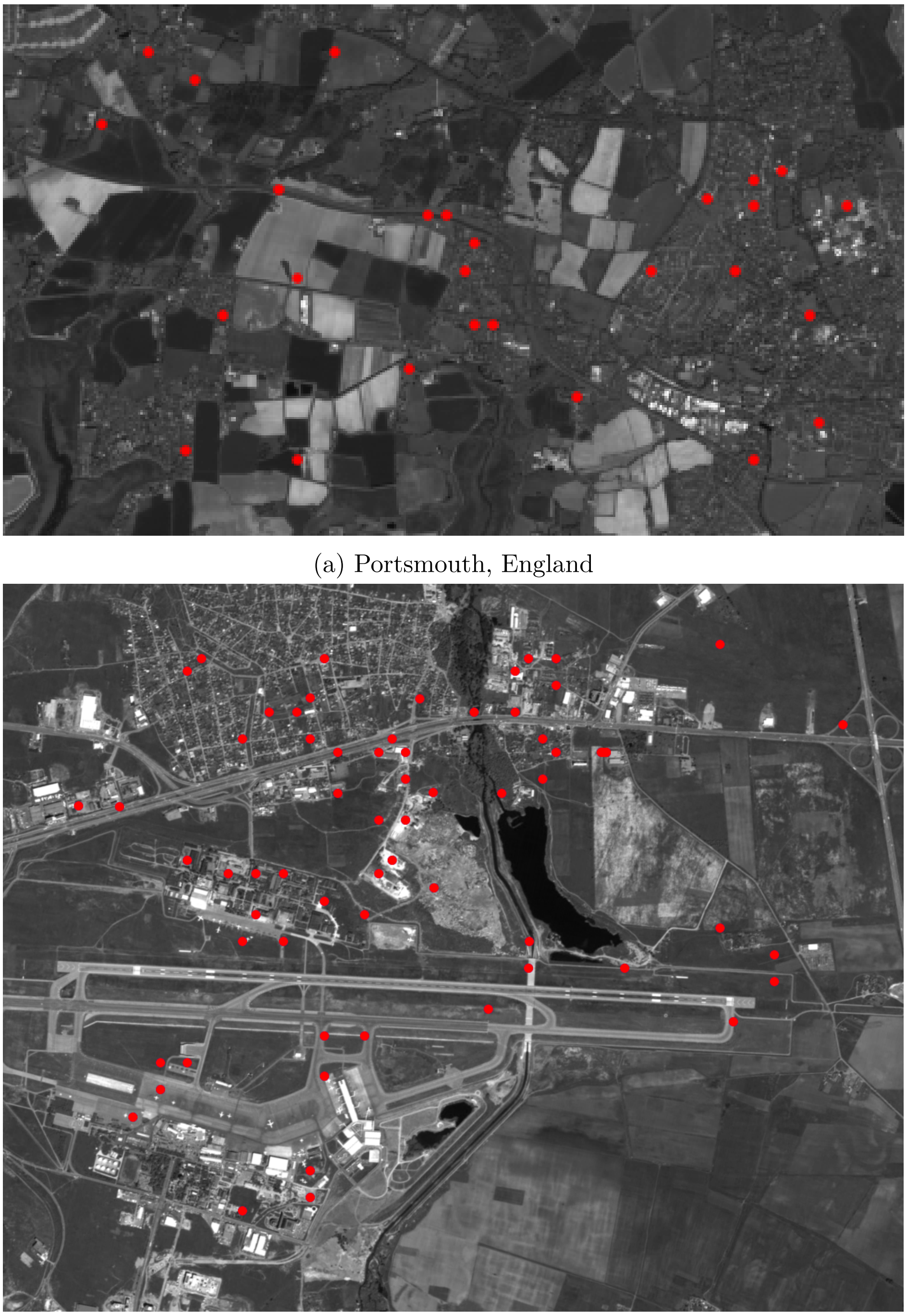
To evaluate the performance and applicability of the SAR-optical matching framework, proposed in Section 3.3.4, under realistic conditions, the correspondence network used in the framework needs to be selected. Based on the previous investigations, the CorrASL network (see Section 3.1.3) was selected to fulfil this role due to its significantly higher matching performance and computational efficiency.

Although the individual components of the framework have been evaluated, the experimental conditions under which these evaluations took place are not comparable to the conditions present in real-world matching scenarios. Thus to gain an insight into the suitability of the proposed framework for identifying correspondences under real-world conditions, it is applied to the problem of determining corresponding points between large-scale test scenes which have not undergone manual co-registration.

To this end two large-scale test scenes were extracted from the areas surrounding the cities of Portsmouth, England and Sofia, Bulgaria, both of which belong to the selection of test cities in the Urban Atlas dataset. The test scenes have a spatial extent of approximately 0.8 × 1.8km, and 4.0 × 5.0km, respectively. The optical image of each scene, with the final set of correspondences overlaid, is depicted in Figure 5.5.

From Figure 5.5, it can be seen that while the proposed matching framework does not produce a large set of correspondences, the resultant set is well spatially distributed. The spatial diversity of the identified correspondences is an important property for a number of data fusion tasks, particularly in applications where the correspondences are to be used as tie-points.

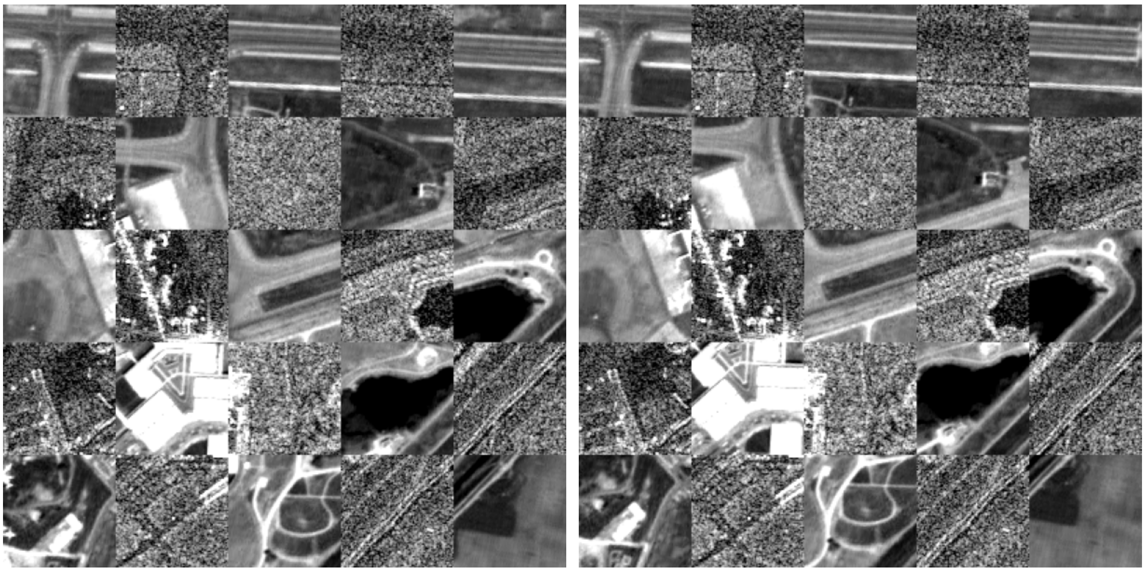
To further examine the suitability of the final correspondence set for use in SAR-optical data fusion endeavours, a qualitative evaluation of the accuracy of the points is performed within the frame of SAR-optical co-registration. As the test scenes have not been manually co-registered, the derived correspondences are used to improve the geo-referencing between the SAR and optical imagery. This is achieved by computing the mean shift between the corresponding SAR and optical point sets. The resultant mean-shift is then applied to the optical scene in order to align it with the SAR image. For the Portsmouth scene the mean-shift (x, y) was found to be $(11.03, -12.74)$ pixels



(a) Portsmouth, England

(b) Sofia, Bulgaria

FIGURE 5.5: The final set of identified correspondences overlaid, in red, on the corresponding optical image of (a) Portsmouth, England with a spatial extent of 0.8 × 1.8km and (b) Sofia, Bulgaria with an extent of 4.0 × 5.0km. The final correspondence set size for (a) is 27 points, and for (b) is 68 points.



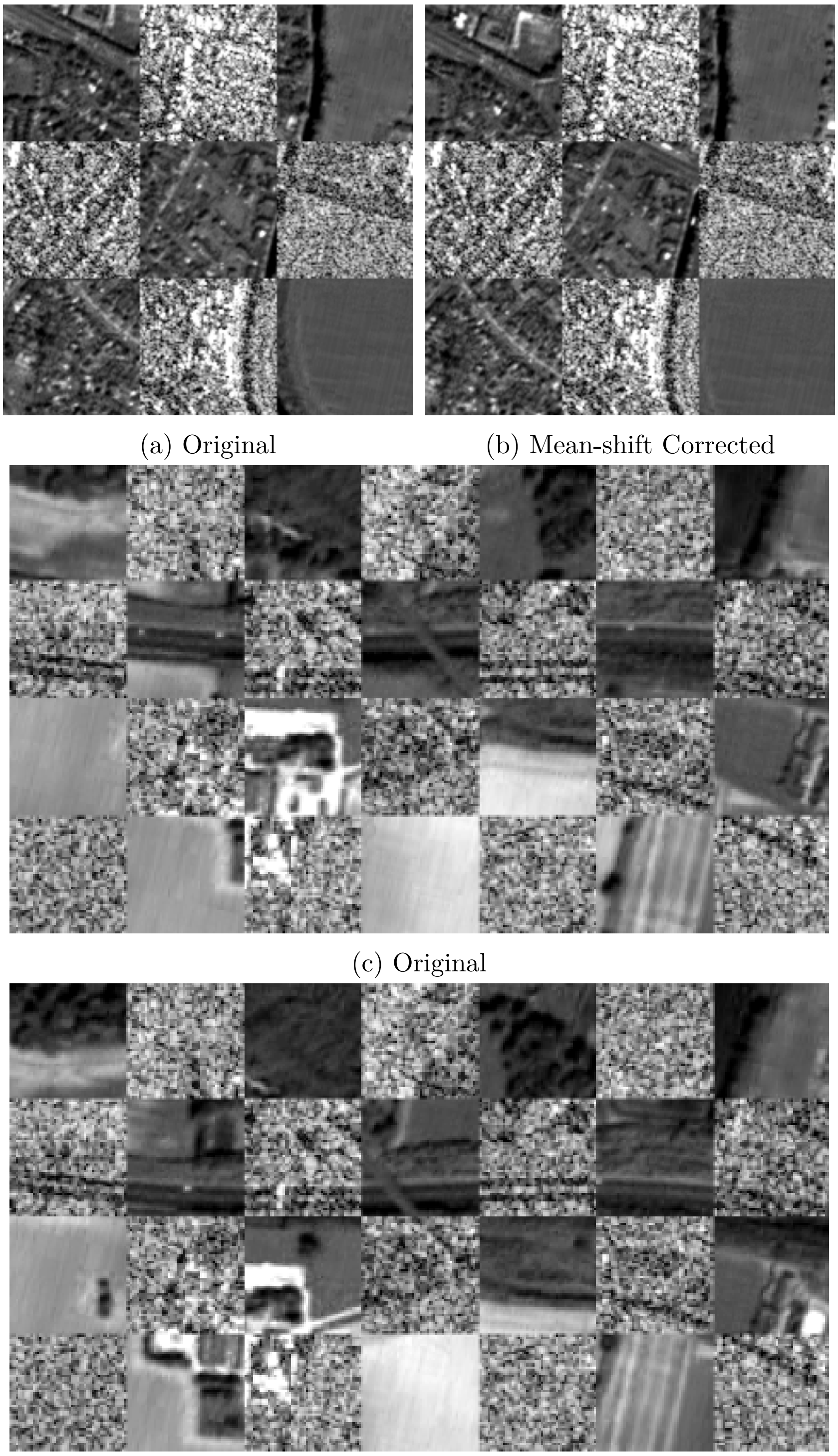
(a) Original

(b) Mean-shift Corrected

Figure 5.6: Checkerboard overlays comparing the alignment of a TerraSAR-X image to the original (non-coregistered), and mean-shifted optical PRISM imagery for two subsets of the Sofia, Bulgaria test scene. The original imagery is depicted in (a) and (c), while the mean-shift, correct imagery is shown in (b) and (d). All images have a pixel spacing of 2.5 meters.

with a standard deviation of $(1.99, 2.20)$ pixels. Similarly, for the Sofia test scene the mean-shift was determined to be $(8.48, 9.12)$ pixels with a standard deviation of $(1.74, 3.01)$ pixels. Checkerboard overlays of sub-regions within each of the test scenes are depicted in Figure 5.6 and Figure 5.7 for Sofia and Portsmouth respectively.

Figure 5.6 and Figure 5.7 highlight the accuracy and utility of the proposed matching framework in being able to identify spatially diverse correspondences which are sufficiently accurate to enable co-registration of high-resolution SAR and optical imagery. Although the co-registration method employed in this evaluation is simplistic, the results still lead to a noticeable improvement in co-registration of the imagery. Furthermore, more advanced techniques such as using the correspondences as GCPs to correct the overall optical sensor model, as described by Müller et al. (2012), could lead to further improved co-registration accuracies; however, this investigation is beyond the scope of this thesis.



(a) Original

(b) Mean-shift Corrected

(c) Original

(d) Mean-shift Corrected

Figure 5.7: Checkerboard overlays comparing the alignment of a TerraSAR-X image to the original (non-coregistered), and mean-shifted optical PRISM imagery for two subsets of the Portsmouth, England test scene. The original imagery is depicted in (a) and (c), while the mean-shift, correct imagery is shown in (b) and (d). All images have a pixel spacing of 2.5 meters.

6. Conclusion and Outlook

6.1 Summary and Conclusion

This dissertation has investigated the applicability of deep learning as a toolkit for formulating solutions and sub-components in the challenging task of creating a fully automatic SAR-optical matching pipeline. In this process, several sub-objectives were defined and investigated to address the various sub-problems encountered within the larger frame of SAR-optical image matching.

As an initial contribution, two large-scale datasets containing corresponding pairs of SAR-optical imagery were created. These datasets were designed and validated to be suitable for training and evaluating deep learning-based SAR-optical image matching architectures, although they are applicable to a wide range of SAR-optical data fusion endeavours. The first dataset is a global-scale medium resolution SEN1-2 dataset, while the second dataset is based on the Urban Atlas dataset which contains high-resolution SAR and optical imagery from 23 cities around Europe.

Centred around this high-resolution data, numerous methodological contributions were summarized within this thesis. Firstly, two supervised deep learning-based matching architectures were proposed. The first of which was a pseudo-siamese architecture which was inspired by the seminal deep matching work of Mou et al. (2017). Although the experimental evaluation of this approach showed promising results when framing matching under the assumption of an ideal cross-domain feature detector, later investigations conducted under more realistic operating conditions highlighted some shortfalls of relying solely on high-level features and having a fixed size fusion network. Based on these lessons, a second architecture (CorrASL) was proposed, centred on the concept of convolutional hypercolumns. These hypercolumns capture feature representations at different scales and are matchable via a standard cross-correlation operator. Through experimental evaluation, this architecture was found to significantly outperform existing SAR-optical matching methodologies in terms of accuracy and precision. Furthermore, the correspondence heatmaps produced by the correlation operator were found to encode the structure of the matching result, as hypothesized and later validated, in such a way that they could be used to identify inaccurate matching results.

Based on the premise that the ability of deep matching networks to generalize is largely dependant on the diversity of the data used to train them, it was hypothesized that matching data from different sensors than those in the dataset would require new datasets or model fine-tuning. Under this hypothesis, it was seen that many formulations of SAR-optical matching can still be considered small data problems. Thus two deep learning-based methodologies were proposed which addressed different challenges of matching SAR and optical imagery under small data constraints. On the one hand, a generative adversarial network was used to create artificial hard negative samples which were in turn used to augment the training of the pseudo-siamese network. This strategy was shown to improve the discriminability of the trained network without requiring any actual additional training data.

On the other hand, a semi-supervised deep matching architecture was proposed to exploit the masses and diversity of unlabelled training data available in Earth observation data archives. An unsupervised autoencoder architecture was used to learn descriptive modality-specific feature latent spaces, which were aligned using an adversarial loss and a small number of labelled training samples. During the evaluation, it was found that even under low levels of supervision (25% labelled data) the network can learn feature representations which are similar to those learned under full supervision. However, the features encoded in the latent space were found not to be descriptive enough to enable accurate matching. One possible reason for this is due to the non-complementary nature of the losses used in the supervised and unsupervised training iterations, thus causing the network to converge to a solution which is not particularly well suited towards either task. However, the preliminary results are promising and warrant further research and investigation.

Much of the research contained in this thesis and the existing literature focuses on developing solutions to the correspondence problem. However, determining the point of correspondence between two images does not constitute the whole image matching pipeline. Thus in order to address the main objective of this dissertation, to create a fully automatic deep learning-based SAR-optical matching pipeline, mechanisms for cross-domain feature detection and outlier removal were developed. Firstly, the feature detection sub-task was reframed as a region proposal problem, whereby the aim was to detect regions in each modality which has a high likelihood of being salient and visible in the other modality. To this end, the *goodness* was proposed to identify these regions which could then be used as candidate input patches for the correspondence network. This approach of feature detection led to an increase in the accuracy and precision of the final set of identified correspondences over the accuracy achieved when using feature points extracted from a single modality. Secondly, an outlier detection network was developed to learn to identify unsuccessful correspondences based on the structure of the heatmaps produced by the correspondence network. This approach to outlier detection was then experimentally validated and found to be effective at identifying outliers without the need for explicitly modelling the feature point transformations across the images. However, more research is required to reduce the number of successful matches which are discarded by the ORN, and to improve the density of feature points proposed by the *goodness* network.

Finally, the goodness network, multi-scale correspondence network and outlier reduction network were linked together to create a fully automatic SAR-optical matching pipeline. This pipeline was then evaluated in an end-to-end manner by determining correspondences between poorly geo-referenced high-resolution SAR and optical imagery across various scenes. When using the resultant set of correspondences to better align the images, the overall geo-referencing error was substantially reduced. Thus, achieving the main objective set out in this thesis to develop a novel, fully automatic deep learning-based SAR-optical matching pipeline capable of matching high-resolution SAR and optical imagery across a diverse range of scenes.

Despite the significant advancements made within the frame of this thesis, the robust and large-scale matching of high-resolution SAR and optical imagery remains an open problem and thus will remain an active area of research for many years to come. However, this thesis has provided a strong case for the continued use of deep learning as the go-to methodological framework for the continued pursuit and development of a generalizable and globally applicable SAR-optical matching pipeline.

6.2 Open Problems

Although the matching accuracy and number of detectable correspondences remain limited by the unfavourable conditions for joint scene visibility between high-resolution SAR and optical imagery, there is still significant room for improvement over what is achievable by the proposed SAR-optical matching framework. Thus, numerous avenues for future investigation remain open, the most immediate of which, as seen by the author, are:

- The extension of the goodness network to a full resolution region proposal network such that more candidate patches can be extracted which in turn should lead to a higher number of detected correspondences.
- The further exploration of semi-supervised learning, and more specifically a robust formulation of the SAR-optical matching problem as a semi-supervised learning task. This will allow for the exploitation of the vast amounts of existing, unlabelled Earth observation data.
- The continued research and development of multi-scale CNN architectures for determining SAR-optical correspondences. The use of multi-scale hypercolumns, proposed in this thesis, significantly improved the matching performance obtainable by deep matching architectures and thus warrants further investigation.
- The formulation of the entire SAR-optical matching pipeline as an end-to-end trainable network. Based on recent trends and results in conventional deep matching literature, end-to-end trainable matching pipelines further remove human bias from the various sub-tasks and thus allow the network to learn stronger representations and formulations for matching.
- The inclusion of prior scene and sensor knowledge into the matching pipeline. As information about the sensor state is known, this information could theoretically be included into the matching process to resolve ambiguities caused by the vastly different geometries of the sensors.

6.3 Outlook

Given the rise of the New Space era, the number of SAR and optical remote sensing sensors is growing rapidly, with many companies striving to create large clusters of high-resolution Earth observation satellites with high frequency revisit times. This increase in the accessibility and availability of SAR and optical remote sensing data has become a driving factor behind the need to develop algorithms and mechanisms to extract valuable insights from this data in an automated and efficient manner.

Furthermore, as the information obtainable from these modalities is highly complementary, the increase in the availability of data has further increased the relevance and importance of SAR-optical data fusion research. This is nowhere better seen than in the fact that the two largest remote sensing-based research competitions, namely, the IEEE Data Fusion Contest (Yokoya et al., 2020) and the SpaceNet Challenges (Shermeyer et al., 2020), both provided large-scale SAR-optical datasets for use in their respective 2020 competitions.

With these trends likely to continue into the foreseeable future, the need for efficient, high-resolution SAR-optical matching methodologies, required to enable data fusion endeavours, will continue to grow. To this end, the research presented within the frame of this thesis has shown the great potential for the application of modern deep learning techniques to the challenging task of matching high-resolution SAR and optical imagery.

It will therefore continue to be of utmost importance to strengthen the connection between the fields of deep learning and remote sensing. The deep learning community need to better understand the complexity of working with vastly heterogenous data such that they can develop models and methodologies which are better suited to non-optical data sources and scarce training data. While, the remote sensing community need to combine their strong domain expertise with deep learning techniques to develop modern methodologies for the extraction of valuable insights from the masses of data being produced, thus ushering in the New era of global Earth observation.

Bibliography

Auer, S., Schmitt, M., & Reinartz, P. (2017). Automatic alignment of high resolution optical and SAR images for urban areas. *Proc. International Geoscience and Remote Sensing Symposium Conference*, 5466–5469.

Bagheri, H., Schmitt, M., d'Angelo, P., & Zhu, X. X. (2018). A framework for SAR-optical stereogrammetry over urban areas. *ISPRS Journal of Photogrammetry and Remote Sensing*, *146*, 389–408.

Balntas, V., Johns, E., Tang, L., & Mikolajczyk, K. (2016). PN-Net: Conjoined triple deep network for learning local image descriptors. *arXiv:1601.05030.*

Breit, H., Fritz, T., Balss, U., Lachaise, M., Niedermeier, A., & Vonavka, M. (2009). TerraSAR-X SAR processing and products. *IEEE Transactions on Geoscience and Remote Sensing*, *48*(2), 727–740.

Burger, W., & Burge, M. J. (2016). Image matching and registration. *Digital image processing* (pp. 565–585). Springer.

Bürgmann, T., Koppe, W., & Schmitt, M. (2019). Matching of TerraSAR-X derived ground control points to optical image patches using deep learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, *158*, 241–248.

Butler, D. (2014). Earth observation enters next phase. *Nature*, *508*(7495), 160–161.

Chapelle, O., Schölkopf, B., & Zien, A. (2006). Semi-supervised learning. *Ieee transactions on neural networks* (pp. 542–542).

Chen, H., Arora, M. K., & Varshney, P. K. (2003). Mutual information-based image registration for remote sensing data. *International Journal of Remote Sensing*, *24*(18), 3701–3706.

Cheng, H., Zheng, S., Yu, Q., Tian, J., & Liu, J. (2004). Matching of sar images and optical images based on edge feature extracted via svm. *Proc. International Conference on Signal Processing*, *2*, 930–933.

Chopra, S., Hadsell, R., & LeCun, Y. (2005). Learning a similarity metric discriminatively, with application to face verification. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, *1*, 539–546.

Citak, E., & Bilgin, G. (2019). Visual saliency aided SAR and optical image matching. *Proc. Innovations in Intelligent Systems and Applications Conference*, 1–5.

Cumming, I. G., & Wong, F. H. (2005). Digital processing of synthetic aperture radar data. *Artech house*, *1*(3).

Curlander, J. (1982). Geometric and rediametric distortion in spaceborne SAR imagery. *Proc. NASA Workshop on Registration and Rectification*, 163–197.

Dai, Z., Yang, Z., Yang, F., Cohen, W. W., & Salakhutdinov, R. R. (2017). Good semi-supervised learning that requires a bad gan. *Proc. Advances in Neural Information Processing Systems*, 6510–6520.

Dare, P., & Dowman, I. (2000). A new approach to automatic feature based registration of SAR and SPOT images. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *33*(B2; PART 2), 125–130.

Dellinger, F., Delon, J., Gousseau, Y., Michel, J., & Tupin, F. (2015). SAR-SIFT: A SIFT-like algorithm for SAR images. *IEEE Transactions on Geoscience and Remote Sensing*, *53*(1), 453–466.

Denis, G., Claverie, A., Pasco, X., Darnis, J.-P., de Maupeou, B., Lafaye, M., & Morel, E. (2017). Towards disruptions in Earth observation? New Earth observation systems and markets evolution: Possible scenarios and impacts. *Acta Astronautica*, *137*, 415–433.

DeTone, D., Malisiewicz, T., & Rabinovich, A. (2018). Superpoint: Self-supervised interest point detection and description. *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshop*, 224–236.

Dusmanu, M., Rocco, I., Pajdla, T., Pollefeys, M., Sivic, J., Torii, A., & Sattler, T. (2019). D2-Net: A trainable CNN for joint description and detection of local features. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 8092–8101.

Eineder, M., Minet, C., Steigenberger, P., Cong, X., & Fritz, T. (2010). Imaging geodesy – Toward centimeter-level ranging accuracy with TerraSAR-X. *IEEE Transactions on Geoscience and Remote Sensing*, *49*(2), 661–671.

Fan, B., Huo, C., Pan, C., & Kong, Q. (2012). Registration of optical and sar satellite images by exploring the spatial relationship of the improved sift. *IEEE Geoscience and Remote Sensing Letters*, *10*(4), 657–661.

Fan, J., Wu, Y., Wang, F., Zhang, Q., Liao, G., & Li, M. (2014). Sar image registration using phase congruency and nonlinear diffusion-based sift. *IEEE Geoscience and Remote Sensing Letters*, *12*(3), 562–566.

Felzenszwalb, P., McAllester, D., & Ramanan, D. (2008). A discriminatively trained, multiscale, deformable part model. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 1–8.

Fischer, P., Schuegraf, P., Merkle, N., & Storch, T. (2018). An evolutionary algorithm for fast intensity based image matching between optical and SAR satellite imagery. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *4*(3).

Fischer, P., Dosovitskiy, A., & Brox, T. (2014). Descriptor matching with convolutional neural networks: A comparison to SIFT. *arXiv:1405.5769.*

Fischler, M. A., & Bolles, R. C. (1981). Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, *24*(6), 381–395.

Gamba, P. (2014). Image and data fusion in remote sensing of urban areas: Status issues and research trends. *International Journal of Image and Data Fusion*, *5*(1), 2–12.

Ghaffary, B. K. (1986). A review of image matching techniques. In F. J. Corbett, H. J. Siegel, & M. J. Duff (Eds.), *Proc. architectures and algorithms for digital image processing* (pp. 164–172). SPIE.

Girard, C. M., & Girard, M.-C. (2003). *Processing of remote sensing data*. CRC Press.

Gong, M., Zhao, S., Jiao, L., Tian, D., & Wang, S. (2014). A novel coarse-to-fine scheme for automatic image registration based on SIFT and mutual information. *IEEE Transactions on Geoscience and Remote Sensing*, *52*(7), 4328–4338.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*.

Gruen, A. (2012). Development and status of image matching in photogrammetry. *The Photogrammetric Record*, *27*(137), 36–57.

Hadsell, R., Chopra, S., & LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, *2*, 1735–1742.

Han, X., Leung, T., Jia, Y., Sukthankar, R., & Berg, A. C. (2015). MatchNet: Unifying feature and metric learning for patch-based matching. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 3279–3286.

Hariharan, B., Arbelaez, P., Girshick, R., & Malik, J. (2015). Hypercolumns for object segmentation and fine-grained localization. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 447–456.

Harris, C., & Stephens, M. (1988). A combined corner and edge detector. *Proc. Alvey Vision Conference*, *15*(50), 10–5244.

Hoffmann, S., Brust, C.-A., Shadaydeh, M., & Denzler, J. (2019). Registration of high resolution SAR and optical satellite imagery using fully convolutional networks. *Proc. International Geoscience and Remote Sensing Symposium Conference*, 5152–5155.

Horn, B. K., & Schunck, B. G. (1981). Determining optical flow. *Techniques and Applications of Image Understanding*, *281*, 319–331.

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2018). Progressive growing of GANs for improved quality, stability, and variation. *Proc. International Conference of Learning Representations*.

Laguna, A. B., Riba, E., Ponsa, D., & Mikolajczyk, K. (2019). Key.net: Keypoint detection by handcrafted and learned CNN filters. *Proc. IEEE/CVF International Conference on Computer Vision*, 5835–5843.

Larsen, A. B. L., Sønderby, S. K., Larochelle, H., & Winther, O. (2016). Autoencoding beyond pixels using a learned similarity metric. *Proc. International Conference on Machine Learning*, 1558–1566.

Leng, C., Zhang, H., Li, B., Cai, G., Pei, Z., & He, L. (2019). Local feature descriptor for image matching: A survey. *IEEE Access*, *7*, 6424–6434.

Li, H., Manjunath, B., & Mitra, S. K. (1995). A contour-based approach to multisensor image registration. *IEEE Transactions on Image Processing*, *4*(3), 320–334.

Li, J., Hu, Q., & Ai, M. (2020). RIFT: Multi-modal image matching based on radiation-variation insensitive feature transform. *IEEE Transactions on Image Processing*, *29*, 3296–3310.

Li, Y., Wang, S., Tian, Q., & Ding, X. (2015). A survey of recent advances in visual feature detection. *Neurocomputing*, *149*, 736–751.

Lillesand, T., Kiefer, R. W., & Chipman, J. (2015). *Remote sensing and image interpretation*. John Wiley & Sons.

Liu, W., Shen, X., Wang, C., Zhang, Z., Wen, C., & Li, J. (2018). H-Net: Neural network for cross-domain image patch matching. *Proc. International Joint Conference on Artificial Intelligence*, 856–863.

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

Ma, W., Wen, Z., Wu, Y., Jiao, L., Gong, M., Zheng, Y., & Liu, L. (2017). Remote sensing image registration with modified SIFT and enhanced feature matching. *IEEE Geoscience and Remote Sensing Letters*, *14*(1), 3–7.

Ma, W., Zhang, J., Wu, Y., Jiao, L., Zhu, H., & Zhao, W. (2019). A novel two-step registration method for remote sensing images based on deep and local features. *IEEE Transactions on Geoscience and Remote Sensing*, *57*(7), 4834–4843.

Makhzani, A., Shlens, J., Jaitly, N., & Goodfellow, I. (2016). Adversarial autoencoders. *Proc. International Conference of Learning Representations*.

Merkle, N., Auer, S., Müller, R., & Reinartz, P. (2017). Exploring the potential of conditional adversarial networks for optical and SAR image matching. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *11*(6), 1811–1820.

Merkle, N., Luo, W., Auer, S., Müller, R., & Urtasun, R. (2017). Exploiting deep matching and SAR data for the Geo-localization accuracy improvement of optical satellite images. *Remote Sensing*, *9*(6), 586.

Merkle, N. M. (2018). *Geo-localization refinement of optical satellite images by embedding synthetic aperture radar data in novel deep learning frameworks* (Doctoral dissertation). University Osnabrück.

Mishchuk, A., Mishkin, D., Radenoviundefined, F., & Matas, J. (2017). Working hard to know your neighbor's margins: Local descriptor learning loss. *Proc. Advances in Neural Information Processing Systems*, 4829–4840.

Mou, L., Schmitt, M., Wang, Y., & Zhu, X. X. (2017). A CNN for the identification of corresponding patches in SAR and optical imagery of urban scenes. *Proc. Joint Urban Remote Sensing Event*, 1–4.

Mukherjee, T., Yamada, M., & Hospedales, T. M. (2017). Deep matching autoencoders. *Computing Research Repository.*

Müller, R., Krauß, T., Schneider, M., & Reinartz, P. (2012). Automated georeferencing of optical satellite data with integrated sensor model improvement. *Photogrammetric Engineering & Remote Sensing*, *78*(1), 61–74.

O'Mahony, N., Campbell, S., Carvalho, A., Harapanahalli, S., Hernandez, G. V., Krpalkova, L., Riordan, D., & Walsh, J. (2019). Deep learning vs. traditional computer vision. *Proc. Science and Information Conference*, 128–144.

OpenStreetMap contributors. (2017). Planet dump retrieved from https://planet.osm.org.

Orth, P. (2018). The geometry of spaceborne synthetic aperture radar. *arXiv:1808.06549*.

Prasad, S., Bruce, L. M., & Chanussot, J. (2011). Optical remote sensing. *Advances in Signal Processing and Exploitation Techniques.*

Qiu, C., Schmitt, M., & Zhu, X. X. (2018). Towards automatic SAR-optical stereogrammetry over urban areas using very high resolution imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, *138*, 218–231.

Salahat, E., & Qasaimeh, M. (2017). Recent advances in features extraction and description algorithms: A comprehensive survey. *Proc. IEEE international Conference on Industrial Technology*, 1059–1063.

Schmitt, M., Tupin, F., & Zhu, X. X. (2017). Fusion of SAR and optical remote sensing data – challenges and recent trends. *Proc. International Geoscience and Remote Sensing Symposium Conference*, 5458–5461.

Schmitt, M., & Zhu, X. X. (2016). Data fusion and remote sensing: An ever-growing relationship. *4*(4), 6–23.

Schneider, M., Müller, R., Krauss, T., Reinartz, P., Hörsch, B., & Schmuck, S. (2010). Urban Atlas – DLR processing chain for orthorectification of PRISM and AVNIR-2 images and TerraSAR-X as possible GCP source. *Internet Proceedings*, 1–6.

Schönberger, J. L., Hardmeier, H., Sattler, T., & Pollefeys, M. (2017). Comparative evaluation of hand-crafted and learned local features. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 1482–1491.

Schulz, K., Boldt, M., & Thiele, A. (2009). CovAmCoh-analysis: A method to improve the interpretation of high resolution repeat pass sar images of urban areas. *Proc. Remote Sensing for Environmental Monitoring, GIS Applications, and Geology*, *7478*, 747–805.

Shapiro, L. S., & Brady, J. M. (1992). Feature-based correspondence: An eigenvector approach. *Image and Vision Computing*, *10*(5), 283–288.

Shermeyer, J., Hogan, D., Brown, J., Van Etten, A., Weir, N., Pacifici, F., Hansch, R., Bastidas, A., Soenen, S., Bacastow, T., et al. (2020). SpaceNet 6: Multi-sensor all weather mapping dataset. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 196–197.

Siddique, M. A., Sarfraz, M. S., Bornemann, D., & Hellwich, O. (2012). Automatic registration of SAR and optical images based on mutual information assisted Monte Carlo. *Proc. International Geoscience and Remote Sensing Symposium Conference*, 1813–1816.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *Proc. International Conference of Learning Representations.*

Simo-Serra, E., Trulls, E., Ferraz, L., Kokkinos, I., Fua, P., & Moreno-Noguer, F. (2015). Discriminative learning of deep convolutional feature point descriptors. *Proc. IEEE/CVF International Conference on Computer Vision*, 118–126.

Steger, C., Ulrich, M., & Wiedemann, C. (2018). *Machine vision algorithms and applications.* John Wiley & Sons.

Sun, C., Shrivastava, A., Singh, S., & Gupta, A. (2017). Revisiting unreasonable effectiveness of data in deep learning era. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 843–852.

Suri, S., & Reinartz, P. (2010). Mutual-information-based registration of TerraSAR-X and Ikonos imagery in urban areas. *IEEE Transactions on Geoscience and Remote Sensing*, *48*(2), 939–949.

Suri, S., Schwind, P., Uhl, J., & Reinartz, P. (2010). Modifications in the SIFT operator for effective SAR image matching. *International Journal of Image and Data Fusion*, *1*(3), 243–256.

Szeliski, R. (2010). *Computer vision: Algorithms and applications.* Springer.

Thenkabail, P. (2018). *Remote sensing handbook - three volume set.* CRC Press.

Tupin, F. (2010). Fusion of optical and SAR images. *Radar remote sensing of urban areas* (pp. 133–159). Springer.

United Nations. (2015). Transforming our world: The 2030 agenda for sustainable development. *General Assembly 70 session.*

Wang, Y., Yu, Q., & Yu, W. (2012). An improved normalized cross correlation algorithm for SAR image registration. *Proc. International Geoscience and Remote Sensing Symposium Conference*, 2086–2089.

Wang, Y., & Zhu, X. X. (2018). The sarptical dataset for joint analysis of sar and optical image in dense urban area. *Proc. International Geoscience and Remote Sensing Symposium Conference*, 6840–6843.

Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). CBAM: Convolutional block attention module. *Proc. European Conference on Computer Vision*, 3–19.

Wu, J., Cui, Z., Sheng, V. S., Zhao, P., Su, D., & Gong, S. (2013). A comparative study of SIFT and its variants. *Journal of the Institute of Measurement Science*, *13*(3), 122–131.

Xiang, Y., Wang, F., & You, H. (2018). OS-SIFT: A robust SIFT-like algorithm for high-resolution optical-to-SAR image registration in suburban areas. *IEEE Transactions on Geoscience and Remote Sensing*, *56*(6), 3078–3090.

Xu, C., Sui, H., Li, H., & Liu, J. (2015). An automatic optical and sar image registration method with iterative level set segmentation and sift. *International Journal of Remote Sensing*, *36*(15), 3997–4017.

Ye, Y., & Shen, L. (2016). HOPC: A novel similarity metric based on geometric structural properties for multi-modal remote sensing image matching. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *3*, 9.

Ye, Y., Shen, L., Hao, M., Wang, J., & Xu, Z. (2017). Robust optical-to-SAR image matching based on shape properties. *IEEE Geoscience and Remote Sensing Letters*, *14*(4), 564–568.

Yi, K. M., Trulls, E., Lepetit, V., & Fua, P. (2016). LIFT: Learned invariant feature transform. *Proc. European Conference on Computer Vision*, 467–483.

Yokoya, N., Ghamisi, P., Haensch, R., & Schmitt, M. (2020). 2020 IEEE GRSS Data Fusion Contest: Global land cover mapping with weak supervision [Technical Committees]. *IEEE Geoscience and Remote Sensing Magazine*, *8*(1), 154–157.

Zagoruyko, S., & Komodakis, N. (2015). Learning to compare image patches via convolutional neural networks. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 4353–4361.

Zhang, J. (2010). Multi-source remote sensing data fusion: Status and trends. *International Journal of Image and Data Fusion*, *1*(1), 5–24.

Zhaohui, Z., Chunhong, P., & Songde, M. (2004). An automatic procedure for sar-optical satellite image registration based on multi-layer feature matching strategy. *Proc. International Symposium on Intelligent Signal Processing and Communication Systems*, 574–580.

Zhu, R., Yu, D., Ji, S., & Lu, M. (2019). Matching rgb and infrared remote sensing images with densely-connected convolutional neural networks. *Remote Sensing*, *11*(23).

Zitová, B., & Flusser, J. (2003). Image registration methods: A survey. *Image and Vision Computing*, *21*(11), 977–1000.

Acknowledgements

Over the last years many people have provided guidance and inspiration in both a professional and personal capacity, without them the doctoral process and completion of this thesis would not have been possible.

Firstly, I would like to thank my supervisor Dr. Michael Schmitt, whose enthusiasm towards this topic, scientific guidance and consistent positive outlook kept me pushing forwards even on the hardest of days. Without his unwavering support for both myself and my work, and compassionate nature, this thesis would not have been realised.

Secondly I am grateful to Prof. Xiaoxiang Zhu for providing me with the opportunity to work under her guidance as a part of the SiPEO research group, and broader DLR community. Furthermore, for always ensuring that we had the resources we needed succeed at the cutting edge of Earth Observation data science.

Additionally, I would like to thank Prof. Florence Tupin for her role in acting as an external examiner of this thesis, and Prof. Martin Werner for taking the time to act as the Chair of my doctoral defence. I am highly appreciative of both of your time and availability in assessing the work done, and presented as part of this thesis.

Similarly, I would like to recognise and thank all the co-authors and collaborators I had the opportunity to work along side over the course of my doctoral journey. Working with each of you was a pleasure which brought new insights and experiences that have helped me grow as a researcher and scientist. I would also like to extend a special thank you to Prof. Devis Tuia, who provided me with the invaluable opportunity to visit and work along side his research group at Wageningen University and Research, this was a deeply rewarding experience. Furthermore, to Prof. Peter Reinartz for providing me with the invaluable data, in the form of the Urban Atlas dataset, required for the successful completion of this thesis.

Over the years I was also fortunate enough to work along side a wonderful group of people, who are not only excellent researchers in their own sense, but also kind individuals who were open to sharing their thoughts about science, ethics, travel or anything in-between. I am thankful to all of you for the additions you made to my world view, and for your openness in sharing the human experience. More specifically, I would like to acknowledge my office-mates, Chunping Qiu, Lukas Liebel, Marc Russwurm, Sandra Aigner and part-time resident Tobias Koch. I could not have asked for better people to share an office and the doctoral journey with.

Last, but most certainly not least, I would like to thank those closest to me. To my partner, Janien, thank you for your continuous support and love, you know most the ups and downs of this journey and never faltered to be there for me when I have needed you. To my parents, Debby and Justin, for being a constant source of support in my life, and for all the sacrifices you have made over the years to ensure we always had what we needed to achieve our goals. My sister, Ashleigh for always being supportive of me and providing me with a sounding board for all of life's decisions. Finally, to my closest friends, both old and new, you have been instrumental in this journey and in my personal development - thank you for always being there for me and for pushing me to constantly better myself in all aspects of life.