UNIVERSITÄT OSNABRÜCK

# Geo-localization Refinement of Optical Satellite Images by Embedding Synthetic Aperture Radar Data in Novel Deep Learning Frameworks

Dissertation

Zur Erlangung
des Doktorgrades der Naturwissenschaften (Dr. rer. nat.)
des Fachbereichs Mathematik/Informatik
der Universität Osnabrück

Vorgelegt von

*Nina Marie Merkle*

Prüfer der Dissertation:

Prof. Dr. Peter Reinartz, Universität Osnabrück
Prof. Dr. habil. Stefan Hinz, Karlsruher Institut für Technologie

Osnabrück, 2018

# ABSTRACT

Every year, the number of applications relying on information extracted from high-resolution satellite imagery increases. In particular, the combined use of different data sources is rising steadily, for example to create high-resolution maps, to detect changes over time or to conduct image classification. In order to correctly fuse information from multiple data sources, the utilized images have to be precisely geometrically registered and have to exhibit a high absolute geo-localization accuracy. Due to the image acquisition process, optical satellite images commonly have an absolute geo-localization accuracy in the order of meters or tens of meters only. On the other hand, images captured by the high-resolution synthetic aperture radar satellite TerraSAR-X can achieve an absolute geo-localization accuracy within a few decimeters and therefore represent a reliable source for absolute geo-localization accuracy improvement of optical data. The main objective of this thesis is to address the challenge of image matching between high resolution optical and synthetic aperture radar (SAR) satellite imagery in order to improve the absolute geo-localization accuracy of the optical images.

The different imaging properties of optical and SAR data pose a substantial challenge for a precise and accurate image matching, in particular for the handcrafted feature extraction stage common for traditional optical and SAR image matching methods. Therefore, a concept is required which is carefully tailored to the characteristics of optical and SAR imagery and is able to learn the identification and extraction of relevant features. Inspired by recent breakthroughs in the training of neural networks through deep learning techniques and the subsequent developments for automatic feature extraction and matching methods of single sensor images, two novel optical and SAR image matching methods are developed. Both methods pursue the goal of generating accurate and precise tie points by matching optical and SAR image patches. The foundation of these frameworks is a semi-automatic matching area selection method creating an optimal initialization for the matching approaches, by limiting the geometric differences of optical and SAR image pairs. The idea of the first approach is to eliminate the radiometric differences between the images trough an image-to-image translation with the help of generative adversarial networks and to realize the subsequent image matching through traditional algorithms. The second approach is an end-to-end method in which a Siamese neural network learns to automatically create tie points between image pairs through a targeted training. The geo-localization accuracy improvement of optical images is ultimately achieved by adjusting the corresponding optical sensor model parameters through the generated set of tie points.

The quality of the proposed methods is verified using an independent set of optical and SAR image pairs spread over Europe. Thereby, the focus is set on a quantitative and qualitative evaluation of the two tie point generation methods and their ability to generate reliable and accurate tie points. The results prove the potential of the developed concepts, but also reveal weaknesses such as the limited number of training and test data acquired by only one combination of optical and SAR sensor systems. Overall, the tie points generated by both deep learning-based concepts enable an absolute geo-localization improvement of optical images, outperforming state-of-the-art methods.

## ZUSAMMENFASSUNG

Aufgrund der steigenden Zahl an Anwendungen, die sich auf hochauflösenden Satellitendaten stützen, gewinnt auch die kombinierte Nutzung mehrerer Datenquellen immer mehr an Bedeutung. Beispielsweise können mit der Fusion mehrerer Datenquellen hochauflösende Karten erstellt oder Bildklassifikationen durchgeführt werden. Voraussetzung für eine korrekte Fusion von Informationen aus mehreren Datenquellen ist eine präzise Registrierung aller Bilder verbunden mit einer hohen Lagegenauigkeit der einzelnen Datenquellen. Aufgrund ihres Bildaufnahme-Verfahrens weisen optische Satellitenbilder in der Regel jedoch eine geringere Lagegenauigkeit im zweistelligen Meterbereich auf. Dagegen können Bilder des hochauflösenden Radarsatelliten TerraSAR-X eine absolute Lagegenauigkeit innerhalb weniger Dezimeter erreichen und stellen somit eine zuverlässige Quelle zur Verbesserung der absoluten Lagegenauigkeit dar. Das Hauptziel dieser Arbeit ist die Verbesserung des Bild-Matchings zwischen hochauflösenden optischen und Synthetic Aperture Radar (SAR) Satellitenbildern, um damit die absolute Lagegenauigkeit von optischen Bildern zu verbessern.

Die unterschiedlichen Abbildungseigenschaften von optischen und SAR-Bildern stellen für die gängigen Methoden zur Bildregistrierung eine große Herausforderung dar, insbesondere für den Schritt der Merkmalsextraktion. Aus diesem Grund ist ein Verfahren erforderlich, das auf die speziellen Eigenschaften von optischen und SAR-Bildern zugeschnitten ist und die Identifizierung und Extraktion von relevanten Merkmalen erlernen kann. Inspiriert durch die jüngsten Durchbrüche beim Training neuronaler Netze mit Hilfe von Deep Learning Techniken und den resultierenden Entwicklungen bei der automatischen Bild-Merkmalsextraktion und beim Bild-Matching, werden in dieser Arbeit zwei neuartige Verfahren für die Erzeugung von Verknüpfungspunkten zwischen optischen und SAR-Bildern vorgestellt. Das Ziel beider Bild-Matching Verfahren ist die Erzeugung präziser Verknüpfungspunkte zwischen optischen und SAR-Bildpaaren. Für eine optimale Ausgangssituation des Bild-Matchings sorgt ein halbautomatisches Verfahren zur Auswahl der Matching-Gebiete. Hierdurch werden die geometrischen Unterschiede zwischen optischen und SAR-Bildpaaren auf ein Minimum reduziert. Die Idee der ersten Methode besteht darin, die radiometrischen Unterschiede zwischen den Bildpaaren durch Anwendung einer Bild-zu-Bild Transformation durch sogenannte „generative adversarial networks“ zu beseitigen und danach das eigentliche Bild-Matching durch traditionelle Methoden zu realisieren. Die zweite Methode ist ein „end-to-end“ Ansatz, bei dem ein siamesisches neuronales Netzwerk durch ein gezieltes Training lernt, automatisch Verknüpfungspunkte zwischen Bildpaaren zu erzeugen. Schließlich werden mit Hilfe der generierten Punkte die optischen Sensormodellparameter angepasst und somit die Lagegenauigkeiten der optischen Bilder verbessert.

Die Qualität der vorgeschlagenen Verfahren zur Bildregistrierung wird anhand voneinander unabhängiger und über ganz Europa verteilter optischer und SAR-Bildpaaren ausgewertet. Dabei liegt der Fokus auf einer sowohl quantitativen als auch qualitativen Auswertung der beiden Verfahren zur Generierung von Verknüpfungspunkten. Die Ergebnisse belegen das Potenzial der entwickelten Methoden, zeigen aber auch Schwächen wie beispielsweise die begrenzte Anzahl von Trainings- und Testdaten, erstellt aus den Bildern eines optischen bzw. SAR Sensors. Insgesamt ermöglichen die von beiden Verfahren erzeugten Verknüpfungspunkte eine Verbesserung der absoluten Lagegenauigkeit optischer Bilder und sind dabei genauer als State-of-the-Art Methoden.

# Contents

# 1

# Introduction

## Contents

## 1.1 Motivation and Scope

In the last years, the technical advances in sensors systems and the increasing number of national and international space programs led to a strongly growing volume of remote sensing data. At the same time, the increased processing power and the development of new tools and algorithms boosted the applications of these data in terms of efficiency, reliability and robustness. In particular, the joint use of multi-modal data such as optical and synthetic aperture radar (SAR) images contain complementary information on objects on the Earth surface, which enriches the information conveyed by an object for several applications such as the generation of high-resolution maps for autonomous driving, the monitoring and modeling of changes over time for precision farming and urban planning. Despite the latest developments and technological advances, an accurate geo-referencing and co-registration step is still a prerequisite for the successful joint use of images acquired from different data sources. Very often, such steps are difficult to carry out with satisfactory precision in an unsupervised way, due to the intrinsic limitations in the sensors' specific acquisition modes.

The goal of this thesis is performing multi-modal image registration of high-resolution optical and SAR satellite images in order to enhance the absolute geo-localization accuracy of the former. To this end, deep learning techniques are utilized to develop two novel tie point generation methods, which enable an enhanced registration of optical and SAR image pairs and therefore a highly improved orthorectification of optical image data.

## 1.2 Scientific Relevance of the Topic

The collection of complementary information from aligned multi-modal image data enables a more detailed and more robust understanding of an image scene or specific object, and is important for several applications in the fields of medical imaging, computer vision or remote sensing [1–4]. In the particular case of optical and SAR satellites, the images acquired by these sensors exhibit different behavior (see example illustrated in Figure 1.1). These different and often complementary characteristics have been proven to be conducive for diverse applications in the field of remote sensing. More specific, several research studies investigated possibilities of their combined usage for tasks such as earthquake damage assessment of buildings [5], road network extraction [6], land cover classification [7], change detection [8], urban surface model generation [9–12] and stereogrammetric 3D analysis of urban areas [13]. However, optical and SAR satellite images are affected by acquisition related influences leading to local or global distortions, which lower the accuracy of the extracted information affecting in particular the absolute geo-localization accuracy of the optical images, thus hindering the use of optical images in any data fusion application.

To overcome this fusion problem, the absolute geo-localization accuracy of optical images has to be improved beforehand. A common approach is the use of ground control points (GCPs) obtained from tedious in-situ GPS measurements or from very exact maps. The generation of GCPs is time consuming and expensive and therefore only available in a minority of cases. Another possibility is to align optical images to an image with a high absolute geo-localization accuracy. As images captured by the high-resolution SAR satellite TerraSAR-X can reach an absolute geo-localization accuracy within a few decimeters, or

**Figure 1.1:** Illustration of an optical (top) and SAR image (bottom) covering the same area. Both images have a ground sampling distance of 1.25 m.

centimeters for specific targets [14], they represent a reliable source for the geo-localization accuracy improvement of optical images. Over the last years, different research studies investigated the geo-localization accuracy improvement of optical satellite images based on SAR reference data, achieving promising results [15–19]. These works rely on suitable image registration techniques, which are tailored to the problem of optical and SAR images matching. Due to the different acquisition concepts it is difficult to find identical features in both image modalities or reliable similarity measures. More precisely, the sideways-looking acquisition of SAR sensors causes typical geometric distortion effects (layover, foreshortening) and shadowing for 3D objects such as buildings or trees. These effects have a strong influence on the appearance of all objects above ground level. As a consequence, the boundary of an elevated object in a SAR image does not match the object boundary in the optical image, even if the imaging perspective is the same for both sensors. Additionally, the different wavelengths measured by the two sensors lead to different radiometric properties in the optical and SAR images. This is due to the fact that the response of an object depends on the

signal properties (wavelength, polarization), the surface properties (roughness, randomness of local reflectors and reflectance properties) and sensor perspective. The same object may therefore appear with high intensity for one sensor and with low intensity in another. The multiplicative noise in SAR images (speckle) further complicates the human and automatic interpretation of SAR images and, hence, the matching of optical and SAR images. As an example, Figure 1.1 shows the difference between an optical and a high-resolution SAR image (e.g. the different intensity of streets), which are acquired over the same area containing man-made structures and vegetation. As a consequence, a suitable registration approach has to be carefully developed or adapted in order to fulfill the particular characteristics of optical and SAR image matching.

Several methods have been developed over the years to find a solution to the problem of optical and SAR image registration. The so-called intensity- or area-based approaches mainly utilize similarity measures like the cluster reward algorithm [20, 21], mutual information [20, 22–24] and the cross-cumulative residual entropy [25] and are often computationally expensive and sensitive to multiplicative noise in the SAR image. As they mainly infer image correspondences on the basis of pixel intensity values they are hindered by the different radiometric properties of optical and SAR images. On the other hand, feature-based approaches are focusing on the detection, extraction and matching of image features such as lines [26–29], contours [30, 31] or regions [32], or utilize point feature detector and descriptor methods such as the scale-invariant feature transform (SIFT) [33–35]. Due to their higher robustness to radiometric changes and geometric inconsistencies between the images, feature-based approaches often outperform intensity-based algorithm. However, even by combining feature- and intensity-based approaches, the development of a single approach, which is not tailored to one particular kind of image feature or a certain image scene, and therefore able to reliable co-register optical and SAR image pairs in a general way, is still an open problem.

## 1.3 Our Contributions and Focus of the Thesis

In this thesis we present a novel and automatic framework for the improvement of the absolute geo-localization accuracy of optical satellite images via tie points generated from high-resolution TerraSAR-X images. For the first time, this problem is tackled with the help of neural networks and deep learning techniques in order to avoid problems frequently occurring in former approaches. Neural networks in combination with deep learning techniques have demonstrated their potential through a variety of successful applications in research fields, such as medicine, biology, computer vision and remote sensing. Due to the daily increase of remote sensing data, this research field is becoming suitable for the application of deep learning techniques, since the training of deep neural networks requires are large amount of training data. Furthermore, the progression of deep learning algorithms enabled the modeling and solving of more and more complex problems. Our proposed concept is divided in the following three parts: 1) The identification and extraction of suitable image areas, 2) the generation of reliable and accurate tie points through deep learning boosted matching, and 3) the adjustment of the optical sensor model. The focus of our investigation is on the

first two steps, whereas for the third step we utilize an already well-proven concept without making any adjustments.

The first step of the framework forms the basis for a successful and accurate tie point generation. It is important to take into account the different radiometric and geometric properties of optical and SAR images, while developing a matching method. We present two concepts for the pre-selection of suitable matching areas, both pursuing the goal of identifying areas that only contain salient objects or features that exhibit the same geometrical properties in both images. The semi-automatic algorithm is mainly based on the usage of the CORINE land cover layer and a manual refinement. The developed automatic concept is based on the usage of existing road network information from OpenStreetMap data in combination with a deep learning-based method for the automatic segmentation of road networks in SAR images. Note that only the semi-automatic method is used in later steps. However, the automatic approach was developed to overcome the problem of the time consuming manual refinement needed, and hence to open up new possibilities for further developments.

The main part of our work concentrates on step two, consisting of two novel deep learning-based concepts, which pursue the same goal of an accurate and reliable tie point generation. They are tailored to overcome different problems of common optical and SAR registration frameworks. The first approach tackles the problem of the different radiometric properties between optical and SAR images. The possibilities provided by a new machine learning architecture, called generative adversarial networks, for the task of image-to-image translation enable the creation of a novel strategy for the reduction of radiometric differences between the images to a minimum. As a result, traditional algorithms for the matching can be applied, and hence the creation of tie points between artificially generated images and corresponding target images become feasible. The second approach pursues the idea of an end-to-end tie point generation, which does not require any handcrafted feature detection and extraction. Siamese neural networks have already proven their potential for the matching of single sensor image pairs. In this work, they are adapted in order to achieve an accurate and reliable tie point generation method for optical and SAR image pairs, through a careful adaption of the network architecture towards the particular characteristics of these images and the development of a target-oriented training procedure. The two methods are trained on a set of training images, the hyperparameters of the neural networks tuned on a set of validation images and subsequently, the set of generated tie points validated on an independent set of test images, where their accuracies and precisions are compared to state-of-the-art techniques.

In order to complete our image registration framework, we utilized an already existing technique for the sensor model adjustment of optical images through the generated sets of tie points. Since this has already proven its effectiveness, we apply it without conducting any changes. Using the adjusted sensor models for the geo-referencing of the optical images lead to the pursued goal of an improvement in the absolute geo-localization accuracy. Summarizing, our framework reduces the effort for handcrafted processing steps to a minimum, is applicable to generic optical and SAR image pairs, and outperforms state-of-the-art approaches.

## 1.4 Organization of the Thesis

The thesis is divided in six chapters. In Chapter 1, the topic of this thesis is introduced, its scientific relevance discussed and the focus and our contributions outlined. The theoretical background of optical and SAR sensors and a comparison between the different imaging properties of both sensors are provided in Chapter 2. In addition, this chapter also contains a detailed introduction on supervised machine learning principles, focusing on artificial neural networks and generative adversarial networks. The basic principle of image registration techniques, state-of-the-art optical and SAR image registration methods and an overview on recently developed deep learning-based image registration frameworks are described in detail, and research gaps are discussed in Chapter 3. Chapter 4 covers the methodological contribution of our work, which includes an image registration framework and two novel tie point generation approaches based on conditional adversarial networks and Siamese neural networks, respectively. Then, the results of both tie point generation approaches and their potential for a precise optical and SAR image registration are evaluated and discussed in Chapter 5. At last, the overall results and findings of this thesis and on outlook on future work are summarized in Chapter 6.

# 2

# Theoretical Background

*This chapter provides the theoretical foundations of the developed concepts, which are presented in later chapters of this thesis. In the first part, the basic concepts of optical and SAR sensors are introduced. Additionally, a comparison of the different imaging properties between optical and SAR sensors is outlined. In the second part, the principles concepts and required terms of supervised machine learning are introduced. The theory about convolutional neural networks and conditional adversarial networks represents here the fundamental frameworks of the later developed multi-modal image registration concepts.*

## Contents

## 2.1 Optical and Synthetic Aperture Radar Satellite Imagery

In the field of remote sensing different sensors are utilized in order to acquire spatial, spectral, and temporal information on objects or areas. In this thesis, images of two sensors are utilized, namely optical (passive) and SAR (active). Each of these sensors follows a particular acquisition process and comes with specific advantages and disadvantages. As outlined in Chapter 1, the pursued aim of this thesis is to improve the absolute geo-localization accuracy of optical images through the use of high-resolution SAR data. To understanding why this undertaking is necessary and possible, but on the other hand difficult, the principles of optical and SAR image acquisition and relevant image properties are presented in the remainder of this section.

### 2.1.1 Principles of Optical and SAR Sensors

In this subsection the relevant principles of optical and SAR sensors will be shortly introduced. For a detailed summary of remote sensing principles (including optical and SAR images) we refer to [37, 38] and for a detailed overview on the foundations of SAR data to [39].

**Optical sensors:** Optical satellite sensors are passive systems that measure the sunlight reflected from ground objects with a strong dependence on atmospheric and local weather conditions such as cloud and haze. More precisely, they detect the reflected or emitted electromagnetic radiation from objects on the ground in the visible and infrared (near infrared, intermediate infrared, thermal infrared) range of the electromagnetic spectrum (see Figure 2.2). Each object on ground reflects and absorbs thereby a specific part of the spectrum, and hence shows a specific spectral reflectance signature in the generated images. Depending on the number of spectral bands used in the imaging process, optical sensors can be classified into panchromatic, multispectral and hyperspectral sensors. In this thesis, we utilize images acquired with the high-resolution panchromatic sensor called PRISM mounted on the Earth observing satellite ALOS. Panchromatic sensors acquire images with single wide spectral band usually in the range of 400-900 nm. The nadir optical system of PRISM operates in the range of 520-770 nm and provides images with a spatial resolution

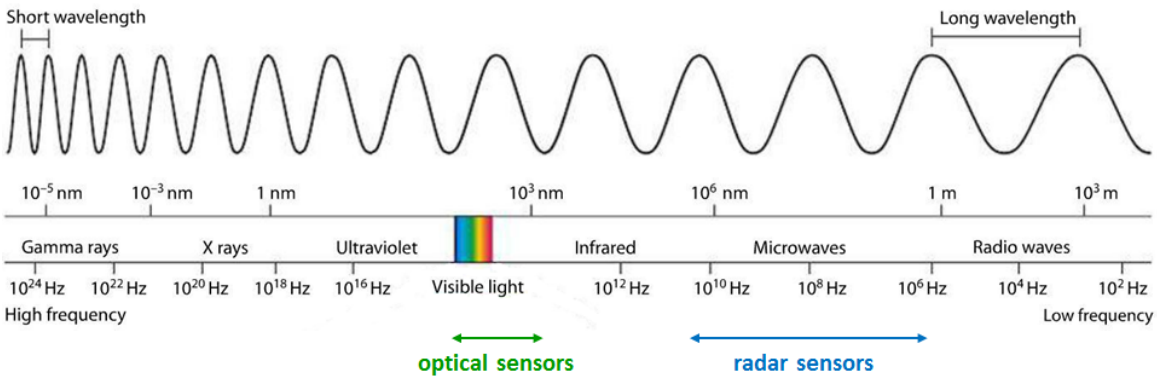



**Figure 2.1:** The electromagnetic spectrum and the operation ranges of optical and radar sensors (image source: [36]).

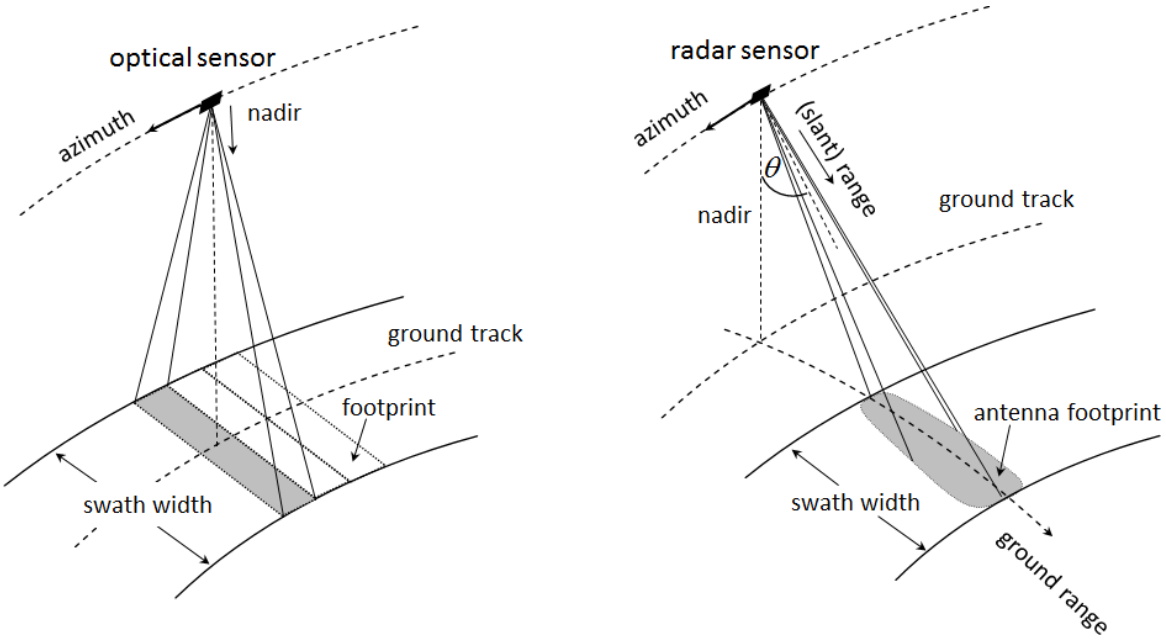



**Figure 2.2:** Comparison of the different acquisition geometries between optical and SAR sensors (source of the right image: [41]).

of 2.5 m and a swath width of 35 km. The images are thereby generated through the use of a pushbroom scanner consisting of a linear array of 14,000 detector elements, which are arranged perpendicular to the flight direction of the satellite and simultaneously receive information from the ground [40]. In contrast to full-frame photography, where the whole image is captured at the same time, such scanner systems scan and record the ground line by line [37]. The particular acquisition geometry of an optical satellite equipped with a nadir looking pushbroom sensor is illustrated on the left side of Figure 2.2. In the later Subsection 4.4.1 more details about the image generation process through a pushbroom scanner system and the used physical sensor model (sets the geometric relation between images and their corresponding ground coordinates) will be presented.

**Radar sensors:** In contrast to optical satellites, radar satellites have an active sensor on board, which emits electromagnetic signals and measures the strength and time delay of the returned signal backscattered from the objects on ground. During image acquisition the range, magnitude and Doppler shift of the reflected signal is collected by an antenna and later processed to a two-dimensional image of the surface. Due to the active emitting of a signal and usage of longer wavelength compared to optical sensors, images can by captured day and night and almost independently from local weather conditions. The term radar stands thereby for radio detection and ranging, which denotes the technique to measure the distance between a target and the sensor by exploiting the electromagnetic radiation-matter interaction. Conventional radar satellites apply the principle of SAR in order to enable the acquisition of radar images from space. The idea behind the concept of SAR is to synthesize a very long antenna by moving a shorter one along the flight path. Thereby, the backscattered signal energy for ground objects along the sensor flight path is integrated and the signal energy compressed in post-processing for a significant increase of the spatial resolution [39].

Here, the data post-processing forms the essential part of the image generation process. In order to avoid ambiguities in azimuth (flight direction) related to the targets on ground, a SAR sensor looks sideways. The look angle of the sensor to an object on ground is called incident angle. The image acquisition geometry of a satellite equipped with a radar sensor on board is illustrated on the right side of Figure 2.2. The commonly used SAR systems use wavelength in the range of 2.4 cm to 20 cm, an incident angle between 20° to 60° (with respect to nadir direction) and can operate in three modes: stripmap, spotlight and scanSAR. In this thesis, stripmap images from the SAR satellite TerraSAR-X with a resolution of 1.25 m are used. In stripmap mode the antenna is pointing along a fixed direction broadside to the platform track. TerraSAR-X is operating with a wavelength of 3.1 cm, an antenna size of $4.8 \times 0.8\,\mathrm{m}^2$, a swath width of 5 km to 10 km (in spotlight mode), and an incident angle between 22° and 55°.

### 2.1.2 Characteristics of Optical and SAR Imagery

In the case of optical and SAR satellites, the images acquired by both sensors exhibit quite different properties that characterize the images. In particular, the specific acquisition principle of a radar sensor and the resulting image effects make the visual interpretation and usage of SAR images a challenging task [42]. The particular characteristics of both sensors have to be taken into account while analyzing the images or, in our case, to develop an optimal image registration strategy. Therefore, the relevant image properties of optical and SAR image are discussed and compared in the following paragraphs.

**Radiometric Properties:** The different wavelengths measured and utilized by optical and SAR sensors lead to different radiometric properties in the images. This is due to the fact that the response of an object depends on the signal properties (wavelength, polarization), the surface properties (roughness, randomness of local reflectors and reflectance properties) and sensor perspective. Here, optical sensors measure the reflected radiation in the visible and near-infrared region of the electromagnetic spectrum in order to generate an image. This particular part of the spectrum enables the collection of information about the chemical structure of an object on ground. The pixel intensity values of optical images therefore contain information about the chemical characteristics of an observed area. SAR sensors, in contrast, utilize electromagnetic signals with a much lower frequency and energy. Therefore, the obtained images mainly capture physical and geometrical properties of the objects on ground, where the pixel intensity values contain information about the roughness, the electrical conductivity and the orientation of an object to the sensor [38]. As a consequence, the same object in an optical and SAR image may appear with high intensity for one sensor and with low intensity for the other. Another effect in SAR images is called speckle, which further complicates the human and automatic interpretation of the images. Speckle is due to the coherent interference of waves that are reflected from many scatterers in each resolution cell. As a consequence, neighboring pixels may show a high variation in their pixel intensity values. An example showing the different radiometric properties is provided in Figure 1.1.

**Geometric Properties:** Due to the different acquisition principles of optical and SAR satellites (measuring signal travel time vs. angles), the corresponding images further exhibit quite different geometric properties for all objects above the ground. The optical images used in the thesis are acquired through the use of a scanner system. For these kind of systems, above ground objects located perpendicular to the direction of flight get projected away from the sensor in the image plane. The opposite is the case for radar systems, where the image geometry is derived through the traveling time of the backscattered signal. Therefore, above ground objects get projected towards the sensor in the image plane [37]. An illustration of the different image geometries between optical and SAR images is shown in Figure 2.4. Here, the different projection (towards and away from the sensor) can be seen for the point $c$ located on the roof of the house. SAR images further shows three typical distortion effects called foreshortening, layover and shadowing. These effects occur along objects above the ground level and have a strong influence on their visual appearance within the image. Foreshortening denotes the shortening of a distance between two points during the projection in the image plan. This effect occurs if a slope is facing the sensor and has an angle $\alpha$ smaller than the incident angle $\theta$ of the SAR sensor or if a slope is facing away from the sensor with an angle smaller than $90° - \theta$. Layover appears if a slope is facing the sensor and has an angle higher than the incident angle of the SAR sensor. This effect is particularly common in urban and

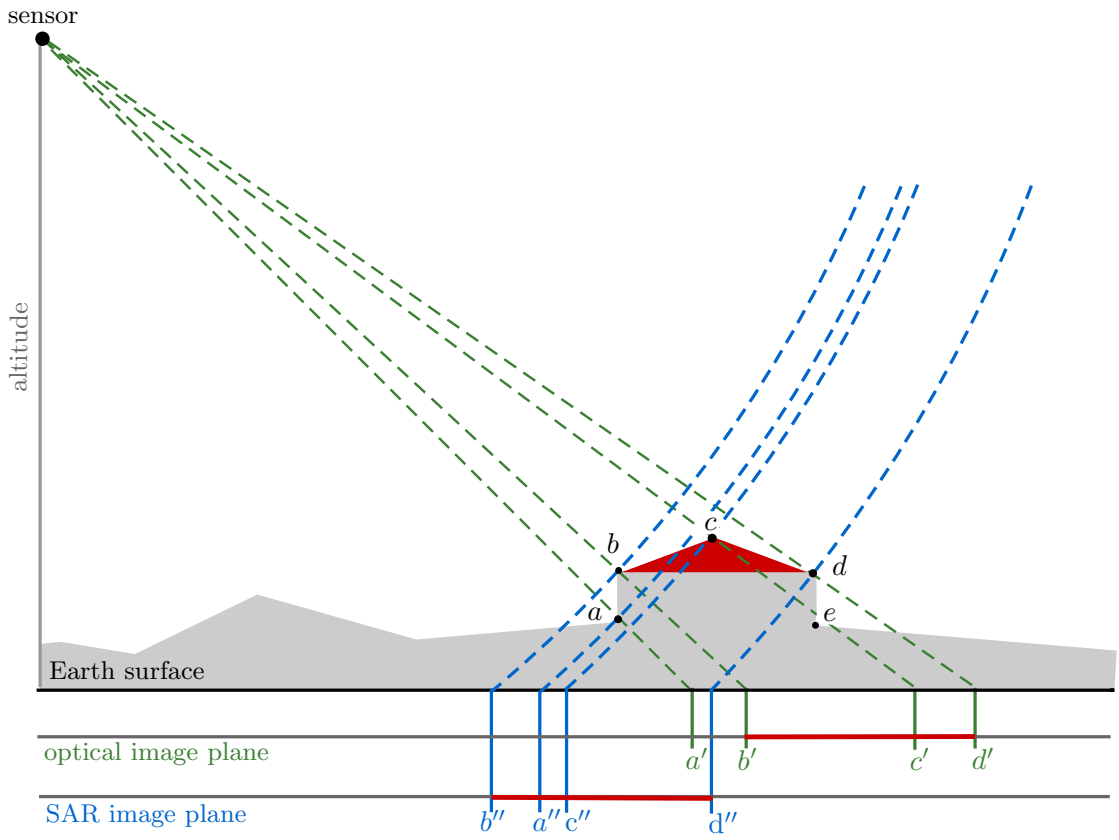



**Figure 2.3:** Comparison of optical and SAR imaging. The green (blue) marked lines illustrate the projection of the four points $a$ to $d$ on the Earth surface into the optical (SAR) image plane. Elevated points such as point $c$ are shifted away from the sensor in the optical image plane and towards the sensor in the SAR image plane. The point $e$ is neither seen by the optical nor SAR sensor, and hence not present in the acquired images.

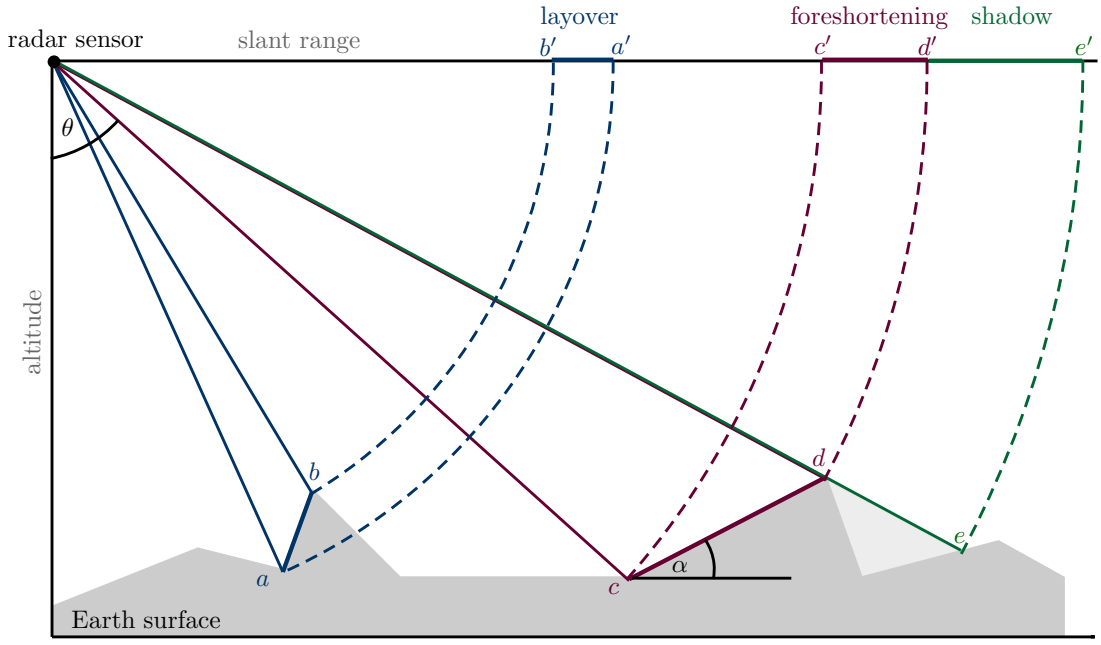
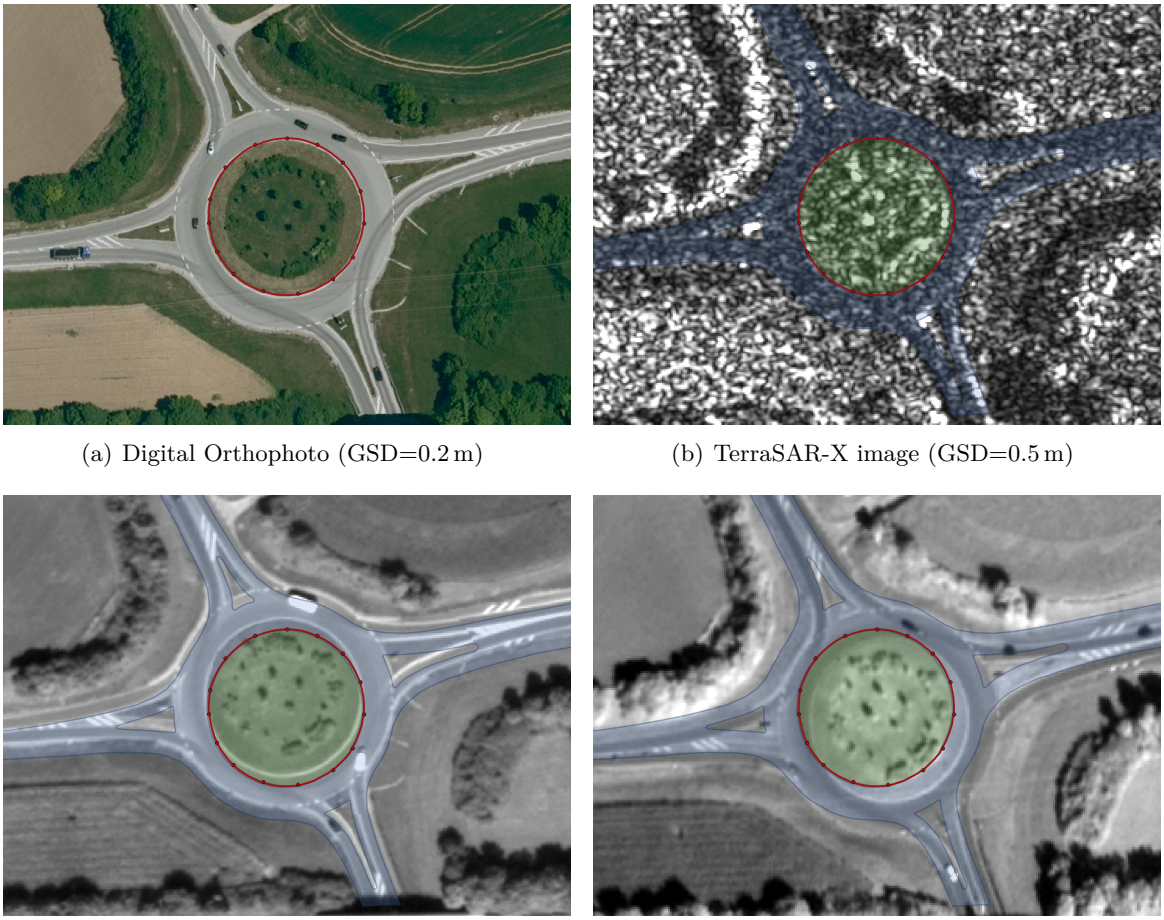



**Figure 2.4:** Illustration of the geometric distortion effects layover (marked blue), foreshortening (marked dark red) and shadowing (marked green) for SAR images. Layover: an observed object appears upside down in the image plan; Foreshortening: an observed object or ground segment appears shortened in the image plan; Shadow: non-visible regions appear as dark areas in the image.

mountainous areas. As a consequence, areas affected by overlay appear relatively bright (due to the overlay of the signal response) and buildings and steep mountains can appear upside down in SAR images. Shadowing on the other hand appears, if a terrain slope is oriented away from the sensor and at an angle higher than the incident angles of the sensor. Since no information can be gained from these shadowed regions, these appear as dark areas in the images. Note that all introduced radar effects depend on the viewing direction of the sensor and the geometry of the targeted object on ground. A visualization of the layover, foreshortening and shadowing effect of SAR sensors is provided in Figure 2.4.

**Positioning Accuracy:** Furthermore, the different acquisition modes have also an effect on the geo-referencing process. The location accuracy of optical satellites depends on a precise knowledge of the satellite orientation in space in order to determine the satellite-viewing direction to ground objects. The required measurements of the attitude angles in space often suffers from insufficient accuracies of the measurements, and are the main reason for a lower geo-localization accuracy of optical satellite data. For example, the absolute geo-localization accuracy of images for optical satellites like PRISM, Worldview-2, or QuickBird ranges from 4 m to 30 m. Images captured by high-resolution SAR satellites on the other hand, exhibit a much higher geo-localization accuracy, mainly due to the availability of precise orbit information and the recent developments in SAR geodesy. More precisely, SAR sensors determine the distance to ground object via the signal traveling time, which can be measured precisely if also atmospheric effects are taken into account, and lead to images with high geo-localization accuracy. The SAR images used in this thesis are acquired by the satellite TerraSAR-X [43] and exhibit an absolute geo-localization accuracy in the range of a few decimeters or centimeters for specific targets [14]. An example of the differences in the



(a) Digital Orthophoto (GSD=0.2 m)

(b) TerraSAR-X image (GSD=0.5 m)

(c) WordView-2 image (GSD=0.5 m)

(d) QuickBird image (GSD=0.6 m)

**Figure 2.5:** Visualization of the absolute geo-localization accuracy of different sensors. The red marked dots and lines represent GPS measurements.

positioning accuracies of optical and SAR satellite images is provided in Figure 2.5. The red dots are GPS measurements along the inner circle of the roundabout and represent with an absolute geometric accuracy within a few centimeters the ground truth. Figure 2.5(a) shows a digital orthophoto (DOP), which is accurately geo-referenced and almost perfectly aligned with the GPS measurements. A similar situation can be seen in Figure 2.5(b). Here, a TerraSAR-X image of the same scene overlaid with the GPS measurements and the street extracted from the DOP is displayed. Figure 2.5(b) and (d) show the same information but underlaid with a WordView-2 and QuickBird image, respectively. The geo-localization error of both optical images is clearly visible. Furthermore, this error depends on the underlying sensor model and varies between images acquired from different satellites.

Based on the discussed optical and SAR sensor principles and the resulting different image properties, two novel methods for the matching of optical and SAR images will be introduced in Chapter 4. In particular, the different radiometric and geometric properties are thereby taken into account in order to enable the exploitation of the high positioning accuracy of SAR data for the positioning improvement of the optical images. In the following Section 2.2, the principles of supervised machine learning are presented. The theoretic concepts discussed here form the basis of our later presented optical and SAR image matching methods.

## 2.2 Principles of Supervised Machine Learning

In 1959, Arthur Samuel coined the term machine learning and defined it as "a field of study that gives computers the ability to learn without being explicitly programmed" [44]. A more formal definition of machine learning, which mentions the term "learning", was given by Tom Mitshell in 1997: "A computer program is said to learn from experience $E$ with respect to some class of tasks $T$ and performance measure $P$ if its performance at tasks in $T$, as measured by $P$, improves with experience $E$" [45]. The experience $E$ is a dataset (a collection of many examples), the task $T$ defines the actually goal of the training, e.g. image denoising or classification, and the performance measure $P$ is usually tailored to the task $T$ with the aim of evaluating the performance of the algorithm, e.g. accuracy measure of a classification problem. The term learning in both definitions refers to the process of attaining the ability to perform a task on the basis of given data or past experience.

The type of learning depends on the kind of given experience and broadly divides machine learning algorithms into three categories. The first class, supervised learning, tries to learn the mapping from input data $X$ to a set of corresponding labels $Y$ (labeled data) and is commonly applied on regression or classification problems. The second class, unsupervised learning, deals in contrast with the problem of analyzing and learning the structure of the input data $X$ without corresponding labels (unlabeled data), e.g. clustering or density estimation. The last class, reinforcement learning, handles the problem of finding suitable actions in a given situation to maximize a reward. The reward is defined by the quality of the action and the learning is realized through a trial and error process (the algorithm is not provided with the optimal actions). The diagram in Figure 2.6 comprises an overview and the general ideas of the three learning types.

The structure of any machine learning algorithm has to be carefully designed before the learning step in order to fulfill the needs and requirements of a particular task. The key of each machine learning algorithm are trainable parameters, often called weights, which are optimized during the learning process. Applying a machine learning algorithm to solve a specific task, commonly involves the following three phases: (1) the learning or training

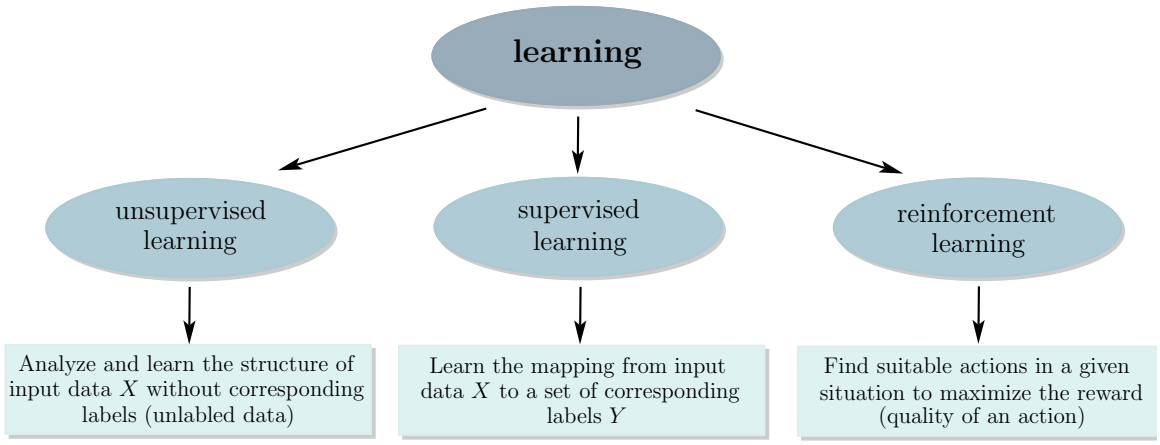



**Figure 2.6:** Overview and general idea of the three types of learning: unsupervised, supervised and reinforcement learning.

phase, (2) the validation phase and (3) the test or inference phase. In the following, we will discuss the aim and process of these phases in the context of supervised learning. More information about unsupervised learning can be found in [46–48] and about reinforcement learning in [47, 49].

**Supervised Learning - the General Goal**

The goal of a supervised learning algorithm is to learn a function

$$f : X \to Y \,, \tag{2.1}$$

when a set of input data $X$ and a set of corresponding labels $Y$ are given. The function $f$ assigns every input $\boldsymbol{x} \in X$ to an output $\boldsymbol{y} \in Y$ given a set of labeled input-output pairs $\mathcal{D} = \left\{ \left( \boldsymbol{x}^{(n)}, \boldsymbol{y}^{(n)} \right) | n = 1, \ldots, N \text{ and } N \in \mathbb{N} \right\}$. The variables $\boldsymbol{x}$ and $\boldsymbol{y}$ can be scalars, vectors or matrices. If the elements of $\boldsymbol{y}$ are continuous, it is called a regression problem, and if the elements are discrete, it is called a classification problems. The function $f$ is often referred to as the model and is defined by the set of trainable parameters $\boldsymbol{\theta}$. The assigned values $\boldsymbol{y} = f(\boldsymbol{x}, \boldsymbol{\theta})$ are commonly called the predictions of the model. Before starting the actual learning process, the set of given input output pairs $\mathcal{D}$ is partitioned into three subsets, a training $\mathcal{D}_{train}$, validation $\mathcal{D}_{val}$ and test set $\mathcal{D}_{test}$.

**Training Phase:** The aim of the training phase is to find the optimal parameters $\boldsymbol{\theta}$ based on the given training data. In order to find an optimal approximation $f(\widehat{\boldsymbol{\theta}})$ of the true but unknown function $f^*$ (given $\mathcal{D}_{train}$) the quality of the model has to be evaluated. In the case of supervised learning the quality of a model can be measured by regarding the error in the model predictions $\widehat{\boldsymbol{y}}^{(n)} = f(\boldsymbol{x}^{(n)}, \widehat{\boldsymbol{\theta}})$. This measure is often called the loss, error or cost function and is defined as

$$\mathcal{L}(\boldsymbol{y}^{(n)}, \widehat{\boldsymbol{y}}^{(n)}) = \begin{cases} 0 & \text{if } \widehat{\boldsymbol{y}}^{(n)} = \boldsymbol{y}^{(n)} \\ > 0 & \text{otherwise} \end{cases} . \tag{2.2}$$

The loss function enables to penalize incorrect model predictions ($\widehat{\boldsymbol{y}}^{(n)} \neq \boldsymbol{y}^{(n)}$) during the training process. The optimal model parameters $\boldsymbol{\theta}^*$ are computed by minimizing the overall error, which gradually increases the quality of the learned model

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \sum_{n=1}^{N} \mathcal{L}(\boldsymbol{y}^{(n)}, \widehat{\boldsymbol{y}}^{(n)}) \,. \tag{2.3}$$

This procedure will lead to a model, which is able to provide accurate predictions for input values from the training set, but not necessary for unseen data. In order to find a model which further provides accurate predictions for unseen data and, hence, is applicable to a real world task, the networks ability to generalize has to be monitored during the training.

**Validation Phase:** The validation phase pursues the goal of estimating the generalization performance of the trained model. Therefore, the performance of the model is measured by evaluating the loss function $\mathcal{L}$ on the validation dataset. The training and validation phase can be executed alternately till the best model is found. Commonly, the model showing the best performance on the validation set is picked as the best and final model. A second purpose

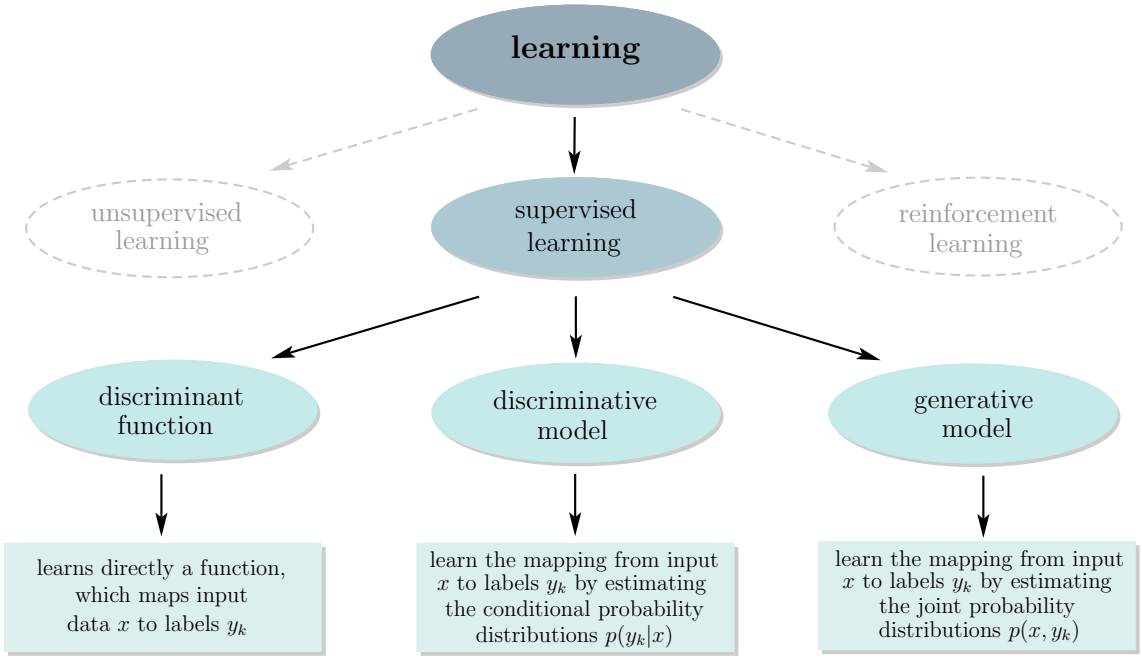



**Figure 2.7:** Overview and general idea of the three branches of supervised learning: discriminant functions, discriminative models and generative models.

of the validation phase is to tune the hyperparameters of the model. Hyperparameters are model configurations, which are set before the actual learning is realized and cannot be directly learned during the training phase. In order to tune them, algorithms like grid search or random search are often used.

**Test Phase:** The aim of the last phase, called test phase, is to evaluate the performance of the final model on an independent test dataset. As in the training and validation phase, the performance is measured by utilizing the loss function $\mathcal{L}$. Since the test set contains only unseen data (not used for the training or validation phase), this step reveals the ability of the model to generalize to unseen data and its quality to perform on the desired task.

The quality of the final model depends next to the chosen type and complexity of the algorithm, the set of hyperparameters, the loss function and the optimization procedure also on the quality and amount of the given training data. It is important that the selection of the set of data pairs $(X, Y)$ represents the real task and that the distribution of the data is a good approximation of the real data distribution. Furthermore, the right balance between the complexity of the model and the amount of training data with respect to the actual task has to be found to limit the generalization error. In this context terms like under- or overfitting a commonly used, which will be thematized in Subsection 2.2.1.

### Supervised Learning - the Types of Learning

Supervised learning algorithms can be further divided into three subcategories depending on how the mapping from $X$ to $Y$ is actually learned during the training phases (shown in Figure 2.7). The first category comprises discriminant functions, which define a class of algorithms that try to find directly a function which maps an input value $\boldsymbol{x}$ to an output label $\boldsymbol{y}$ (non-probabilistic algorithms). The learned model does not provide any confidence

or a probability for an output label $\boldsymbol{y}$. If we consider a simple classification problem (such as illustrated in Figure 2.8), where $X$ is a set of images (represented as colored dotes) and $Y$ a set of kinds of animals shown in the image (cat, dog or elephant), a discriminant function will try to learn the decision boundaries between the classes, which divide the feature space into separate areas. Depending on the learned decision boundaries, new images would be assigned to a label without providing the uncertainty of the model.

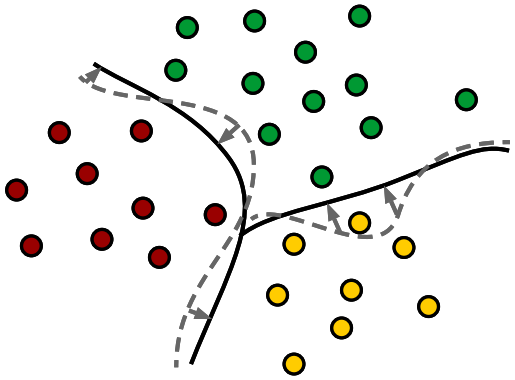


**Figure 2.8:** Illustration of a simple classification problem with three classes (red, yellow and green dots) and the corresponding decision boundaries (black lines). The gray line marks a possible non-optimal decision boundary during the training phase of the discriminant function.

The second class are discriminative models, which define a class of algorithms that try to learn a statistical model to estimate the conditional (posterior) probability distributions $p(\boldsymbol{y}|\boldsymbol{x})$ from input data $\boldsymbol{x} \in X$ to labels $\boldsymbol{y} \in Y$ (probabilistic algorithms). The distributions $p(\boldsymbol{y}|\boldsymbol{x})$ provide the probabilities of each label $\boldsymbol{y}$ given a fixed input value $\boldsymbol{x}$. The goal of assigning or predicting a label to each input value $\boldsymbol{x}$ is realized by evaluating the function

$$f(\boldsymbol{x}) = \arg\max_{\boldsymbol{y} \in Y} p(\boldsymbol{y}|\boldsymbol{x})\,. \tag{2.4}$$

Discriminative models learn $p(\boldsymbol{y}|\boldsymbol{x})$ directly from the data and do not consider the underlying data distribution. In contrast to discriminant functions, discriminative models provide next to the predicted label for a given input data also the corresponding conditional probabilities. This additional information helps to evaluate the confidence of the model regarding a certain prediction.

The last class, generative models try to learn the joint probability distributions $p(\boldsymbol{x}, y_k)$, or in other words, they try to explicitly model the distribution behind the data. In the case of generative models the predicted label of an input value $\boldsymbol{x}$ is determined by evaluating the function

$$f(\boldsymbol{x}) = \arg\max_{\boldsymbol{y} \in Y} p(\boldsymbol{x}, \boldsymbol{y})\,. \tag{2.5}$$

The example in Figure 2.9 shows the same classification example as in Figure 2.8, but this time handled via a generative model. A generative model tries to learn the joint probability distributions $p(\boldsymbol{x}, y_1)$, $p(\boldsymbol{x}, y_2)$ and $p(\boldsymbol{x}, y_3)$ of input data $\boldsymbol{x}$ and the tree classes labels $y_1$, $y_2$ and $y_3$. The red green and yellow area in Figure 2.9 mark the areas in the image space,

which contain (with a high probability) real images showing cats, dogs or elephants. The goal of the training procedure is to minimize the distance to the true data distribution. To measure the distance to the true data distribution a measure such as the Kullback–Leibler divergence can be applied.

The choice of the optimal learning type heavily depends on the actual problem. Generative models are more expensive to compute and learning $p(\boldsymbol{x}, \boldsymbol{y})$ is generally more difficult than learning $p(\boldsymbol{y}|\boldsymbol{x})$ or directly learning the mapping from $X$ to $Y$. On the other hand, generative models perform in principle better when larger training sets are given and are richer in the sense that they implicitly model $p(\boldsymbol{y}|\boldsymbol{x})$ and $p(\boldsymbol{x})$ through $p(\boldsymbol{x}, \boldsymbol{y})$, which is given through the following equation

$$p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{x}) \cdot p(\boldsymbol{y}|\boldsymbol{x}) \qquad \text{(Bayes' Theorem)}. \tag{2.6}$$

Generative models further provide the possibility of sample new data pairs $(\boldsymbol{x}, \boldsymbol{y})$ from $p(\boldsymbol{x}, \boldsymbol{y})$. If we consider the animal image classification example with the three possible classes (cat, dog or elephant) a generative model is able not only to predict a label for a new input image, but also to produce synthetic images (data points) belonging to one of the three classes. However, discriminative models are mostly more robust w.r.t. modeling errors and blunders and regarding a classification problem, where only a decision boundary is needed to separate the classes, a discriminant function or a discriminative model generally perform better.

In the following two subsections, we will introduce the concepts and provide insights of a discriminative model, neural networks, and a generative model, generative adversarial networks (GANs). For more information about discriminant functions and discriminative and generative models we refer to [46, 47] and for a detailed discussion about discriminative and generative classifiers to [50].

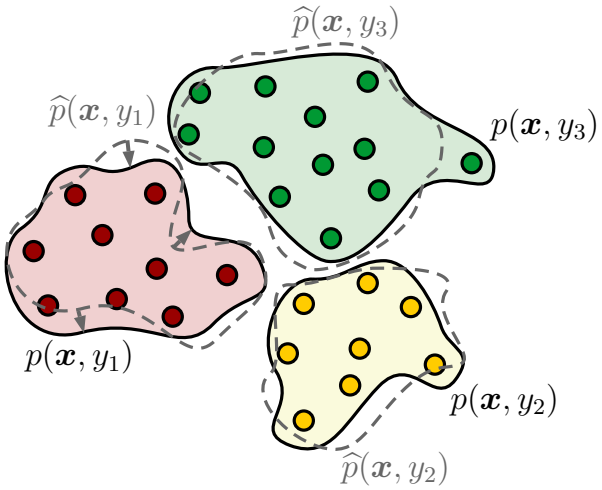



**Figure 2.9:** Illustration of a simple classification problem with three classes (red, yellow and green dots) and the corresponding joint probability distributions $p(\boldsymbol{x}, y_1)$, $p(\boldsymbol{x}, y_2)$ and $p(\boldsymbol{x}, y_3)$. The gray dashed lines illustrate possible (non-optimal) joint probability distributions $\widehat{p}(\boldsymbol{x}, y_1)$, $\widehat{p}(\boldsymbol{x}, y_2)$ and $\widehat{p}(\boldsymbol{x}, y_3)$ during the training phase of the generative model.

### 2.2.1 Artificial Neural Networks

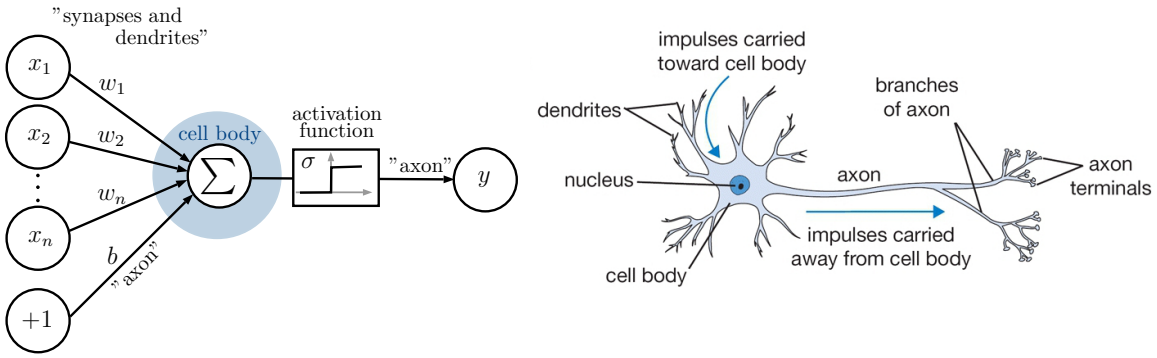
The origin of artificial neural networks (ANNs) dates back to the early 1940s where McCulloch and Pitts [51] developed the first concepts of the functional principles of biological learning system such as the brain. Inspired by this work, Rosenblatt developed the concept of perceptrons [52] in the late 1950s. A perceptron, often called artificial neuron, is a model of a biological neuron and can be understood as a computational unit which produces a single output $y$ given some input $x$. The output $y$ is given through the following equation

$$y = \sigma\left(\sum_{i=1}^{n} w_i x_i + b\right) = \sigma\left(\mathbf{w}^T \boldsymbol{x} + b\right) , \tag{2.7}$$

where the values $x_i$ $(i = 1, \ldots, n)$ are the input elements, $w_i$ the learnable weights, $b$ the learnable bias and $\sigma(\cdot)$ the activation or transfer function. The weights represent the importance of the corresponding input value and are adjusted during the training phase. The activation functions serves as a threshold, which divides the input space into two partitions. Commonly, non-linear activation functions are applied in order to introduce non-linearities into the perceptron and later to the network (otherwise only linear functions can be modeled). The optimal bias is learned during training and enables a shift of the activation function in the image space in order to find the optimal partition of the input data. An illustration of a perceptron and the comparison with the biological neuron counterpart is shown in Figure 2.10.

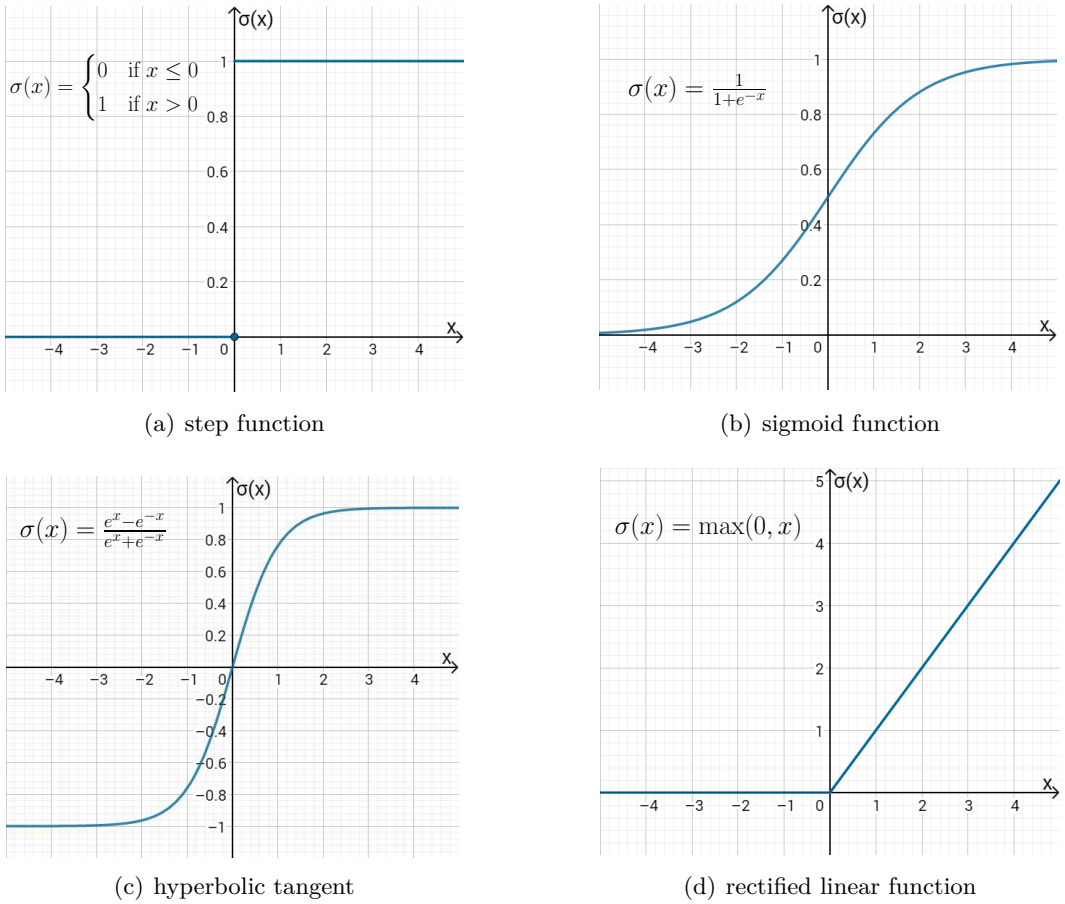
In the original work [52], Rosenblatt proposed the application of a step function as activation function, whereas nowadays non-linear functions such as the sigmoid, the hyperbolic tangent or rectified linear functions are usually applied (see Figure 2.11). This development can be traced back to an unstable behavior caused by a step function: small changes in the input values can lead to huge changes in the output value, which complicates a gradually adjustment of the weights, and hence complicates the training procedure.

A single perceptron is only capable of learning a linear separation of the input data, heavily limiting the number of application cases. This drawback can be overcome by composing




(a) perceptron (artificial neuron) (b) biological neuron (image source: [53])

**Figure 2.10:** Illustration and comparison of a: (a) perceptron (artificial neuron) and (b) biological neuron model. The values $(x_1, \ldots, x_n)$ are the input values, $(w_1, \ldots, w_n)$ the corresponding weights and $b$ the bias of the perceptron.




(a) step function

(b) sigmoid function

(c) hyperbolic tangent

(d) rectified linear function

**Figure 2.11:** Example of four non-linear activation functions: (a) a step function, (b) the sigmoid function, (c) the hyperbolic tangent and (d) a rectified linear function.

several perceptrons (artificial neurons) and connecting them to a directed acyclic graph[i], called an artificial neural network or sometimes a multi-layer perceptron (MLP). Figure 2.12 shows a simple representation of a neural network. The nodes of the graph are usually called units and represent artificial neurons. Commonly, a neural network is built of distinct layers, where the first layer is called the input layer and the last one the output layer. All layers in between are called hidden layers. In the shown example the neurons within one layer share no connections but between two adjacent layers the neurons are fully pairwise connected. Such a layer structure is called fully-connected. If the information is only fed forward through the network (only the output from previous layers is fed as input to the later layer), it is called a feedforward neural network.

The network architecture can vary in the number of input, hidden and output units, the number of hidden layers and the type of layers (overview of different layer types follows below). A network with only one hidden layer is called a shallow neural network and with more than two layers a deep neural network. The task of the layers is to extract information (features) from the input data. The complexity of the features extracted by the layers increases along with the depth, where early layers commonly detect simpler features, such as edges or corners, and later layers more complex features, such as parts of a human face.

[i]directed acyclic graph: a directed graph without cycles

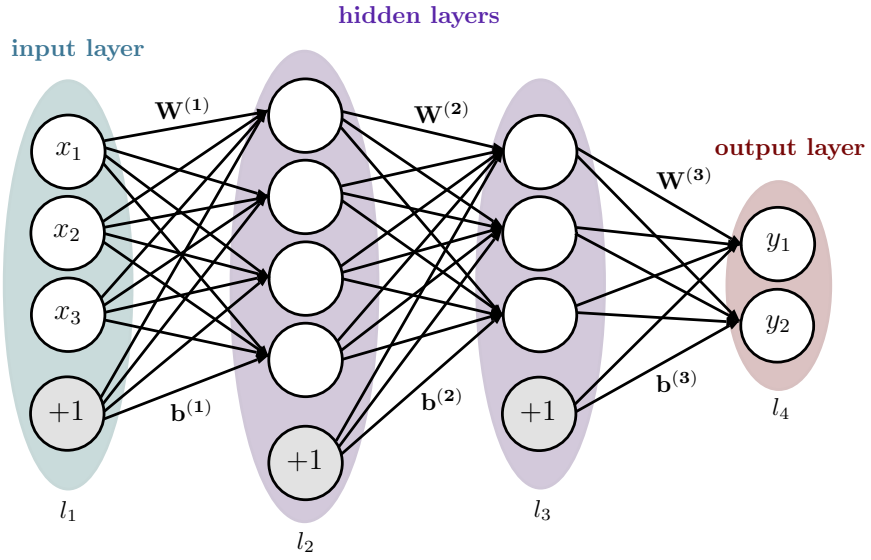



**Figure 2.12:** Example of an artificial neural network with four layers (the input layer $l_1$, two hidden layers $l_2$ and $l_3$ and the output layer $l_4$), which maps the input $\boldsymbol{x} = (x_1, x_2, x_3)$ to the output $\boldsymbol{y} = (y_1, y_2)$. Each circle represents a unit in the network and the arrows the connections between the units of adjacent layers. The unit marked with +1 represents the bias unit of the corresponding layer. The matrix $\mathbf{W}^{(t)}$ contains the weights between layer $t$ and $t+1$ and the vector $\mathbf{b}^{(t)}$ the biases from layer $t$ to $t+1$.

### Neural Networks - The Training Process

In order to train the network to learn meaningful features for solving a specific task, a loss function has to be defined. Since we are still assuming a supervised learning problem, the initial situation is the same as described in Subsection 2.2. The goal is to learn a model that predicts an output $\widehat{\boldsymbol{y}}$ given and input $\boldsymbol{x}$ on the basis of a training dataset $\mathcal{D}_{\text{train}} = \left\{\left(\boldsymbol{x}^{(n)}, \boldsymbol{y}^{(n)}\right) | n = 1, \ldots, N \text{ and } N \in \mathbb{N}\right\}$. Utilizing the mean square error (MSE) to measure the error in the predictions, the loss function for one training example can be expressed as

$$\mathcal{L}\left(\boldsymbol{y}^{(n)}, f_{\mathbf{W},\mathbf{b}}\left(\boldsymbol{x}^{(n)}\right)\right) = \frac{1}{2}\left\|f_{\mathbf{W},\mathbf{b}}\left(\boldsymbol{x}^{(n)}\right) - \boldsymbol{y}^{(n)}\right\|^2, \tag{2.8}$$

where $\mathbf{W}$ is a matrix containing all weights and $\mathbf{b}$ a vector containing the biases of the network. The overall set of learnable parameters is given by $\boldsymbol{\theta} = (\mathbf{W}, \mathbf{b})$. The predictions $\widehat{\boldsymbol{y}}^{(n)} = f_{\mathbf{W},\mathbf{b}}\left(\boldsymbol{x}^{(n)}\right)$ are obtain by computing a forward pass through the network (forward propagation) and can be computed by evaluating a chain of function (vectorized notation)

$$\begin{aligned} f_{\mathbf{W},\mathbf{b}}\left(\boldsymbol{x}^{(n)}\right) = \mathbf{a}^{(L)} \quad \text{with} \quad & \mathbf{a}^{(t+1)} = \sigma\left(\mathbf{h}^{(t+1)}\right) \\ & \mathbf{h}^{(t+1)} = \mathbf{W}^{(t)}\mathbf{a}^{(t)} + \mathbf{b}^{(t)} \quad (t = 1, \ldots, L-1)\,. \end{aligned} \tag{2.9}$$

Here, $L$ denotes the total number of layers, $a^{(t)}$ is a vector containing the so called activation values $a_i^{(t)}$ of unit $i$ in layer $t$ with $a^{(1)} = x^{(n)}$, $h^{(t)}$ is a vector containing the values $h_i^{(t)}$ of unit $i$ in layer $t$ (hidden values), $\mathbf{W}^{(t)}$ is a matrix containing all the weights $w_{i,j}^{(t)}$ between unit $i$ in layer $t$ and unit $j$ in layer $t+1$, and $\mathbf{b}^{(t)}$ is a vector containing the biases $b_i^{(t)}$ from layer $t$ to unit $i$ in layer $t+1$. A visual example of the terms is illustrated in Figure 2.13

and the detailed expression of $\mathbf{h}^{(t+1)}$ is given by the following equation

$$\mathbf{h}^{(t+1)} = \underbrace{\begin{pmatrix} w_{1,1}^{(t)} & w_{1,2}^{(t)} & \cdots & w_{1,n_t}^{(t)} \\ w_{2,1}^{(t)} & w_{2,2}^{(t)} & \cdots & w_{2,n_t}^{(t)} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n_{t+1},1}^{(t)} & w_{n_{t+1},2}^{(t)} & \cdots & w_{n_{t+1},n_t}^{(t)} \end{pmatrix}}_{\mathbf{W}^{(t)}} \cdot \underbrace{\begin{pmatrix} a_1^{(t)} \\ a_2^{(t)} \\ \vdots \\ a_{n_{t+1}}^{(t)} \end{pmatrix}}_{\mathbf{a}^{(t)}} + \underbrace{\begin{pmatrix} b_1^{(t)} \\ b_2^{(t)} \\ \vdots \\ b_{n_{t+1}}^{(t)} \end{pmatrix}}_{\mathbf{b}^{(t)}}, \tag{2.10}$$

where $n_t$ is the number of units in layer $t$. The overall error $\mathcal{E}(\mathbf{W}, \mathbf{b})$ is given by the sum overall training examples in $\mathcal{D}_{\text{train}}$. The goal of the training phase is to find the optimal set of parameters $\boldsymbol{\theta} = (\mathbf{W}, \mathbf{b})$, which minimize the overall error. This optimization problem can be stated as follows

$$\left(\widehat{\mathbf{W}}, \widehat{\mathbf{b}}\right) = \arg\min_{\mathbf{W},\mathbf{b}} \underbrace{\sum_{n=1}^{N} \mathcal{L}\left(\boldsymbol{y}^{(n)}, f_{\mathbf{W},\mathbf{b}}\left(\boldsymbol{x}^{(n)}\right)\right)}_{=\mathcal{E}(\mathbf{W},\mathbf{b})}. \tag{2.11}$$

Before applying an optimization algorithm the parameters $\boldsymbol{\theta}$ of the network have to be initialized. The weights are commonly initialized by small random values, or a specific initialization scheme such as proposed in [54], or with weights from a pre-trained network and the biases set to zero. If all weights would be initialized with the same value, every activation $a_i^{(t)}$ would be the same and, hence, every neuron would learn the same. For a detailed overview of parameter initialization strategies and its advantages and disadvantages we refer to [48] and for the specific ones used in this thesis to Subsections 5.3.1 and 5.2.1.

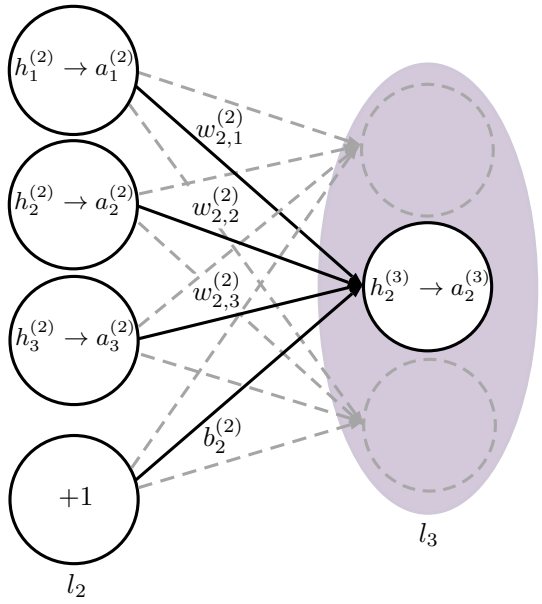



**Figure 2.13:** Illustration of the forward propagation for one unit in a neural network and the computation of the term $h_2^{(3)}$ and the activation value $a_2^{(3)}$ of the second unit in layer 3. The values $a_i^{(2)}$ are the activations of the $i$-th unit in layer 2, $w_{2,j}^{(2)}$ are the weights between the $j$-th unit in layer 2 and the second unit in layer 3 and $b_2^{(2)}$ the bias from layer 2 to the second unit in layer 3.

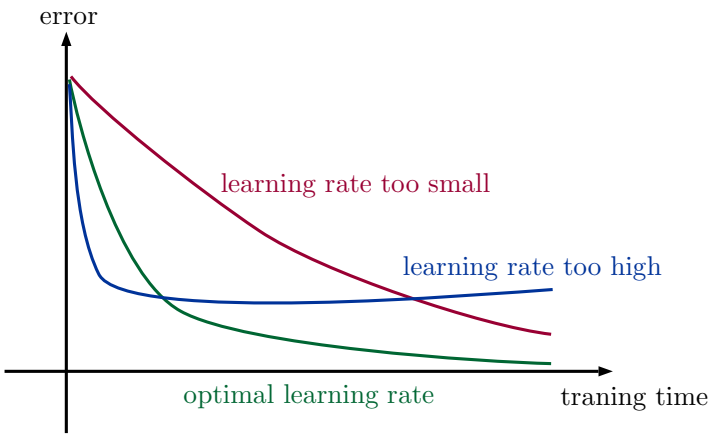



**Figure 2.14:** Influence of the learning rate on the error over the training time.

After the initialization of the parameters, an optimization algorithm such as gradient descent is applied to minimize the overall loss w.r.t. to the parameters. The idea of gradient descent is to learn the optimal parameters by updating the weights and biases in an iterative way

$$\begin{aligned} w_{i,j}^{(t)} &\leftarrow w_{i,j}^{(t)} - \lambda \frac{\partial \mathcal{E}(\mathbf{W}, \mathbf{b})}{\partial w_{i,j}^{(t)}} \\ b_{i}^{(t)} &\leftarrow b_{i}^{(t)} - \lambda \frac{\partial \mathcal{E}(\mathbf{W}, \mathbf{b})}{\partial b_{i}^{(t)}}, \end{aligned} \tag{2.12}$$

where the parameter $\lambda$ is called learning rate and determines the speed of the parameter updates. The learning rate has often a significant influence on the success of the learning and represents a hyperparameter, which has to be determined before the learning phase and tuned by utilizing the validation phase. Figure 2.14 exemplifies the influence of the learning rate on the overall error during training. A too small learning rate lead to a slow convergence of the loss function, whereas a too high learning rate can lead the loss to oscillate around the minimum or to diverge.

So, far the total error $\mathcal{E}(\mathbf{W}, \mathbf{b})$ is computed by summing up the loss function over the whole training set $\mathcal{D}_{\text{train}}$ in one iteration of gradient descent. A common approach for the training of a neural networks is to evaluate the error only over a small subset of $\mathcal{D}_{\text{train}}$, called (mini-)batch. The batches can be chosen randomly from $\mathcal{D}_{\text{train}}$ in each iteration or generated beforehand by partition the set $\mathcal{D}_{\text{train}}$ into several batches. An advantage of this strategy is that it is computationally faster (in the case of a larger training set $\mathcal{D}_{\text{train}}$). Furthermore, it converges faster, while providing a reasonable approximation of the total error due to redundant information in most datasets. This extension of gradient descent is called (mini-batch) stochastic gradient descent (SGD), where the key step of SGD is to compute the partial derivatives $\frac{\partial \mathcal{E}(\mathbf{W},\mathbf{b})}{\partial w_{i,j}^{(t)}}$ and $\frac{\partial \mathcal{E}(\mathbf{W},\mathbf{b})}{\partial b_{i}^{(t)}}$.

In 1986, Rumelhart et al. [55] proposed an algorithm for the computation of these derivatives, which enabled the efficient training of neural networks. The algorithm is called back-propagation and is based on the idea of propagating the error backwards through the network

by applying the chain rule

$$\frac{\partial f(g(x))}{\partial x} = \frac{\partial f(g(x))}{\partial g(x)} \cdot \frac{\partial g(x)}{\partial x} \,. \tag{2.13}$$

We will introduce the idea of back-propagation in the following and simplify the notations without loss of generality by exploiting the following rule

$$\frac{\partial \mathcal{E}(\mathbf{W}, \mathbf{b})}{\partial w_{i,j}^{(t)}} = \frac{\partial \sum_{n=1}^{N_b} \mathcal{L}\left(\mathbf{W}, \mathbf{b}, \boldsymbol{x}^{(n)}, \boldsymbol{y}^{(n)}\right)}{\partial w_{i,j}^{(t)}} = \sum_{n=1}^{N_b} \frac{\partial \mathcal{L}\left(\mathbf{W}, \mathbf{b}, \boldsymbol{x}^{(n)}, \boldsymbol{y}^{(n)}\right)}{\partial w_{i,j}^{(t)}} \tag{2.14}$$

and consider only the error caused by one training example $\left(\boldsymbol{x}^{(n)}, \boldsymbol{y}^{(n)}\right)$ and the corresponding derivatives $\frac{\partial \mathcal{L}(\mathbf{W},\mathbf{b})}{\partial w_{i,j}^{(t)}} \coloneqq \frac{\partial \mathcal{L}(\mathbf{W},\mathbf{b},\boldsymbol{x}^{(n)},\boldsymbol{y}^{(n)})}{\partial w_{i,j}^{(t)}}$ and $\frac{\partial \mathcal{L}(\mathbf{W},\mathbf{b})}{\partial b_i^{(t)}} \coloneqq \frac{\partial \mathcal{L}(\mathbf{W},\mathbf{b},\boldsymbol{x}^{(n)},\boldsymbol{y}^{(n)})}{\partial b_i^{(t)}}$. In the above stated equation the term $N_b$ denotes the size of the batch.

The output of the network is computed through the forward propagation and is defined by a chain of functions (see Equation 2.9). Therefore, the chain rule can be applied and the derivative of $\mathcal{L}(\mathbf{W}, \mathbf{b})$ with respect to a single weight $w_{i,j}^{(t)}$ between unit $i$ in layer $t$ and unit $j$ in layer $t+1$ can be rewritten as

$$\begin{aligned}
\frac{\partial \mathcal{L}(\mathbf{W}, \mathbf{b})}{\partial w_{i,j}^{(t)}} &= \frac{\partial \mathcal{L}(\mathbf{W}, \mathbf{b})}{\partial h_i^{(t+1)}} \frac{\partial h_i^{(t+1)}}{\partial w_{i,j}^{(t)}} = \frac{\partial \mathcal{L}(\mathbf{W}, \mathbf{b})}{\partial a_i^{(t+1)}} \frac{\partial a_i^{(t+1)}}{\partial h_i^{(t+1)}} \frac{\partial h_i^{(t+1)}}{\partial w_{i,j}^{(t)}} \\
&= \left( \sum_{k=1}^{n_{t+2}} \frac{\partial \mathcal{L}(\mathbf{W}, \mathbf{b})}{\partial h_k^{(t+2)}} \frac{\partial h_k^{(t+2)}}{\partial a_i^{(t+1)}} \right) \frac{\partial a_i^{(t+1)}}{\partial h_i^{(t+1)}} \frac{\partial h_i^{(t+1)}}{\partial w_{i,j}^{(t)}} \\
&= \underbrace{\left( \sum_{k=1}^{n_{t+2}} \frac{\partial \mathcal{L}(\mathbf{W}, \mathbf{b})}{\partial h_k^{(t+2)}} w_{k,i}^{(t+1)} \right) \sigma'\left(h_i^{(t+1)}\right)}_{\frac{\partial \mathcal{L}(\mathbf{W},\mathbf{b})}{\partial h_i^{(t+1)}}} a_j^{(t)} \,,
\end{aligned} \tag{2.15}$$

where $n_{t+2}$ denotes the number of units in layer $t+2$. For an efficient computation of the derivatives the error term $\delta_i^{(t)}$ is introduced and recursively defined for $t = 1, ..., L-1$ as follows

$$\delta_i^{(t)} \coloneqq \frac{\partial \mathcal{L}(\mathbf{W}, \mathbf{b})}{\partial h_i^{(t)}} = \left( \sum_{k=1}^{n_{t+1}} \delta_k^{(t+1)} w_{k,i}^{(t+1)} \right) \sigma'\left(h_i^{(t)}\right) \,, \tag{2.16}$$

where the first values of $\delta_i^{(L)}$ (error term of the output layer) are given by

$$\delta_i^{(L)} = \frac{\partial \mathcal{L}(\mathbf{W}, \mathbf{b})}{\partial h_i^{(L)}} = \frac{\partial \mathcal{L}(\mathbf{W}, \mathbf{b})}{\partial a_i^{(L)}} \frac{\partial a_i^{(L)}}{\partial h_i^{(L)}} = \underbrace{\frac{\partial \mathcal{L}(\mathbf{W}, \mathbf{b})}{\partial a_i^{(L)}}}_{\left(\widehat{\boldsymbol{y}}^{(n)} - \boldsymbol{y}^{(n)}\right)} \sigma'\left(h_i^{(L)}\right) \,. \tag{2.17}$$

The error term $\delta_i^{(t)}$ measures the error of each neuron $i$ in layer $t$ and starting from $\delta_i^{(L)}$ the error can be propagate backwards through the network (backpropagation). An illustration of the computation of the error term for a single neuron is illustrated in Figure 2.15.

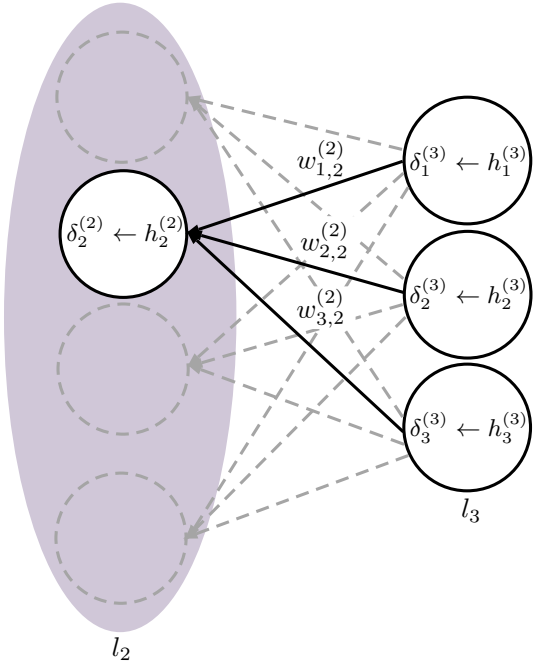



**Figure 2.15:** Illustration of the backward propagation for one unit in a neural network and the computation of the term $h_2^{(2)}$ and the error term $\delta_2^{(2)}$ of the second unit in layer 2. The values $\delta_i^{(3)}$ are the error terms of the $i$-th unit in layer 3, $w_{i,2}^{(2)}$ are the weights between the second unit in layer 2 and the $i$-th unit in layer 3.

Utilizing the concept of the error term and its definition from Equation 2.16 we can rewrite the partial derivatives of $\mathcal{L}(\mathbf{W}, \mathbf{b})$ with respect to a single weight $w_{i,j}^{(t)}$ from Equation 2.15 as follows

$$\frac{\partial \mathcal{L}(\mathbf{W}, \mathbf{b})}{\partial w_{i,j}^{(t)}} = \delta_i^{(t+1)} a_j^{(t)} \tag{2.18}$$

and the partial derivatives of $\mathcal{L}(\mathbf{W}, \mathbf{b})$ with respect to the bias $b_i^{(t)}$ as

$$\frac{\partial \mathcal{L}(\mathbf{W}, \mathbf{b})}{\partial b_i^{(t)}} = \frac{\partial \mathcal{L}(\mathbf{W}, \mathbf{b})}{\partial h_i^{(t+1)}} \underbrace{\frac{\partial h_i^{(t+1)}}{\partial b_i^{(t)}}}_{=1} = \delta_i^{(t+1)} . \tag{2.19}$$

The final algorithm can be briefly summarized into the following five steps:

1. Initialize the parameters of the network, e.g. set weights to small random values and the biases to zero.

2. Compute the forward propagation as stated in Equation 2.9 by calculating $\mathbf{a}^{(t+1)}$ and $\mathbf{h}^{(t+1)}$ for each neuron of every layer and store the results for the later backpropagation of the error.

3. Compute the error terms $\delta_i^{(t)}$ for each neuron of every layer according to Equations 2.16 and 2.17.

4. Compute the derivative of $\mathcal{E}(\mathbf{W}, \mathbf{b})$ with respect to the weights and biases according to Equations 2.18, 2.19 and 2.14.

5. Update the parameters according to Equation 2.12 and start the next iteration with step 2.

The error function stated in Equation 2.11 is non-convex and, hence, it is not guaranteed for the described algorithm to find the global minimum. However, Choromanska et al. [56] showed that in practice the local minima, which can be found by the algorithm, tend to be very similar to the global one.

The training of a deep and large neural network is often difficult even through the application of the back-propagation algorithm and SGD. Several research studies have proposed specific techniques for the training of such networks, commonly known as deep learning. These techniques include the developments of optimal initialization schemes for the network parameters [54, 57], strategies for tuning the hyperparameters [58, 59], faster and better optimization algorithms [60–63] or optimal data pre-processing approaches [57]. However, one of the major research efforts were and are still spent on the development of techniques to increase the ability of the network to generalize (providing reasonable predictions to unseen data).

In 1991, Hornik [64] demonstrated that a neural network with only one hidden layer can approximate any function. More precisely, by training a shallow neural network an optimal model can be found, which fits perfectly to the training data. However, it is not guaranteed that this (optimal) model performs reasonable on an unseen datasets or is able to learn features at various levels like a deep neural network. The problem, where a model shows a low error on the training set, but has not learned to generalize (high error on the test set) is called overfitting. Especially deep neural networks with a large number of parameters are vulnerable to this problem. The illustration in Figure 2.16 shows the effects of overfitting on the training and validation set error.

During the training phase the complexity of the learned model increases successively. As a consequence, the model fits better and better to the training data, which leads to a reduction in the training error (illustrated by the blue curve). The period, where the complexity of the model is too low to fit to the data and, hence, the training and generalization error (computed over the validation set) are high, is called underfitting. At the beginning of the training, the generalization error shows a similar behavior to the training error and decreases over time. The moment the model starts overfitting to the training data, the validation error starts to increase. If the training will be continued after this point, the model ability to generalize will successively decrease. By monitoring the loss function of the training and validation set over the training time the occurrence of overfitting can be identified and the training procedure can be stopped the moment the validation error starts to increase (early stopping). In the following further techniques to prevent the network from overfitting will be introduced.

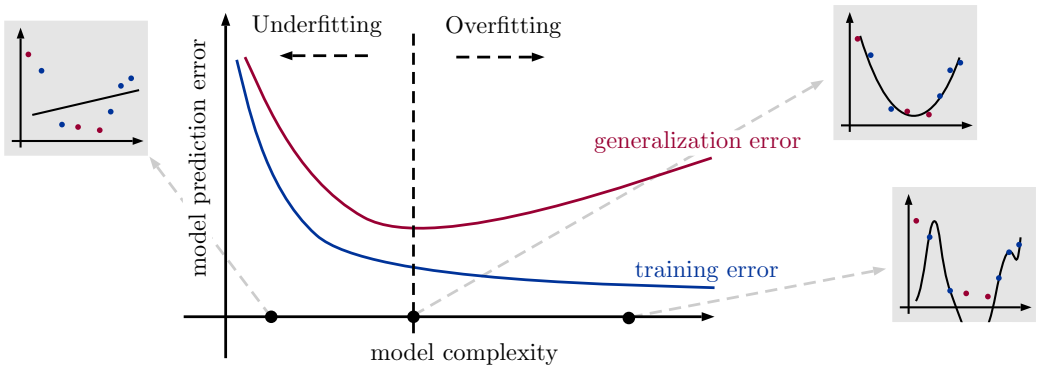



**Figure 2.16:** Illustration of the overfitting problem of neural networks. The blue and red curves show the training and generalization error (computed over the validation set) of the network over the training time and with respect to the learned model complexity. The gray framed images show examples of learned models during the training given some training data (blue points). The longer the training, the higher the model complexity and its ability to fit to the training data, but the higher the generalization error on a validation set (red points).

### Neural Networks - Types of Regularization

All techniques with the goal of preventing the network from overfitting are summarized in general under the term of regularization. One way of regularization is to increase the size of the training dataset. This can be achieved by adding new data to the training set or by generating new data from the existing data (data augmentation). Commonly used forms of data augmentations are the addition of noise or the application of transformations, such as shifting, rotation or scaling on the existing data.

A further way of regularization is to put constraints to the parameters of the network. A common approach is to extend the loss function by an additional regularization term $\mathcal{R}$, which lead to the modified loss function

$$\mathcal{L}\left(\boldsymbol{y}^{(n)}, f_{\mathbf{W},\mathbf{b}}\left(\boldsymbol{x}^{(n)}\right)\right) = \frac{1}{2}\left\|f_{\mathbf{W},\mathbf{b}}\left(\boldsymbol{x}^{(n)}\right) - \boldsymbol{y}^{(n)}\right\|^2 + \mu \cdot \mathcal{R}(\mathbf{W})\,. \tag{2.20}$$

The value $\mu$ is called the regularization parameter and defines the influence of the regularization term. It represents also another hyperparameter, which as to be adjusted with the help of the validation set. The idea of $\mathcal{R}$ is to restrict the values of the weights by penalizing too large weights. In practice, the $L_1$- and $L_2$-norm are commonly utilized, where $\mathcal{R}$ is given as $\mathcal{R} = \|\mathbf{W}\|_1$ or $\mathcal{R} = \frac{1}{2}\|\mathbf{W}\|_2^2$, respectively.

Another form of regularization is to reduce the number of parameters and, hence, limit the model complexity. This can be achieved by changing the network architecture through the reducing the number of hidden layers or the number of units per layer. However, in practice it has been proven that deeper networks in combination with regularization techniques, which keep the basic network structure fixed, perform better. One of these techniques is called dropout [65]. Dropout is a technique, which tries to prevent the network from adapting too much to the training data by randomly dropping units and their corresponding connections during the training. As a result, a slightly different network is trained at each step of the training procedure.

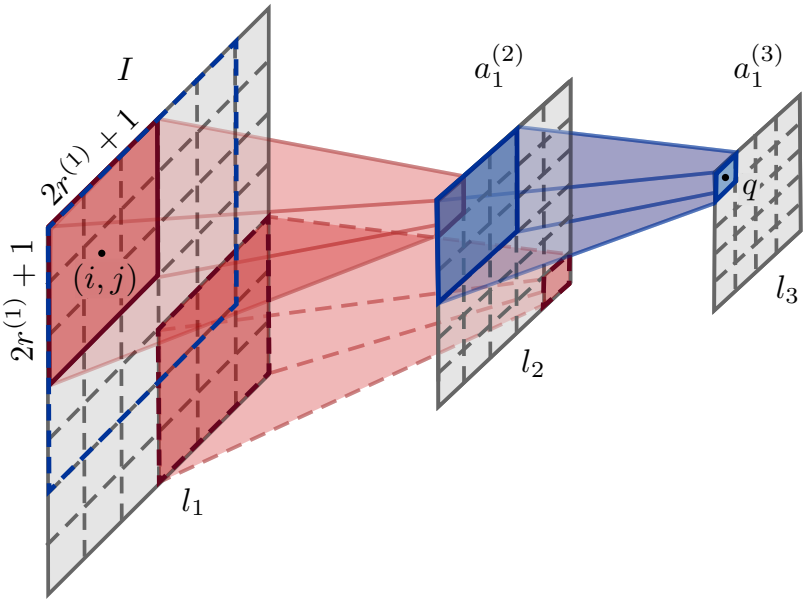



**Figure 2.17:** Illustration of convolutional layers with filters of size $3 \times 3$. The filter between layer $l_1$ and $l_2$ is represented by the red marked square and the filter between layer $l_2$ and $l_3$ by the blue marked square. $a_1^{(2)}$ and $a_1^{(3)}$ denote the feature maps of layer $l_2$ and $l_3$, respectively. The blue dashed line in the input image $I$ illustrates the receptive field of the point (pixel) $q$ in the feature map $f_1^{(3)}$.

The number of parameters can further be reduced by decreasing the number of connections between layers (sparse connections) or by sharing parameters between units. A possible way to achieve this is to consider prior knowledge about dependencies between input units or model parameters, and thereby adjust the layer structure. The most widely used type of network, which is based on this idea, is called convolutional neural network (CNN). CNNs are specifically developed for the task of image analyzes and taking the structure and properties of an image as input into account. Commonly, a CNN consists of a set of convolutional layers, which are followed by several fully-connected layers. The task of the convolutional layers is to detect and extract features from the input image (the deeper the network the more complex are the learned features), whereas the task of the fully-connected layers is to find the mapping to the output labels from the extracted features.

A convolutional layer consists of a multi-dimensional rectangular grid of neurons and a set of $s$ filters sometimes called kernels, which contain the weights of the network. Each filter $\mathbf{k}^{(t,s)}$ has a size of $(2r^{(t)}+1) \times (2r^{(t)}+1)$. In contrast to fully-connected layers, only a small region in layer $t$ is connected to a neuron in layer $t+1$ (sparse connected layers). The region in the input image $I$, which is indirectly connected to a neuron (unit) $q$ in layer $t$ is called the receptive field. The size of the receptive field depends on the size of the applied filter and increases with depth. An example of convolutional layers and the receptive field of a point $q$ is illustrated in Figure 2.17.

The concept of sparse connected layers is reflected in the computation of the activations $\mathbf{a}^{(t+1)}$. Given as an example the 3 dimensional input $\mathbf{a}^{(t)}$, where the third (spectral) dimension represents the different channels, the activation $\mathbf{a}^{(t+1)}_{i,j,c_{t+1}}$ for a unit at position $(i,j)$ of channel

$c_{t+1}$ in layer $t$ is given as

$$\mathbf{a}_{i,j,c_{t+1}}^{(t+1)} = \sigma\left(\left(\mathbf{a}_{c_t}^{(t)} * \mathbf{k}^{(t,s)}\right)_{i,j} + \mathbf{b}_{i,j,c_t}^{(t)}\right) = \sigma\left(\sum_{u=-r}^{r}\sum_{v=-r}^{r} \mathbf{a}_{i-u,j-v,c_t}^{(t)} \mathbf{k}_{u,v}^{(t,s)} + \mathbf{b}_{i,j,c_t}^{(t)}\right), \quad (2.21)$$

where $*$ denotes the convolution operator. The matrix $\mathbf{a}_{c_{t+1}}^{(t+1)}$ of channel $c_{t+1}$ is called feature map and is generated by shifting the filter $\mathbf{k}^{(t,s)}$ over the input feature map $\mathbf{a}_{c_t}^{(t)}$. The length of these shifts is called stride and can vary from one to several units. If the input matrices are not padded with zeros around the boundaries (zero-padding) the size of the feature maps will decrease with the depth of the network. In practice, a set of up to several hundred filter is applied to each channel of the input of layer $t$, where every filter learns to extract a particular feature and produces a corresponding feature map.

In contrast to fully connected layers, where every connection has a distinct weight, convolution layers utilize only one filter (one set of weights) to generate one feature map. The idea behind this concept of parameter sharing is that a particular feature occur at several locations within the image and can always be extracted with the same filter. Parameter sharing and sparsely connected layers tremendously decreases the number of parameters of a network and have been proven to be a very efficient regularization technique in practice. The number of parameter in a CNN can be further decreased by inserting pooling layers between the convolutional layers. The aim of pooling layers is to reduce the dimensionality of the feature maps and to make the network invariant to local translations. Common used pooling operators are max pooling [66] or average pooling, where the features maps are downsampled by taking the maximum or the average value of a defined region. Pooling layers are useful for tasks, such as object recognition, where the exact location of the object in the image is less important. For further regularization strategies we refer to [48].

**Neural Networks - Types of Application Possibilities**

The ways of deploying neural networks for particular task can be categorized into three strategies: The first strategy is to design a network particular for a specific task and train it from scratch given a labeled dataset. This strategy is only applicable if a large dataset is available. The second strategy is based on the idea of transfer learning, which means to use knowledge gained from one problem to solve a similar one. The idea is to utilize layers from an already trained network (trained for a different task) and combine them with new and untrained layers. Commonly, only the last few layers of pre-trained models are exchanged. For the subsequent training the old weight can either be fixed and only the weights of the new layers are adjusted, or all weights are adjusted (fine-tuning). The last strategy is to use the features learned by a neural network and feed it as input to a different machine learning algorithm such as SVMs or decision trees. In this thesis we will follow the first strategy and always train the networks from scratch.

### 2.2.2 Generative Adversarial Networks

Generative adversarial networks (GANs) are a new machine learning architecture introduced by Goodfellow et al. [67] in 2014. The concept of GANs earned a lot of attention in the field of machine learning and offers new possibilities for several research problems through the generation of high quality samples. The application fields of GANs range from computer vision problems, e.g. semantic segmentation [68], single image super-resolution [69], text to image synthesis [70], to the problem of discovering new drugs for specific diseases in the field of medicine [71]. A further use case in medicine is the utilization of GANs for the generation of computer tomography (CT) images from magnetic resonance imaging (MRI) to reduce the radiation exposure to patients during acquisition [72]. In the context of remote sensing, Guo et al. [73] investigated the application of GANs for the synthesis of SAR image patches.

GANs belong to the class of generative models and pursue the goal of learning the data distribution of a given dataset, commonly images, in order to generate new data from the learned distribution. As described in Subsection 2.2 generative models are difficult to learn and come with high computational costs. Moreover, commonly no labeled datasets exist for these kinds of tasks, which means that GANs initially deal with an unsupervised learning problem. Despite these problems, the development of generative models and especially GANs made decisive progress over the last years. The success originates from the idea of reformulating the image generation problem and to integrate it into a newly defined task, which can be learned through supervised learning techniques. Through this strategy, the original task will be learned as a by-product and difficulties of unsupervised learning can be avoided. The only additional requirement is a labeled dataset in order to realize the supervised learning task.

In the specific case of GANs the learning procedure is realized through an adversarial process, which is based on the simultaneous training of two counteractive neural networks, the generator $G$ and the discriminator $D$. The overall (unsupervised) goal is to train the generator network $G$ to map random noise $\boldsymbol{z}$ to output images $\boldsymbol{y}$ (the artificial image samples). More specific, the aim of $G$ is to estimate the real data distribution $p_{\text{data}}$ of a given training dataset $\mathcal{D}_{\text{train}}$ as good as possible and to generate artificial images samples from the learned distribution. In order to reach this goal, a discriminator network $D$ is added to the system for the time of the training. This network is trained through supervised learning techniques, where the goal of $D$ is to distinguish as good as possible between real images $\boldsymbol{y}$ and images $\widetilde{\boldsymbol{y}} = G(\boldsymbol{z})$ generated by $G$. The training objective of $G$ on the other hand is reformulated to the goal of producing more and more realistic images to "fool" $D$ as often as possible. The described problem can be expressed through a two-player minimax game

$$\min_G \max_D \mathcal{L}_{\text{GAN}}(G, D) = \min_G \max_D \Big( \overbrace{E_{\boldsymbol{y} \sim p_{\text{data}}(\boldsymbol{y})}[\log D(\boldsymbol{y})]}^{\text{predicted log probability of } D \text{ that } y \text{ is real}} + \underbrace{E_{\boldsymbol{z} \sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))]}_{\text{predicted log probability of } D \text{ that } G(z) \text{ is fake}} \Big), \tag{2.22}$$

where $E$ denotes the expected value, $p_{\text{data}}$ the real data distribution and $p_{\boldsymbol{z}}$ a noise distribution, e.g. a uniform or normal distribution. $D$ is commonly realized by a binary classification

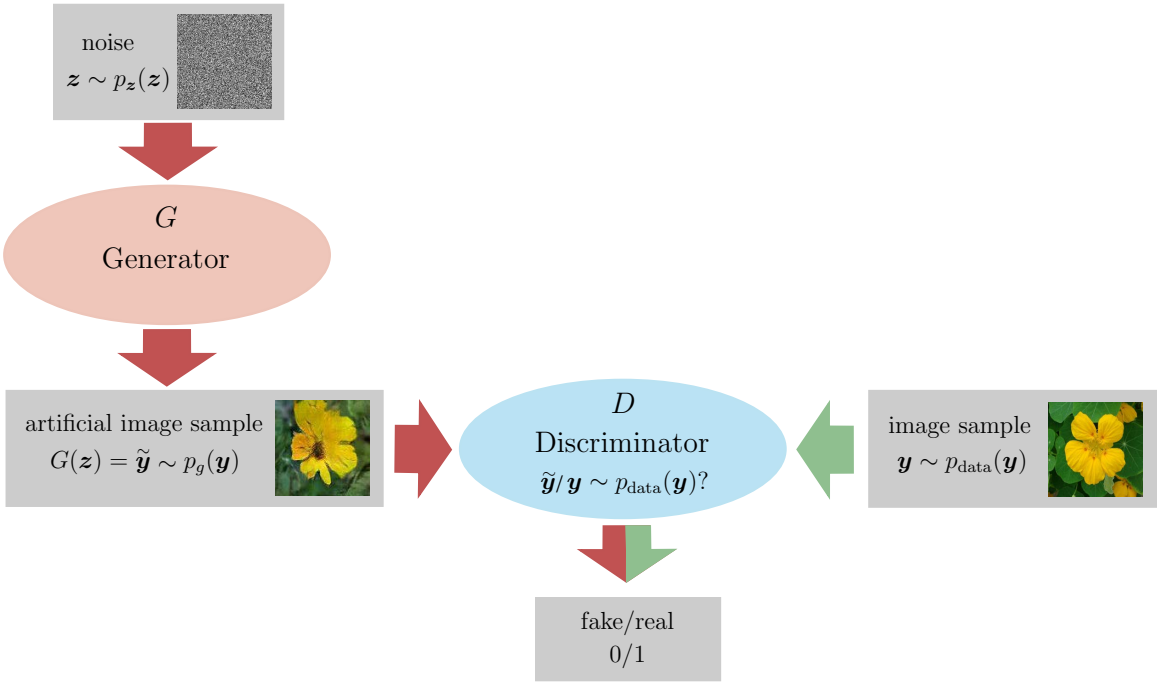



**Figure 2.18:** Illustration of the general GAN concept. The task of the generator network $G$ is to produce artificial image samples $y = G(z)$ from random noise samples $z$ with a distribution $p_g(y)$ as close as possible to the real data distribution $p_{\text{data}(y)}$. The task of the discriminator network $D$ is to distinguish as good as possible between real image samples $y \sim p_{\text{data}(y)}$ and artificial generated samples $\widetilde{y} \sim p_g(y)$.

network and outputs the probability that an input image belongs either to the class 0 ("fake") or to the class 1 ("real"). To ensure that the output values of $D$ lie in the range of $[0, 1]$ a sigmoid layer can be used as the last layer of $D$. Generally, the network architecture of $D$ and $G$ underlies not many restrictions and a variety of different designs find a practical application. Note that after the training phase $D$ will be neglected and only the learned skills of $G$ will be evaluated during the test phase.

In practice, the two networks are trained at the same time by alternating the training of $D$ and $G$, by first maximizing the GAN loss with respect to the discriminator parameters $\boldsymbol{\theta}^{(D)}$ and then minimizing the same loss with respect to the generator parameters $\boldsymbol{\theta}^{(G)}$. The intuition behind these two steps is that $D$ tries to get $D(G(\boldsymbol{z}))$ close to 0, which means to detect all images generated by $G$ and correctly label them as "fake". In contrast, $G$ aims to get $D(G(\boldsymbol{z}))$ close to 1, which means that $D$ does not identifies the artificial images generated by $G$ and wrongly label them as "real". Due to this learning strategy $G$ will consequently learn an estimation of the real data distribution $p_{\text{data}}$. The discriminator network $D$ is trained on two different kinds of training image samples. Half of the training samples are "fake" examples, generated by $G$, and the other half are "real" examples from the training dataset $\mathcal{D}_{\text{train}}$. These two cases are represented by the red and green arrows in Figure 2.18, which illustrated the overall concept of GANs.

Commonly, gradient descent is utilized to optimize the GAN loss from Equation 2.22, where one gradient descent step of $D$ is followed by one gradient descent step of $G$. There exist several algorithms to optimize gradient descent, such as [63] or [62], with the aim of finding

the optimal update rule for the parameters $\boldsymbol{\theta}^{(D)}$ and $\boldsymbol{\theta}^{(G)}$ of $D$ and $G$, respectively. The specific algorithms utilized in this thesis will be discussed in detail in Sections 4.2 and 4.3.

A frequent problem at the beginning of the training procedure is the low quality of the image samples generated by $G$. Due to the low image sample quality, $D$ will quickly learn to distinguish artificially generated images from real image samples, which means that $D(G(\boldsymbol{z}))$ is close to zero. As a consequence, $\log(1 - D(G(\boldsymbol{z})))$ will be close to 0, and hence the loss of the generator and its gradients are close to 0. Small gradients hamper the generator from learning and extremely slow down the training procedure. This problem is often referred to as vanishing gradients problem. To avoid vanishing gradients of $G$ a common course of action is to minimize $-\log(D(\boldsymbol{y}, G(\boldsymbol{z})))$ instead of $\log(1 - D(\boldsymbol{y}, G(\boldsymbol{z})))$ with respect to $\boldsymbol{\theta}^{(G)}$ (the first term of $\mathcal{L}_{\text{GAN}}$ is independent from $G$). This change provides stronger gradients for $G$, even if $D(G(\boldsymbol{z}))$ is small, while providing the same optimum. Note, this change is practical motivated to ensure strong gradients for both networks to facilitate the learning process. In the later theoretical discussion of the GAN concept the original generator loss will be considered. A summary of the described GAN training procedure is given in Algorithm 1.

Goodfellow et al. [67] built the described learning strategy on the basis of a theoretical analysis, which shows that through the application of Algorithm 1, $p_g$ converges to $p_{\text{data}}$ if optimal training conditions are given (high enough model complexity and training time). In the following, we will provide an overview of the most important theoretical results and refer for more details and full proofs to [67].

**Algorithm 1:** GAN training procedure with stochastic gradient descent.

**Input:** A training dataset $\mathcal{D}_{\text{train}}$, a noise distribution $p_{\boldsymbol{z}}$, the learning rate $\lambda$, the batch size $N_b$, the number of training iteration $n_{\text{train}}$

**for** $i = 1, ..., n_{\text{train}}$ **do**

- Sample a mini-batch $\left\{\boldsymbol{z}^{(i)}\right\}_{i=1}^{N_b}$ from the noise distribution $p_{\boldsymbol{z}}$ and a mini-batch $\left\{\boldsymbol{y}^{(i)}\right\}_{i=1}^{N_b}$ from the set of real training data $\mathcal{D}_{\text{train}}$ with distribution $p_{\text{data}}$
- Compute the stochastic gradient $\boldsymbol{g}^{(D)}$ of $D$ w.r.t. its parameters $\boldsymbol{\theta}^{(D)}$: $\boldsymbol{g}^{(D)} \leftarrow \nabla_{\boldsymbol{\theta}^{(D)}} \frac{1}{N_b} \sum_{i=1}^{N_b} \left[\log D\left(\boldsymbol{y}^{(i)}\right)] + \log\left(1 - D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right)\right]$
- Update the parameters of $D$ via an optimization algorithm (OptAlg): $\boldsymbol{\theta}^{(D)} \leftarrow \boldsymbol{\theta}^{(D)} + \lambda \text{ OptAlgo}(\boldsymbol{\theta}^{(D)}, \boldsymbol{g}^{(D)})$

- Sample a mini-batch $\left\{\boldsymbol{z}^{(i)}\right\}_{i=1}^{N_b}$ from the noise distribution $p_{\boldsymbol{z}}$
- Compute the stochastic gradient $\boldsymbol{g}^{(G)}$ of $G$ w.r.t. its parameters $\boldsymbol{\theta}^{(G)}$: $\boldsymbol{g}^{(G)} \leftarrow \nabla_{\boldsymbol{\theta}^{(G)}} - \frac{1}{N_b} \sum_{i=1}^{N_b} \left[\log\left(D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right)\right]$
- Update the parameters of $G$ via an optimization algorithm (OptAlg): $\boldsymbol{\theta}^{(G)} \leftarrow \boldsymbol{\theta}^{(G)} + \lambda \text{ OptAlgo}(\boldsymbol{\theta}^{(G)}, \boldsymbol{g}^{(G)})$

**end**

If we assume a continuous space and utilizing the definition of the expected value, the GAN loss from Equation 2.22 can be rewritten as follows

$$\begin{aligned}\mathcal{L}_{\text{GAN}}(G,D) &= \int_{\boldsymbol{y}} p_{\text{data}}(\boldsymbol{y}) \log\left(D(\boldsymbol{y})\right) d\boldsymbol{y} + \int_{\boldsymbol{z}} p_{\boldsymbol{z}}(\boldsymbol{z}) \log\left(1 - D(G(\mathbf{z}))\right) d\boldsymbol{z} \\ &= \int_{\boldsymbol{y}} p_{\text{data}}(\boldsymbol{y}) \log\left(D(\boldsymbol{y})\right) + p_g(\boldsymbol{y}) \log\left(1 - D(\boldsymbol{y})\right) d\boldsymbol{y}\,.\end{aligned} \tag{2.23}$$

Since a function of the form $f(x) = a\log(x) - b\log(1-x)$ has its maximum at $\frac{a}{a+b}$ for $x \in [0,1]$ and $(a,b) \in \mathbb{R}^2 \setminus \{0,0\}$, the optimal solution of $\max_D \mathcal{L}_{\text{GAN}}(G,D)$ is $D_G^*(\boldsymbol{y}) = \frac{p_{\text{data}}(\boldsymbol{y})}{p_{\text{data}}(\boldsymbol{y}) + p_{\text{g}}(\boldsymbol{y})}$ given a fixed generator network $G$. Furthermore, maximizing the GAN loss with respect to $D$ is equivalent to maximizing the log-likelihood of $D$ predicting the correct label to a training sample, whereas maximizing the GAN loss with respect to $G$ is equivalent to minimizing the same log-likelihood. By further assuming an optimal trained discriminator network $D_G^*$, we can reformulate Equation 2.22 as follows

$$\begin{aligned}&\min_G \max_D \mathcal{L}_{\text{GAN}}(G,D) = \\ &\quad= \min_G \underbrace{\left(\max_D \mathcal{L}_{\text{GAN}}(G,D)\right)}_{=\mathcal{L}_{\text{GAN}}(G,D_G^*)} \\ &\quad= \min_G E_{\boldsymbol{y}\sim p_{\text{data}}}\left[\log D_G^*(\boldsymbol{y})\right] + E_{\boldsymbol{z}\sim p_{\boldsymbol{z}}}\left[\log(1 - D_G^*(G(\boldsymbol{z})))\right] \\ &\quad= \min_G E_{\boldsymbol{y}\sim p_{\text{data}}}\left[\log D_G^*(\boldsymbol{y})\right] + E_{\boldsymbol{z}\sim p_{\boldsymbol{z}}}\left[\log(1 - D_G^*(\boldsymbol{y}))\right] \\ &\quad= \min_G E_{\boldsymbol{y}\sim p_{\text{data}}}\left[\log \frac{p_{\text{data}}(\boldsymbol{y})}{p_{\text{data}}(\boldsymbol{y}) + p_{\text{g}}(\boldsymbol{y})}\right] + E_{\boldsymbol{z}\sim p_{\boldsymbol{z}}}\left[\log \frac{p_{\text{g}}(\boldsymbol{y})}{p_{\text{data}}(\boldsymbol{y}) + p_{\text{g}}(\boldsymbol{y})}\right] \\ &\quad= \min_G E_{\boldsymbol{y}\sim p_{\text{data}}}\left[\log \frac{p_{\text{data}}(\boldsymbol{y})}{2} \cdot \frac{2}{p_{\text{data}}(\boldsymbol{y}) + p_{\text{g}}(\boldsymbol{y})}\right] + E_{\boldsymbol{z}\sim p_{\boldsymbol{z}}}\left[\log \frac{p_{\text{g}}(\boldsymbol{y})}{2} \cdot \frac{2}{p_{\text{data}}(\boldsymbol{y}) + p_{\text{g}}(\boldsymbol{y})}\right] \\ &\quad= \min_G -2\log(2) + D_{\text{KL}}\left(p_{\text{data}}(\boldsymbol{y}) \,\middle\|\, \frac{p_{\text{data}}(\boldsymbol{y}) + p_{\text{g}}(\boldsymbol{y})}{2}\right) + D_{\text{KL}}\left(p_{\text{g}}(\boldsymbol{y}) \,\middle\|\, \frac{p_{\text{data}}(\boldsymbol{y}) + p_{\text{g}}(\boldsymbol{y})}{2}\right) \\ &\qquad\uparrow \\ &\text{def. KL divergence} \\ &\quad= \min_G -2\log(2) + 2D_{\text{JS}}(p_{\text{data}}(\boldsymbol{y}) \,\|\, p_{\text{g}}(\boldsymbol{y}))\,, \\ &\qquad\uparrow \\ &\text{def. JS divergence}\end{aligned} \tag{2.24}$$

where the Kullback-Leibler (KL) divergence $D_{\text{KL}}$ is defined as

$$D_{\text{KL}}(p(\boldsymbol{y}) \,\|\, q(\boldsymbol{y})) = E_{\boldsymbol{y}\sim p(\boldsymbol{y})}\left[\log\left(\frac{p(\boldsymbol{y})}{q(\boldsymbol{y})}\right)\right] \tag{2.25}$$

and the Jensen-Shannon (JS) divergence $D_{\text{JS}}$ as

$$D_{\text{JS}}(p(\boldsymbol{y}) \,\|\, q(\boldsymbol{y})) = \frac{1}{2}D_{\text{KL}}\left(p(\boldsymbol{y}) \,\middle\|\, \frac{p(\boldsymbol{y}) + q(\boldsymbol{y})}{2}\right) + \frac{1}{2}D_{\text{KL}}\left(p(\boldsymbol{y}) \,\middle\|\, \frac{p(\boldsymbol{y}) + q(\boldsymbol{y})}{2}\right). \tag{2.26}$$

The JS and KL divergence are two different methods for measuring the similarity between two probability distributions $p$ and $q$. As Equation 2.24 reveal, minimizing the function

$\max_D \mathcal{L}_{\text{GAN}}(G, D)$ with respect to the generator parameters is equivalent to minimizing the Jensen-Shannon divergence between $p_{\text{data}}$ and $p_{\text{g}}$, given an optimal discriminator $D^*_G$. Minimizing the JS divergence between $p_{\text{data}}$ and $p_{\text{g}}$ is equivalent to gradually increasing the similarity between both distributions. It can be further shown, that the global minimum of $\max_D \mathcal{L}_{\text{GAN}}$ is achieved if and only if $p_{\text{g}} = p_{\text{data}}$ (see [67]). Additionally, this optimal solution can be found through the application of Algorithm 1 if the model complexity of $D$ and $G$ is high enough and the discriminator is trained long enough to reach its optimum given $G$. In contrast to this theoretical results, where $D$ should be trained until the optimal solution, given $G$, is found (one gradient descent step of $G$ is followed by many of $D$), it has been proven in practice to be more efficient to train $G$ and $D$ alternately, where one gradient descent step of $G$ is followed by one of $D$.

We introduced the overall training problem of Equation 2.22 as a two-player minimax game. In game theory the optimal solution of such a problem is called Nash Equilibrium. Based on the theory presented in [74], a Nash Equilibrium in the context of GANs can be defined as a tuple $(\theta_D, \theta_G)$ that is a local maximum of $\mathcal{L}_{\text{GAN}}$ with respect to $\theta_D$ and a local minimum of $\mathcal{L}_{\text{GAN}}$ with respect to $\theta_G$. At this points the both players $G$ and $D$ reach there optimal strategy. In relation to the above discussion this point is reached if $p_{\text{g}} = p_{\text{data}}$.

A drawback of GAN training in comparison to the training of a common neural network is that finding the Nash equilibrium is commonly more difficult than optimizing a loss function. On the other hand GANs have the advantage of not tending to overfit. Since the generator is getting information about the training data only indirect through the discriminator it cannot learn to just replicate the training samples. In Section 4.2 we will introduced some GAN variants, further discuss some advantages and disadvantages and introduce the specific network architecture of $G$ and $D$ applied in this thesis.

## 2.3 Summary

The theoretical concepts presented in this chapter lay the foundation for the multi-modal image registration methods presented later in this thesis. In summary the following aspects have been discussed in the context of optical and SAR imagery in Section 2.1:

- Optical and SAR satellites are built on different acquisition concepts (synthetic aperture with distance measurements in SAR; perspective projection in optical), viewing perspectives (respectively off-nadir and usually near-nadir) and utilize different wavelengths (respectively cm and nm) for the acquisition of images.

- The use of different wavelengths lead to different radiometric properties in the optical and SAR images, as the response of an object depends on the signal properties (wavelength, polarization), the surface properties (roughness, randomness of local reflectors and reflectance properties) and sensor perspective. The speckle effect in SAR images further complicates the human and automatic image interpretation.

- The different image acquisition principles also affect the geometry of the observed objects. In particular, the sideways-looking acquisition of SAR sensors introduces typical geometric distortion effects (layover, foreshortening) and shadowing for 3D objects such as buildings or trees. These effects have a strong influence on the appearance of all objects above the ground level in SAR images. As a consequence, the boundary of an elevated object in a SAR image does not fit the object boundary in the optical image, even if the imaging perspective is the same for both sensors.

- The differences in image acquisition further affect the geo-referencing process of the images. Optical satellite images commonly have a geo-localization accuracy in the order of tens of meters only, due to inaccurate measurements of the attitude angles in space. High-resolution SAR images such as TerraSAR-X images on the other hand, exhibit an absolute geo-localization accuracy within a few decimeters.

In the context of supervised machine learning (see Section 2.2) the most important aspects can be summarized as follows:

- Depending on the type of learning, machine learning algorithms can be broadly divided into three categories: supervised, unsupervised and reinforcement learning.

- The focus of this thesis is on the application of neural networks trained through supervised learning algorithms. Such learning algorithms pursue the goal of learning a mapping from input data $X$ to a set of corresponding labels $Y$ (labeled data).

- The learning process of the networks is thereby divided into three phases, training, validation and test, where each phase pursues a certain goal and therefore requires an independent dataset.

- For an efficient training of neural networks the so called backpropagation algorithm in combination with regularization techniques and adjusted network architectures, such as usage of convolutional layers, is suggested.

- Through the design of a suitable network architecture and training procedure, neural networks are able to learn all kind image features and can therefore be used for a variety of tasks such as the detection and extraction of objects or the classification of images.
- A novel machine learning architecture is the so called generative adversarial network (GAN), which pursues the goal of learning the data distribution of a given dataset, commonly images, in order to generate new data from the learned distribution.
- GANs are trained through an adversarial training process, which is based on the training of two counteractive neural networks, the generator and the discriminator.
- Through the specific network architecture and particular training process the generation of high quality image samples from noise became feasible and open ups new possibilities for the generation of artificial images.

# 3

# Image Registration

*This chapter briefly describes the principles of image registration and its application in the context of multi-modal images. Furthermore, state-of-the-art concepts for the problem of optical and SAR image registration are introduced, and their advantages and disadvantages discussed and summarized. In contrast to traditional image registration methods, typically used in remote sensing, we introduce two novel image matching concepts, based on deep learning techniques, which provides new opportunities for the improvement of multi-modal image registration. The chapter is concluded by a short summary about previous research studies and the research gaps regarding the image matching of optical and SAR satellite images.*

## Contents

Image registration has applications in various fields such as medical imaging, computer vision and remote sensing. Commonly, image registration techniques are required, if information from multi-model, multi-temporal or multi-viewpoint images has to be compared on a point-to-point basis or combined in order to fuse information, to find changes or to derive three-dimensional information. Zitová and Flusser [75] roughly divided image registration problems into four groups (according to the image acquisition mode): Registration of images acquired 1) from different viewpoints, 2) at different times, 3) from different sensors and, 4) the registration of images and models of a scene. Besides this separation, image registration techniques can also be categorized with respect to other aspects such as the transformation models, which align one image with another, or based on the frameworks utilized for the detection of correspondences between the different images. However, due to the circumstance that registration techniques are commonly developed for a specific kind of application and not for a specific problem (e.g. multi-temporal or multi-modal image alignment), it is difficult to assign the different image registration techniques only to one class (in relation to one of these categorization types). The focus of this thesis is on the registration of multi-modal image data, which are acquired at varying times. In particular, we focus on the development of concepts for the registration of image pairs from optical and SAR satellites.

Accurate geo-referenced and precisely co-registered optical and SAR image pairs are a prerequisite for any image fusion application such as earthquake damage assessment of buildings [5], road network extraction [6] or change detection [8] (further details presented in Section 1.2). Commonly such data are not available, entailing the need of image registration techniques. Besides the usage for image fusion tasks, image registration of high-resolution optical and SAR image pairs has an additional benefit. Assuming the case of multi-modal image data, where one of the images exhibits a higher absolute geo-localization accuracy, image registration techniques can further be employed to improve the localization accuracy of the second image. As discussed in Subsection 2.1.2, high-resolution SAR satellites like TerraSAR-X exhibit an absolute geo-localization accuracy in the range of a few decimeters or centimeter for specific targets [14], whereas high-resolution optical sensors still require ground control points (GCPs) to reach similar accuracies.

Therefore, the successful registration of optical and SAR images has two major advantages: first, the provision of aligned multi-modal image data as a foundation for several image fusion applications and second, the provision of accurate absolute geo-referenced optical images (under the assumptions of utilizing high-resolution SAR images for the registration process). The main application aimed in this thesis is the geo-localization accuracy enhancement of optical images through the registration with SAR images. Therefore, we will provide a brief introduction to image registration principles in Section 3.1. Section 3.2 gives an overview of state-of-the-art optical and SAR image registration techniques and a discussion about open problems and challenges of these methods. A novel concept for the registration of images, which represents a promising alternative to traditional methods by providing solutions for several common challenges, is introduced and discussed in Section 3.3. We will conclude this chapter with a summary of the research gaps regarding the problem statement and research questions of this thesis in Section 3.4.

## 3.1 Principles of Image Registration

Zitová and Flusser [75] defined image registration as the process of precisely overlaying two or more images of the same scene taken at different times, from different viewpoints, and/or by different sensors. This process geometrically aligns two images, by transforming an input images (or sensed image) to a reference image. A slightly different definition can be found in [76], where image registration is defined as the procedure to determine the best spatial fit between two or more images of the same scene by geometrical matching of two or more images acquired with the same or different sensor, with or without the same ground resolution or at the same or different time. Mathematically, image registration between an input image $\boldsymbol{I}$ and a reference image $\boldsymbol{R}$ can be stated as the following equation

$$\boldsymbol{R}(x, y) = g(f(\boldsymbol{I}(u, v))) \,, \tag{3.1}$$

where $(u, v)$ are coordinates in the image space of $\boldsymbol{I}$, $(x, y)$ coordinates in the image space of $\boldsymbol{R}$, $f$ a spatial transformation function (e.g. translation, affine or polynomial function) and $g$ a radiometric interpolation function to resample the transformed input image $\boldsymbol{I}$ to $\boldsymbol{R}$ (e.g. nearest neighbor, bilinear or cubic interpolation). A visualization of Equation 3.1 is presented in Figure 3.1.

The overall goal of each image registration method is to find the optimal transformation functions $f$ and $g$ to precisely align $\boldsymbol{I}$ and $\boldsymbol{R}$. In order to achieve this goal a suitable registration framework has to be developed, which takes the particular properties of the input-reference image pair into account. Commonly, every image registration framework consists of the following four essential steps:

1. **Feature detection and extraction:** Prominent and salient image features such as corners, line intersections or small image patches containing distinctive structures, are detected independently in the spatial or frequency domain of the input and reference image. The important information about the detected features is subsequently extracted from the images and often represented in form of feature descriptors.

2. **Feature matching:** In order to identify corresponding matching or tie points in the input and reference image the extracted features are matched through the application of a suitable matching approach (feature- or intensity-based).

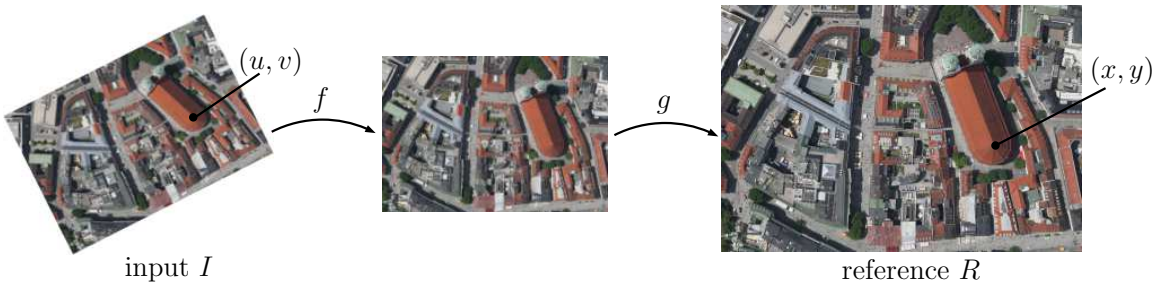



**Figure 3.1:** Illustration of the image registration process. The input image $\boldsymbol{I}$ is mapped to the reference image $\boldsymbol{R}$ by applying a spatial transformation $f$ and a radiometric transformation $g$.

3. **Transformation model estimation:** Based on the obtained image correspondences in form of tie points and eventually existing knowledge about the specific kind of distortion (e.g. none, global or local) between the input and the reference image the optimal type of the transformation model and its parameters are estimated.

4. **Image transformation and resampling:** The input image is transformed based on the computed transformation function and resampled through the application of a suitable interpolation method in order to be precisely aligned with the reference image.

The four steps are usually well adapted to fulfill the needs and requirements of a particular application, given an available image dataset, and can therefore vary quite a lot between different registration approaches. Each step of the registration framework pursues a particular goal and has to overcome certain challenges. The most challenging and most widely researched part of the image registration framework are the first two steps, which also form the focus of this thesis.

The purpose of the feature detection, extraction and matching steps is to link the input and reference image by generating a set of tie points, sometimes called matching points. Tie points are points, which represent the same locations in the input and in the reference image and are utilized to estimate the type of the transformation model and its parameters. Therefore, the accuracy and precision, with which the tie points are extracted, have an enormous influence on the final registration outcome. As a consequence, the detection and extraction of a reliable set of tie points play a central role in every registration framework. In particular, if the input image exhibit local distortions, the tie points should be uniformly spread over the whole image scene. Automatic techniques for the provision of such a set of tie points roughly divides image registration techniques into two groups: feature-based and intensity-based approaches. In the following, the general idea and specific challenges of feature- and intensity-based tie point generation approaches will be discussed (see Subsections 3.1.1 and 3.1.2). In Subsection 3.1.3 we will present common practices for the determination of a suitable transformation model and resampling function utilizing the obtain tie points.

### 3.1.1 Intensity-based Tie Point Generation

The first group of automatic techniques for the generation of tie points are often called intensity- or area-based approaches. Generally, intensity-based approaches skip the feature detection and extraction step, and instead focus on feature matching without explicitly link particular features between the images. The principle idea is to define and utilize a similarity metric to measure the similarity between image regions from the input and reference image based on pixel information (e.g. intensity values in the spatial domain). Image correspondences are found by searching region pairs, which achieve the highest similarity value among all region pairs.

A common practice to find corresponding image regions is to crop small image patches, often called templates, from the input image around a regular grid of location (see left side of Figure 3.2). Afterwards, the similarities between the templates and image regions from the reference image are computed. Without additional information about the specific kind of

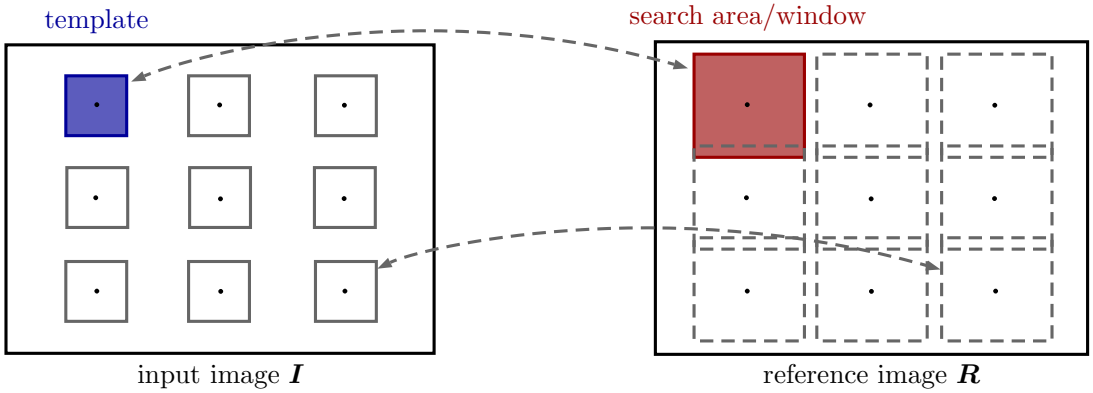



**Figure 3.2:** Illustration of a search strategy to find correspondences between images within an intensity-based matching framework. Templates are cropped around a regular grid of locations from the input image $\boldsymbol{I}$. For every template a search areas (windows) around the same locations in the reference image $\mathbf{R}$ is defined. The search areas can be adjusted and reduced in size by taking additional information about the image distortion between both images into account.

distortion, the whole reference image has to be searched in order to find a corresponding image region for every template. If additionally information exist, it can be used to adjust and limit the search areas within the reference images. In the case of registering orthorectified high-resolution optical and SAR image it can be assumed that there is only a local offset of less than one hundred meters between the images. Therefore, the search space can be significantly reduced in size by determining a small search area (window) in the reference image for every extracted template from the input image (see right side of Figure 3.2).

The success of intensity-based methods heavily depends on the selected similarity metric and its ability to measure the similarity between the given images. More specific, in the case of multi-temporal and multi-modal image data the selection of suitable metrics requires particular care. In order to gain a better understanding for different types of similarity metrics, we will briefly introduced two frequently utilized metrics in the following. Both metrics will later form the basis of two baseline approaches that we will utilize to assess our results (see Subsection 5.1.4).

**Normalized Cross-Correlation:** The first similarity metric, often applied for the task of single sensor image matching, is called normalized cross-correlation (NCC). The idea of NCC is to measure the similarity of two images or image patches based on a pixelwise comparison of their intensity values. The NCC-value between a template $\boldsymbol{T}$ with size $N_x \times N_y$ cropped from $\boldsymbol{I}$ and an image patch located around the position $(m, n)$ in $\boldsymbol{R}$ is defined as follows

$$\mathrm{NCC}(m,n) = \frac{\sum_{i=1}^{N_x}\sum_{j=1}^{N_y}\Big(\boldsymbol{R}(m+i,n+j)-\overline{\boldsymbol{R}}\Big)\Big(\boldsymbol{T}(i,j)-\overline{\boldsymbol{T}}\Big)}{\sqrt{\sum_{i=1}^{N_x}\sum_{j=1}^{N_y}\Big(\boldsymbol{R}(m+i,n+j)-\overline{\boldsymbol{R}}\Big)^2\Big(\boldsymbol{T}(i,j)-\overline{\boldsymbol{T}}\Big)^2}}\,. \tag{3.2}$$

Here, $\boldsymbol{R}(m+i, n+j)$ and $\boldsymbol{T}(i, j)$ are the intensity values of $\boldsymbol{R}$ and $\boldsymbol{T}$ at the locations $(m+i, n+j)$ and $(i, j)$, respectively, $\overline{\boldsymbol{T}}$ is the mean intensity value of $\boldsymbol{T}$, and $\overline{\boldsymbol{R}}$ the mean

intensity value of the overlapping image patch between $\boldsymbol{T}$ and $\boldsymbol{R}$. The NCC-value ranges from $-1$ to 1, where a value close to 1 indicates a high similarity between the images. In order to find the image regions of $\boldsymbol{R}$, showing the highest similarity to the template $\boldsymbol{T}$, the template is moved over the corresponding search area in $\boldsymbol{R}$. The search area has a size of $(N_x + 2 * \Delta_x) \times (N_y + 2 * \Delta_y)$, where $\Delta_x$ and $\Delta_y$ is the search space in $x$- and $y$-direction. The length of one stride is $s_x$ and $s_y$ in $x$- and $y$-direction, respectively (commonly one pixel in each direction). At each position the NCC-value between the overlaying areas is computed and the position with the highest NCC-value (within the search area) is the position with the best match between $\boldsymbol{T}$ and $\boldsymbol{R}$. This procedure is often called template matching and is illustrated in Figure 3.3.

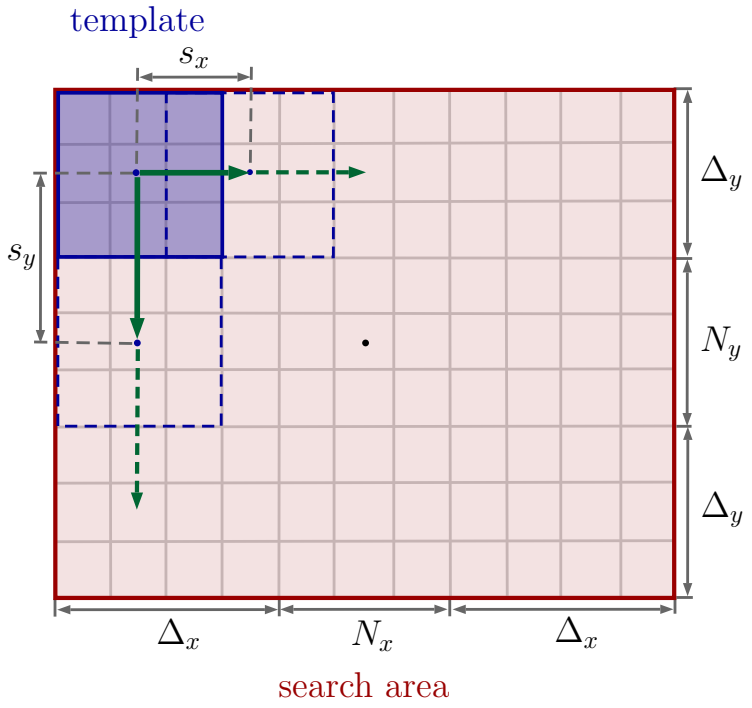



**Figure 3.3:** Illustration of an intensity-based matching between a template $\boldsymbol{T}$ and a reference image $\boldsymbol{R}$, often called template matching. The template is moved over $\boldsymbol{R}$ (within the search area) with a striding length of $s_x$ and $s_y$ in $x$- and $y$-direction, respectively. The search area has a size of $(N_x + 2 * \Delta_x) \times (N_y + 2 * \Delta_y)$ where $\Delta_x$ and $\Delta_y$ is the search space in $x$- and $y$-direction, respectively.

**Mutual Information:** A further similarity measure is mutual information (MI). In contrast to NCC, MI is measuring the similarity between images or image patches based on the comparison of their local intensity distributions. The normalized MI-value between a template $\boldsymbol{T}$ and an image patch $\boldsymbol{R}_i$ cropped from the search area in $\boldsymbol{R}$ is defined as

$$\mathrm{MI}(\boldsymbol{T}, \boldsymbol{R}_i) = \frac{H(\boldsymbol{T}) + H(\boldsymbol{R}_i)}{H(\boldsymbol{T}, \boldsymbol{R}_i)}, \tag{3.3}$$

where $H(\boldsymbol{T})$ and $H(\boldsymbol{R}_i)$ are the marginal entropies and $H(\boldsymbol{T}, \boldsymbol{R}_i)$ the joint entropy between $\boldsymbol{T}$ and $\boldsymbol{R}_i$. The marginal and joint entropy between two images $\boldsymbol{X}$ and $\boldsymbol{Y}$ are defined based on the marginal and joint probability distribution of the intensity values in the images and can be stated as follows:

$$
\begin{aligned}
H(\boldsymbol{X}) &= -\sum_{x} p_{\boldsymbol{X}}(x) \log p_{\boldsymbol{X}}(x) \\
H(\boldsymbol{X}, \boldsymbol{Y}) &= -\sum_{x,y} p_{\boldsymbol{X},\boldsymbol{Y}}(x, y) \log p_{\boldsymbol{X},\boldsymbol{Y}}(x, y) \,.
\end{aligned}
\tag{3.4}
$$

Furthermore, the marginal and joint probability distribution can be computed using the joint histogram $h(x, y)$ of the two image patches $\boldsymbol{X}$ and $\boldsymbol{Y}$:

$$
\begin{aligned}
p_{\boldsymbol{X},\boldsymbol{Y}}(x, y) &= \frac{h(x, y)}{\sum\limits_{x,y} h(x, y)} \\
p_{\boldsymbol{X}}(x) &= \sum_{y} p_{\boldsymbol{X},\boldsymbol{Y}}(x, y) \\
p_{\boldsymbol{Y}}(y) &= \sum_{x} p_{\boldsymbol{X},\boldsymbol{Y}}(x, y) \,.
\end{aligned}
\tag{3.5}
$$

The joint histogram is a two-dimensional matrix containing correspondences between the intensity values of both images. The normalized MI-value ranges from 0 to 1, where a value close to 1 indicates a high similarity between two images. As for NCC, the detection of image correspondences is based on maximizing the MI-value between image regions from the input and reference image and can be realized through the above explained template matching procedure. Additional information about MI can be found in [77, 78].

Generally, intensity-based matching approaches suffer from particular problems, which need to be taken into account when utilizing them for the tie point generation. A major drawback is the reliability of the detected correspondences. By skipping the feature detection step of the registration chain, there is a high chance that the selected templates and the corresponding search areas do not show any salient features or structures. Due to ambiguities in such homogenous areas it can happen that unrelated image areas are linked. Moreover, if the measure of the similarity is only determined based on the pixel intensity values without analyzing the structure of the values, as in the case of NCC, the results will be sensitive to noise and radiometric differences between the images. As a consequence, relevant points or areas are normally identified beforehand through feature detection algorithms such as the Förstner operator [79] or Harris corner detector [80]. Additionally, the computation of tie points through an intensity-based approach, comes commonly with high computational costs. Therefore, sophisticated search strategies have to be applied or developed to lower the computational costs and, hence, speed up the template matching without loss in accuracy.

### 3.1.2 Feature-based Tie Point Generation

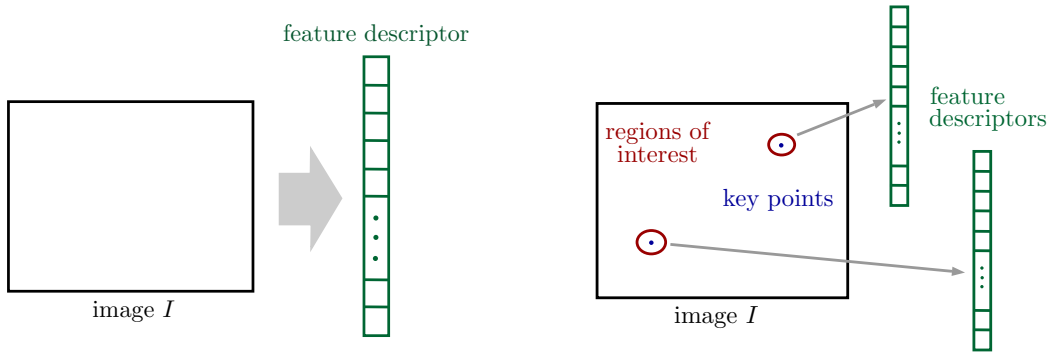
In contrast to an intensity-based tie point generation, the focus of feature-based approaches lie on the detection and matching of salient image features. An image feature is often defined as a pattern or an object that differs from its neighborhood and exhibits a salient and distinctive structure, which capture important image information [75, 81]. Features are often divided into three categorize: point features (e.g. line intersections, road crossings, corners, centroids of closed boundary regions such as centers of building roofs or inner circles of roundabouts), line features (e.g. line segments, roads, object contours, coastal lines) and areal or region features (e.g. lakes, small islands, buildings, forest, fields). The detection

and extraction of image features is an important step for several applications, e.g. image classification, object recognition or matching. Here, features are utilized to represent images either through a single feature vector representing the whole image (global features) or through smaller image regions (local features) (see Figure 3.4 for a visualization).

In order to find image correspondence and to generate a set of tie points, distinctive image features have to be detected and extracted in both images. For the particular task of registering multi-temporal and multi-modal images and to handle local distortions between the images, it is important that local features are additionally visible and detectable in both images and are constant over time. In practice man-made infrastructure, such as airports, road networks and intersections, field patterns and boarders, corners of agricultural fields, buildings and inner circles of roundabouts have proven to be reliable sources for the detection of features.

Frequently used methods for the detection of suitable features are corner detectors such as the Harris corner detectors [80], edge detectors such as Canny [82] or the Laplacian of Gaussian (LoG) [83] and segmentation techniques [84] for the detection of areas or regions. Due to the essential role of the detected features in the image registration process, a feature detector should fulfill various requirements [81]. Firstly, a feature detector should be robust against image noise and local image deformations (e.g. rotation, scaling, shifting). Secondly, the feature detection procedure should be repeatable (one detector should always detect the same features in one scene independent from viewing conditions). Lastly, the detector should be as accurate as possible in determining the exact feature locations.

Independent of the feature type (point, line, area) each detected feature can be represented by a point (end of a line, centroid of an area). These points are often called control points (CPs) or keypoints. One possible approach to match two images is to compare the spatial distribution of the detected keypoints. Another and more frequently used approach, is the matching via a symbolic feature descriptions. Therefore, local patterns (regions of interest) around each keypoint are extracted and represented by so called feature descriptors or feature vectors (see Figure 3.4). Similar to a feature detector, a feature descriptor should fulfill certain characteristics. These can be summarized according to Zitová and Flusser [75] as




(a) Global image features (b) Local image features

**Figure 3.4:** Comparison between local and global image features and a visualization of feature descriptors, regions or areas of interest and keypoints.

follows: 1) The same feature detected in the input and reference image should be represented by the same descriptors (invariance), 2) if two features are different, then the corresponding descriptors should be different as well (uniqueness), 3) small changes of a feature should lead to small changes of the descriptor (stability) and 4) if the feature is described by a vector, the elements of the vector should be functionally independent (independence). In practice, no feature detector or descriptor will fulfill all the required characteristics. Therefore, it is particular important to select feature detectors and descriptors that are suitable or adaptable to the needs of the specific application.

The next step, after obtaining a set of keypoints and associated descriptors from the input and reference image, is to identify corresponding keypoints between the images. A common approach towards this goal is to compare the distances between the feature descriptors from the input and reference image. A pair of keypoints $(p, q)$, where $p$ is a keypoint from the input image $\boldsymbol{I}$ and $q$ a keypoints from the reference image $\boldsymbol{R}$, is set to be a match, if the distance between their descriptors is the minimum among all distances between the descriptor from $p$ and all descriptors from $\boldsymbol{R}$ and the minimum among all distances between the descriptor from $q$ and all descriptors from $\boldsymbol{I}$. To find corresponding pairs among all possible pairs of keypoints and the measure the distance between descriptors, methods such as the nearest-neighbor search using the Euclidean or the Hamming distance are commonly applied [81]. An important requirement for every matching approach is to provide a reliable set of tie points. This means that keypoints should only be linked during the matching procedure, if they represent the same feature in the input and reference image.

In the following, we will briefly introduce two feature-based tie point generation approaches, the scale-invariant feature transform (SIFT) [85] and the binary robust invariant scalable keypoints (BRISK) [86]. Both methods will serve later as baselines for the assessment of the methods presented in this thesis (see Subsection 5.1.4) and will help to gain a better insight into common feature detection, description and matching techniques.

**Scale-Invariant Feature Transform (SIFT):** SIFT was introduced by Lowe [85] in 2004 and has proven to be a robust technique for the generation of tie points ever since. The four major steps of SIFT are: 1) keypoint detection, 2) keypoint localization and outlier removal, 3) orientation assignment and 4) keypoint description. In order to detect local features a space scale is constructed by convolving the image with Gaussian filters. In order to enhance the detectability of the image features, the differences of Gaussian (DoG) images are computed by subtracting the filtered images from each other. An illustration of the procedure is shown in Figure 3.5. This process is followed by the detection of local extrema over scale and space by comparing neighboring pixels within a scale and between the adjacent scales (next and previous scale). A pixel with a larger or smaller value compared to all of its neighbors is set to be keypoint candidate. The location of the keypoint candidates is determined with sup-pixel accuracy by utilizing the interpolation technique described in [87]. This method is based on the idea of fitting a quadratic surface to the neighborhood of each keypoint and computing the peak of the surface. Afterwards, unstable keypoints (with low contrast or located on edges) are removed by performing two threshold-based stability checks. The next step is to determine one or more dominant orientations for each keypoint. Therefore, the gradient magnitude and orientation of all pixels around a keypoint

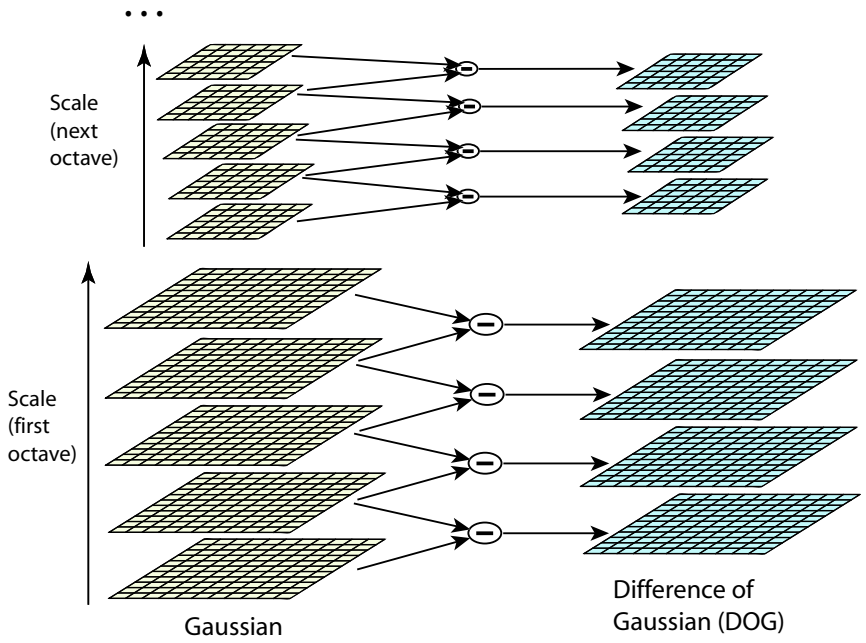



**Figure 3.5:** Illustration of the scale space and the computation of the differences of Gaussian (DoG) (image source: [85]).

are computed. Then, a histogram over all orientations (surrounding the keypoint) and weighted by their gradient magnitudes is computed. The most dominant orientations of each keypoint are determined by the highest peak and all peaks larger than 80% of the highest peak of the corresponding histograms. Finally, the set of detected keypoints is described by the keypoint descriptors. Therefore, a $16 \times 16$ descriptor window around each keypoint is selected and the gradient magnitudes and orientations are computed. To increase the influence of gradients close to the keypoints the gradient magnitudes within the descriptor windows are weighted by a Gaussian kernel. Additionally, the gradient magnitudes and coordinates are rotated with respect to the determined keypoint orientations in order to achieve rotation invariance. Then, the descriptor windows are divided into sixteen $4 \times 4$ sub-windows and for each of these sub-windows an 8 bin orientation histogram is computed. The histogram values of the sixteen sub-windows form the final feature descriptor, which is represented by a vector with 128 elements. Even though the main objective of the SIFT operator is to detect and describe local keypoints, Lowe proposed to match the descriptors based on searching the minimum of the Euclidean distances between the descriptors [85].

**Binary Robust Invariant Scalable Keypoints (BRISK):** The BRISK algorithm was introduced by Leutenegger et al. [86] in 2011 and can be divided in the same four steps as SIFT (keypoint detection, keypoint localization and outlier removal, orientation assignment and keypoint description). The first step of the algorithm is to create a scale space that consist of $n$ octaves $c_i$ and $n$ intra-octaves $d_i$. The octaves are generated by half-sampling the previous octave staring from the input image, whereas the intra-octaves are generated by half-sampling the previous intra-octave starting from $d_0$, which is the input image downsampled by a factor of 1.5. Subsequently, the FAST 9-16 detector from [88] is applied in order to select keypoint candidates. To remove outliers from the keypoint candidates, non-maxima suppression is

performed within the scale-space pyramid. In order to determine the sub-pixel locations of the maximum in each layer, a 2D quadratic function is fit to the $3 \times 3$ patch surrounding the remaining keypoints in the corresponding and the adjacent layers. For each keypoint, the tree corresponding maxima are then interpolated applying a 1D parabola across the scale-space and the 3D maximum determined. The first step of the orientation assignment is the computation of sampling patterns around the detected keypoints. The points of the patterns are subsequently paired and divided into short- and long-distance pairs depending on the distance between points (shorter or longer than a certain threshold). Additionally, the gradients between the long-distance pairs are computed and the sum of the gradients used to determine the orientation of the keypoints. The computed orientations are then applied to rotate the short-distance pairs and used in a final step to construct the binary descriptor of each keypoint.

In practice, feature-based approaches are preferred over intensity-based ones, when image properties are better represented by structural information than by image intensities [75] and when computational costs should be kept to a minimum. However, feature-based tie point generation approaches have to overcome certain problems in order to generate a reliable set of tie points, especially when dealing with multi-modal or multi-temporal image data. If the images additionally exhibit local distortions, the two sets of extracted features from input and reference image should have enough common elements in order to find a proper amount of tie points. Ideally, the tie points are spread over the whole image scene to enable an accurate transformation model estimation. As a consequence, the feature detection, extraction and matching techniques have to be accurate and should not be sensitive to noise, intensity changes or to radiometric differences between the images.

### 3.1.3 Transformation Model Estimation and Image Alignment

After generating a set of reliable tie points between the images, a suitable type of transformation model has to be chosen and its parameters have to be estimated. The task of the transformation function, sometimes called mapping function, is to map the input image onto the reference image. The proper type of transformation function depends on the kind of image acquisition processes, the type of the assumed or known geometric deformation between the images and on the type of application and its required accuracy. In cases, where the image acquisition processes and the cause of the geometric distortions are not known, an empirical transformation model has to be used, where the transformation model parameters are estimated from the set of detected tie points. Such transformation models can be roughly divided into two categories: global and local transformation models.

Global transformation models utilize a single transformation function for the mapping of the entire image, where the function parameters are estimated from the whole set of tie points. Commonly applied global transformation models are translation, affine, projective or higher polynomial transformations. Each of these transformation models requires a minimum amount of given tie points to determine the transformation parameters. In the case that more tie points are given than required, least square estimations are usually applied. As a consequence, not every tie point will be mapped exactly on the corresponding point in the

reference image. The computational costs of global transformation models are low but such models can not accurately handle local distortions.

Local transformation models are preferred over global ones, when local distortion between the images are assumed [89, 90]. Local transformation models map image regions differently, where the parameters of the applied mapping functions depend on the spatial location of the image regions. Frequently used local transformations models are piecewise linear or cubic functions, local weighted mean [89] and elastic models [91]. Local transformations are only able to accurately handle local image distortion if a large number of uniformly spread tie points are available. Detailed descriptions of common utilized global and local transformations models in the context of remote sensing image registration can be found in [75, 91–93].

On the other hand, if information about the image acquisition processes and the cause of the image distortions exist, a physically accurate transformation model can be constructed. Such models can be directly adapted to the cause of the distortions. For example, the main cause for the inaccuracy of the absolute geo-localization of the optical satellite data are inaccurate measurements of the satellite attitude and thermally affected mounting angles between the optical sensor and the attitude measurement unit (for details see Subsection 2.1.2). This insufficient pointing knowledge leads to local geometric distortions of orthorectified images caused by the height variations of the earth surface. Based on this knowledge, manually extracted GCPs or tie points automatically generated from the matching with an accurate geo-localized image can be utilized to adjust the parameters of the physical sensor model, and hence correct the geometric distortions.

**Image Transformation and Resampling:** After a suitable type of transformation model and its parameters are determined, the transformation function is utilized to map the input image onto the reference image and thereby aligning the images. Two possible methods exist to perform the transformation: The first one, often called forward mapping, directly transforms each pixel from the input based on the estimated transformation model. The problem of a forward mapping is the possible occurrence of holes and/or overlapping regions in the transformed image. The second method, often called inverse or backward mapping, generates the transformed image by applying the inverse transformation function on the pixel of the reference image in order to determine the corresponding intensity values from the input image. The advantage of the inverse mapping is that it assigns one intensity value to each pixel of the transformed image and thus, avoid holes and overlaps.

In general, the image grid of the input and reference image do not correspond to each other. In order to achieve a fine registration between the input and reference images, the transformed image has to be resampled to the reference image grid. Therefore, interpolation techniques are usually applied (in addition to the transformation function) to estimate the intensity values for all locations within the transformed image that lie between grid points in the input image. Commonly, methods such as nearest neighbor, bilinear and cubic convolutions, and B-splines are utilized for this task. An overview and detailed description of frequently applied interpolation techniques can be found in [94].

## 3.2 Traditional Multi-modal Image Registration Concepts - A Review

After introducing the fundamentals of image registration, we will now discuss the latest research studies regarding the problem of multi-modal image registration of optical and SAR satellite images and outline the current challenges of these traditional registration approaches. Since this thesis focuses on the development of accurate and reliable methods for tie point generation, we also set the focus of the following discussion on the various developed methods for the construction of image correspondences between optical and SAR image rather than on the subsequent process of image alignment. For the registration of optical and SAR images we follow a concept, which is based on an approach introduced in [95] and will be outlined in detail in Section 4.4. This approach pursues the idea of enhancing the physical sensor model of the optical satellite images through a set of GCPs in order to improve their geo-localization accuracy. In our case, the GCPs are represented by the part of tie points that is extracted from the high-resolution SAR images complemented by height information from a d digital elevation model (DEM). By utilizing these data for the sensor model enhancement of the optical images, the optical and SAR images get aligned.

### 3.2.1 Intensity-based Optical and SAR Image Registration Methods

Intensity-based concepts for the registration of optical and SAR images determine the transformation between two images by optimizing a corresponding similarity measure that assign image similarities based on a relation between pixel intensity values. Early registration approach such as [20, 96], investigated the applicability of correlation-based similarity measures to find image correspondences between optical and SAR images. However, correlation-based methods (e.g. NCC) or the squared intensity differences (SID) cannot always handle radiometric differences between multi-model images and are therefore unsuitable for the matching of multi-modal image data [97].

Influenced by the field of medical image processing, the cluster reward algorithm (CRA) [20, 21, 96], mutual information (MI) [20–24, 96] and the cross-cumulative residual entropy (CCRE) [25] have been repeatedly investigated for their applicability to optical and SAR image registration. Several of these studies [20, 21, 96] compared CRA and MI and came to the conclusion that MI is more robust against noise and radiometric differences and is therefore more suitable for matching optical and SAR image pairs. Suri and Reinartz [21, 23] further investigated the influence of the image content (type of objects within the scenes) on the image registration results. Due to the different geometric imaging properties of optical and SAR sensors, which are particularly pronounced for all 3D objects (e.g. buildings), images correspondences obtained from areas that contain such objects are not reliable. Therefore, they propose the usage of an segmentation approach, which is based on the image intensities of the SAR images, in order to discard all image regions containing above ground objects. Hassan et al. [25] on the other hand, investigated the use of the CCRE to measure the similarity between the images. Their results revealed that a matching of optical and SAR images based on CCRE is more robust compared to MI and in addition, computationally faster.

A particular difficulty of every similarity measure in the spatial domain is to handle the non-linear radiometric differences between optical and SAR images and the speckle in the SAR images. Liu et al. [98] try to overcome this problem by finding image correspondences between the images in the frequency domain. They therefore computed local frequency information in Log-Gabor wavelet transformation space utilizing the mean local phase angle and the frequency spread phase congruency. Subsequently, image correspondences are computed by applying a confidence aided similarity measure, which measures the similarity between the image pairs and provides a corresponding confidence score. The evaluation of this method showed a higher robustness to image noise and radiometric differences and a higher matching accuracy in comparison to MI (applied in the spatial domain).

### 3.2.2 Feature-based Optical and SAR Image Registration Methods

Feature-based approaches rely, in contrast to intensity-based approaches, on the detection and matching of robust and salient images features. A majority of feature-based methods are adapted to the detection, extraction and matching of one particular feature within the image scenes. Early approaches are often adapted to the detection and matching of line or regional features. Typical utilized line features are edges [26–28, 99–102], straight line segments [29] and contours [30, 31, 103, 104]. Contours are often extracted applying thresholding strategies [31, 103] or edge detection algorithms [30, 104]. To enable a robust detection of edges in the images it is normally recommended to apply different algorithms to optical and SAR images (e.g. Canny [82] for optical images and D1 method [105] for SAR images). However, the detection of edges or lines in SAR images is in general a difficult task and strongly influences the success of subsequent matching process. Region-based registration approaches on the other hand, commonly utilize areas such as larger fields or water areas such as lakes, rivers or flooded areas [32, 106, 107]. In [32, 106] suitable regions are detected using segmentation strategies and in [107] a supervised classification algorithm is applied for the detection of water bodies. However, a common problem of all feature detection methods that rely on water levels is the stability over time. In the case of extreme weather conditions, images that are acquired only a few hours apart can exhibit great differences in the shape and size of waters bodies.

Another category of feature-based approaches investigated the applicability of point feature detector and descriptor methods such as SIFT [85] or the local self-similarity (LSS) [108] for the registration of optical and SAR images. The feature descriptors provided by these methods commonly suffer from the speckle in the SAR image and the non-linear radiometric differences between the images and are therefore usually not effective for the generation of tie points [24, 109, 110]. As a consequence, several research groups investigated the adaptation of the SIFT and LSS operator to optical and SAR images [33, 34] and the application of SIFT in combination with other features detectors [35]. More precisely, Fan et al. [33] introduced a modified version of SIFT, which enables a fine registration for coarsely registered images, but on the downside is not applicable to image pairs with large geometric distortions. Ye et al. [34] on the other hand investigated the applicability of an adjusted LSS descriptor. The evaluation of results revealed the merits of the proposed approach in comparison to intensity-based approaches such as MI, but is only applicable if the images exhibit enough

shape or contour information. Xu et al. [35] successful utilized SIFT for the identification of corresponding points between optical and SAR images by combining SIFT features with a level set segmentation procedure for the detection of area features. A drawback of this approach is the need for sharp edges from runways, rivers or lakes.

If a feature-based approach utilizes only one particular feature it always exist the possibility that this features has a low occurrence or a poor distribution within the image scene. In such situations an accurate estimation of a transformation model will be difficult, especially in case of local image distortions. Therefore, several research studies investigated the use of a combination of different feature types [32, 111, 112]. Long et al. [111] proposed a single stage registration approach, which utilize the combination different features (points, straight lines, free-form curves or regions). The approach shows good performance for the registration of optical and SAR images, but a drawback is the need of a manual feature extraction from the SAR images. In [112] a two stage registration framework was proposed, where extracted image regions are utilized for a coarse registration and line and point features for a fine registration of the images. A similar approaches was proposed in [113], where the coarse registration was carried out through the matching of closed contours and the fine registration by corners detected through the Harris operator [80]. Although these methods utilize several kinds of features, they rely on a successful coarse registration and will fail if the first step provides an inaccurate transformation model.

To overcome the problem of misaligned images caused by unprecise detected and extracted features two research groups [29, 35] proposed iterative registration procedures. The approach in [29] is based on an iterative coarse-to-fine Voronoi spectral point matching procedure, which pursues the goal of finding point correspondences between extracted line-intersection. Here, for the coarse registration only the main spatial structures are utilized and extracted at lower resolutions. Due to the iterative detection and extraction strategy during the coarse and fine registration stage, the feature matching is more robust and yields to a more reliable set of tie points. On the other hand, the proposed approach is only applicable for image scenes that exhibit salient straight line features. The iterative approach introduced by Xu et al. [35] showed its effectiveness for high-, mid- and low-resolution images but (as mentioned above) requires the occurrence of sharp edges within the image scenes.

Instead of performing the feature-based registration in the spatial domain a few studies investigated the use of Fourier [114, 115] or wavelet transformations [116] to derive features in the frequency domain. The main reason for this course of action is to improve the robustness of feature detectors and descriptors to noise and the radiometric differences between optical and SAR images. Shi et al. [116] introduced a feature point extraction method particularly developed for the registration of islands. This method is based on the non-subsampled wavelet transform and a threshold shrink operator in order to extract robust and accurate key points from islands. In [114] a new feature descriptor was developed to extract local shape properties based on the amplitude and orientation of phase congruency. The obtained descriptors tend to be more robust to noise and radiometric differences, but are not invariant to scale or rotation changes between the images and hence unsuitable for images that exhibit larger geometric distortions. Another feature descriptor was introduced by Chen et al. [115]. In order to improve the robustness against speckle, features are detected based on a

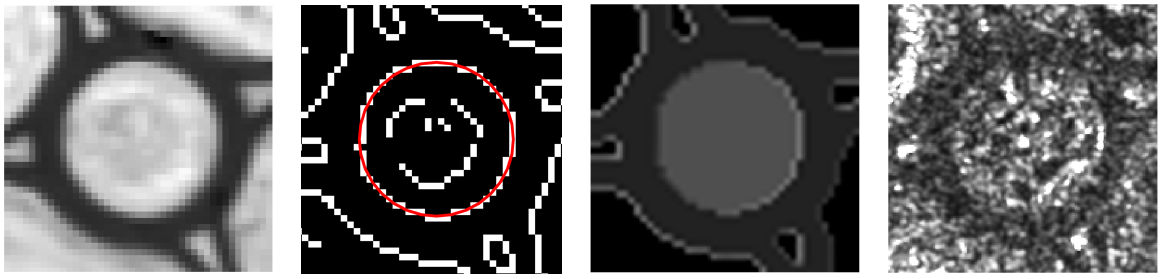
logarithmic phase congruency and the corresponding descriptors are constructed subsequently on the basis of Gaussian-Gamma-shaped bi-windows. A comprehensive evaluation showed the advantages of the proposed descriptor compared to SIFT and improved SIFT descriptors proposed in [33].

A common difficulty of all feature-based approaches is the robust detection of features especially in the SAR images (independent of the type of feature). Many feature detectors that perform well on optical images provide unreliable or unstable results on SAR images due to speckle within the images (e.g. [109, 110]). As mentioned above, several research groups tried to customize feature detectors to SAR images. Hänsch et al. [117] in contrast, tried to overcome this problem by developing a machine learning-based framework, which follows the goal of learning how a keypoint detected in the optical image appears in the SAR image by reformulating the problem of tie points generation to a classification problem. Therefore, SIFT is utilized to detect and describe a set of keypoints in the optical images, which serve as positive examples. Another set of negative keypoint examples is drawn from random position in the optical image. For every training sample (positive and negative) a patch around the corresponding location in the SAR image is extracted. The pixel intensity values of these patches are used to build the descriptors of the corresponding SAR image features. Note that for the generation of the training data aligned optical and SAR image pairs are needed. The last step includes the training of a random forest classifier in order to learn the identification of corresponding features descriptors in the optical and SAR images.

### 3.2.3 Hybrid Optical and SAR Image Registration Methods

Hybrid registration approaches try to overcome drawbacks of intensity and feature-based approaches by combining beneficial aspects of both registration schemes. A global coarse registration using mutual information on selected areas followed by a fine local registration based on linear features is proposed in [118]. An interesting aspect of the proposed approach is the selection of proper regions for the MI information-based coarse registration. Similar to [21, 23] image regions that usually provide tie points with high localization errors such as dense urban and heterogeneous areas are not taken into account. As a drawback, the method highly depends on the coarse registration. If the coarse registration fails, the fine registration will lead to unreliable results.

However, the majority of existing hybrid optical and SAR image registration approaches follow a two stage coarse-to-fine registration strategy, where a feature-based approach is utilized to coarsely register the images and an intensity-based approach for the fine registration. For the coarse feature-based registration the utilized features are straight lines [119], contours [112] or point features extracted with the help of the BRISK [120] or SIFT algorithm [121]. The fine intensity-based registration is realized by applying the correlation coefficient [112] and mutual information [119–121] as similarity measures. This registration concept is motivated by two observations from previous research studies. First, the accurate localization and assignment of features between optical and SAR images is prone to errors. Second, similarity-based approaches show difficulties in handling larger geometric distortions between the images to be registered. Although the difficulties of intensity and feature-based approaches are



(a) optical image (NDVI) (b) edge image (c) roundabout template (d) SAR image

**Figure 3.6:** Illustration of the artificial roundabout generation. From left to right: The roundabout in the normalized difference vegetation index (NDVI) image, the edge image and the detected central island (red marked), the artificial generated roundabout template and the roundabout in the SAR image.

partly remedied an accurate registration is not guaranteed. For example, if the coarse registration fails (overall alignment error larger than required for the algorithms used for the fine registration), e.g. due to missing contour or straight line features, the fine registration will probably also yield in inaccurate results.

The hybrid approach proposed in [122] combines an intensity- and a feature-based algorithm in another way. Here, a coarse transformation model is estimated by the use of MI, but in contrast to the other approaches it is subsequently utilized to remove unreliable tie points obtained from an improved SIFT-based image matching. Additionally, the influence of a speckle filter on the matching quality was investigated. Similar pre-processing concepts were proposed in [29, 100, 118, 120], where pre-filtering or de-noising techniques were applied in order to reduce the influence of noise or speckle on similarity measures and on feature detection methods. A risk of such pre-processing steps is a possible loss of the exact location of feature or the addition of artifacts. Due to a possible influences on the sub-pixel feature localization, Suri et al. [122] suggested the utilization of a speckle filter only for a coarse but not for a fine registration.

Another possible hybrid matching concept is an artificial templates-based matching approach [123]. In Merkle et al. [124] we introduced such a matching approach, which focuses on the generation of artificial roundabout templates. The three main steps of this method are: 1) detect and extract roundabouts in the optical image, 2) generate artificial SAR-like templates out of the detected roundabout information (see Figure 3.6) and, 3) apply an intensity-based matching approach to match the artificial templates with the SAR images. The advantage of this approach is that features only have to be extracted in the optical images and by adapting the radiometric properties of the artificial templates to SAR images, even similarity measures such as NCC yield accurate results. A drawback of the proposed method is the hand-crafted generation of the artificial templates, which requires the knowledge of the common radiometric properties of SAR image features and lead to unsatisfying results if the visual appearance of the extracted features deviates from the norm.

### 3.2.4 Challenges of Traditional Optical and SAR Image Registration Methods

Every optical and SAR image registration framework has to face challenges, mainly due to the different sensor acquisition concepts and different times, viewpoints and weather conditions during the image acquisition. In particular, different radiometric and geometric properties caused by the different acquisition concepts and speckle in the SAR images hamper the accurate registration of optical and SAR images. In order to develop suitable optical and SAR image registration methods, these challenges have to be taken into account and insights obtained from the state-of-the art approaches should be exploited. Therefore, we will summarize in the following the major challenges a traditional optical and SAR image registration approaches has to face.

As shown in Subsection 3.2.1 many intensity-based approaches have been developed and investigated over the past years and several of them showed their potential for the registration of optical and SAR images. Nevertheless, these approaches mainly deduce image correspondences on the basis of pixel intensity values and hence suffer from the different radiometric properties of optical and SAR images. Additionally, these approaches are often computationally expensive and sensitive to speckle in the SAR image and, usually only achieve high registration accuracies if the distortions between the images are small. Besides, several comparative studies have demonstrated the higher suitability of feature-based approaches for the problem of optical and SAR image registration, due to their higher flexibility and less sensitivity to illumination, reflectance, and geometry inconsistency between the images.

Nevertheless, feature-based approaches are not free of challenges and difficulties when applying them on optical and SAR images. In general, it is difficult to develop one approach which is able to reliable extract features among images with various imaging properties. Such methods have to take into account that features might change over time (multi-temporal images) or are dissimilar in both images (radiometric differences between multi-modal image data or different image acquisition conditions). In order to handle local image distortions the chosen approach has to be capable of selecting features that are visible in both images, stable over time, frequently spread and occur in a large number and further, is not sensitive to noise. Even though such features occur in the images these features have to be accurately extracted, which have proven to be very difficult especially in SAR images, and robustly matched in order to obtain reliable and precise tie points.

One way to overcome some of the existing problems is the development of a hybrid-based registration approach. By combining valuable characteristics of intensity- and feature-based approaches promising concepts for the registration of optical and SAR images could be developed in the past. Nevertheless, hyprid-based approaches, like intensity- and features-based ones, are handcrafted in which case every processing step has to be carefully developed or adapted to fulfill the particular needs of optical and SAR images. Often, these approaches are specialized for the detection, description and matching of one specific type of feature, which limits the applicability to particular image scenes.

## 3.3 Deep Learning-based Image Matching Concepts

A variety of research studies indicate the high potential of deep learning-based methods for tasks such as image classification [125], object recognition [126] and detection [127]. Over the last years, these methods outperformed traditional methods in various tasks and over different research areas due to their impressive ability of feature detection, representation and discrimination. In the field of remote sensing neural networks find already application for problems such as the classification of hyperspectral data [128], enhancement of existing road maps [129] or high-resolution SAR image classification [130]. Furthermore, neural networks were successfully trained for tasks that rely on the detection of image correspondences between two images such as stereo evaluation [131–135], optical flow estimation [136–138], oblique aerial image matching [139], ground to aerial image matching [140] or image registration [141, 142].

In contrast to traditional approaches for the tie-point generation, deep learning-based approaches open up the possibility of finding image correspondences without the need of handcrafting similarity measures or feature detectors, descriptors and matching concepts. Fischer et al. [143] investigated the quality of features extracted with CNNs by comparing their matching performance to SIFT features for manual pre-selected regions of interest. Their results showed a higher quality of the extracted CNN features in comparison to SIFT features and hence the potential of CNNs for the task of image matching. In [144] the utility of learned features from a pre-trained CNN in combination with SIFT features for the matching of remote sensing data was investigated. More precisely, the obtained SIFT features and the CNN features, computed from an area around the SIFT features, are fused and afterwards matched between the images by utilizing the Euclidean distance. The evaluation over several image pairs showed an improved matching performance through the combination of SIFT and CNN features in comparison to a matching based on SURF or SIFT features only. Yi et al. [145] went one step further and proposed a framework where CNNs are not only used for the description of features from pre-selected areas, but are also used for the step of feature detection and orientation estimation. They compared the proposed framework with a number of feature detection and description algorithms (e.g. SIFT, BRISK, SURF [146] and FREAK [147]), which revealed the effectiveness of the proposed approach and the advantages over these traditional matching concepts.

So far, the introduced deep learning-based approaches still require a handcrafted feature matching framework to find image correspondences between images. In [148], a deep learning-based method is proposed to detect and match multiscale keypoints with two separated networks. The detection network is trained on multiscale patches to identify regions including good keypoints. The description network is trained to match extracted keypoints from different images. Here, the problems of feature extraction and matching are regarded separately and as a consequence, the extracted features are might not be the most suitable ones for image matching. Therefore, the latest research studies propose end-to-end concepts on the basis of a Siamese neural network architecture [149]. The basic idea of these concepts is to learn both task in one step by training one neural network, which is composed of two parts: The first part, a Siamese or pseudo-Siamese neural network, is trained to extract features from image patches. In the majority of approaches the two

branches of the Siamese neural networks are realized through convolutional neural network [131, 132, 135, 139, 150–154]. The second part of the network, sometimes called fusion or classification network, is trained to measure the similarity between the extracted features. It is often realized through fully connected layers [132, 135, 139, 150, 151, 153], the $L_2$ distance [152, 154] or the dot product [131, 132]. The input of these networks can be single resolution image patches [139, 150, 152, 154], multi-resolution patches [151, 152] or patches which differ in size for the left and right branch of the Siamese neural network [131, 151, 153]. In order to learn the detection of image correspondences, the networks are commonly trained by using the hinge loss [132, 135, 152–154] or the cross-entropy loss [131, 139, 150].

In general, the just described framework has proven high potential by providing state-of-the-art results on several challenging benchmarking datasets such as KITTI [155] and by offering a high degree of flexibility in terms of the particular input data feed into the network or the particular learning goal. For example, the above introduced frameworks can be applied on a coarse level with the aim of predicting, if two image patches, e.g. acquired from different viewpoints, show the same scene or not, or one fine level with the aim of finding a dense pixel-to-pixel correspondence, e.g. for stereo or optical flow estimation. On the downside, a sufficient amount of training data have to be available in order to train the network and among other things an optimal network architecture, loss function and training procedure have to be carefully chosen.

## 3.4 Research Gaps

The detailed discussion about the relevance of the topic from Section 1.2 underlines the high potential and necessity for accurate and precise optical and SAR image registration frameworks. However, the theoretical comparison of optical and SAR images outlined in Section 2.1, and the summary in Subsection 3.2.4 about the current challenges of traditional registration methods, show the necessity for further investigations and developments of a general concept to handle the problem of optical and SAR image registration.

The basis of such a concept is the creation of an optimal (at least to certain degree) initial situation. Due to fundamental differences in the geometric properties of optical and SAR images and the demands on reliable features for the generation of accurate tie points, not every area within the images is suitable. In order to minimize the impact of the different acquisition modes of optical and SAR satellites and hence enhance the conditions of accurate and reliable tie point generation process, the development of an automatic process that enables the pre-selection of suitable matching areas from optical and SAR images should be researched (see Subsection 4.1).

Traditional intensity- and feature-based approaches have proven their high potential for the generation of accurate and reliable tie points in the case of single sensor image matching but underperform for the precise extraction of corresponding features from optical and SAR images. By a careful pre-selection of suitable areas, the influence of geometric differences can be reduced to a minimum and only radiometric differences have to be taken into account. In order to handle the radiometric differences, an interesting and open research direction is the translation of optical images into SAR images or vice versa by maintaining geometric properties of the input image while synthesizing radiometric properties of the desired output image. Such an image translation could improve the conditions of traditional approaches and could enable the utility of their valuable characteristics (see Subsection 4.2).

However, taking the findings and insights of previous research studies into account, feature-based approaches are to be preferred over intensity-based ones mainly due to the lower susceptibility of the former to the non-linear radiometric differences between the images. Nevertheless, traditional feature-based approaches are handcrafted and many approaches are developed for the extraction of one particular type of feature. The development of an universal concept, which is able to precisely detect and extract diverse and corresponding feature information from images, is very difficult. Considering the high potential and success of deep learning techniques for the automatic extraction and matching of features from optical images (as discussed in Subsection 3.3), the investigation of an deep learning-based approach for the case of optical and SAR image data represents a promising field of study (see Subsection 4.3).

# 4

# Deep Learning-based Optical and SAR Image Registration

*The differences in the geometric and radiometric properties of optical and SAR images pose a substantial challenge for every matching approach. In particular, the handcrafted feature extraction stage of common optical and SAR image matching methods suffers from this circumstance and requires a concept, which is carefully tailored to the characteristics of optical and SAR image pairs. Recent breakthroughs in the training of neural networks through deep learning techniques opened up new possibilities and led to the development of automatic feature extraction and matching methods for the matching of single sensor images. In this chapter two pre-processing chains, a semi-automatic and an automatic one, will be presented in order to create an optimal initial situation for the matching approaches by limiting the geometric differences of optical and SAR image pairs through the extraction of suitable matching areas. Subsequently, two novel deep learning-based optical and SAR image matching methods for the generation of accurate and precise tie points are presented. Finally, a scheme for the registration of optical and SAR images, enhancing the geo-localization accuracy of optical images through the extracted tie points will be discussed.*

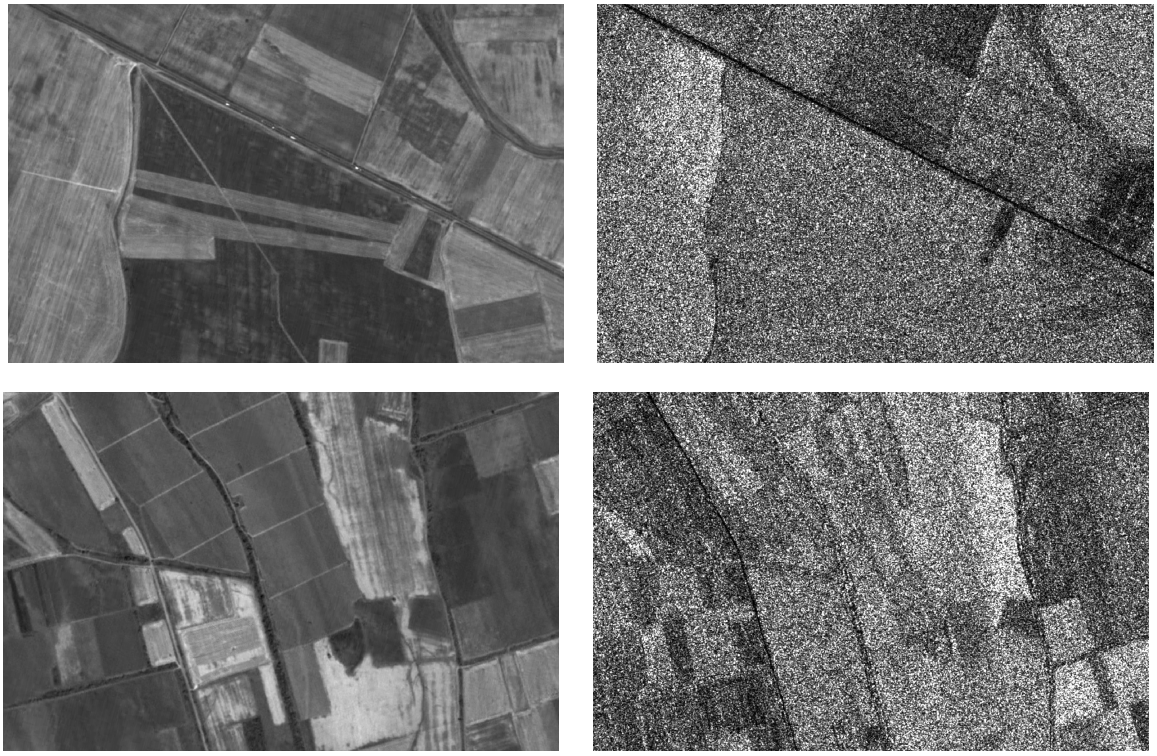
## Contents

## 4.1 Matching Areas Pre-selection

As indicated in Subsection 3.4, reducing the impact of the different acquisition modes through the pre-selection of suitable matching increases the probability to obtain accurate and reliable tie points between optical and SAR images. More precisely, candidates for such suitable areas should contain almost only planar objects that exhibit the same (at least to a certain degree) geometric appearance in the optical and in the corresponding SAR image. In most cases, these features are related to man-made infrastructure objects such as streets, street crossings, roundabouts and borders between agricultural fields. The reason for excluding 3D objects are the different geometric distortions induced by the different acquisition types for optical and SAR images. As described in Subsection 2.1.2 elevated objects like buildings appear differently in optical and SAR images and get projected to different positions within the image. As a consequence, the boundary of an elevated object in a SAR image does not fit the object boundary in the optical image, even if the imaging perspective is the same for both sensors. Therefore, these features can only be utilized for the generation of tie points if additional three-dimensional information (e.g. DEMs or point clouds) for the image scenes is given with high accuracy. Since this is not the case for our investigations we refer to [156, 157] for two research studies examples that investigate the tie point generation between optical and SAR images over urban areas by taking three-dimensional information into account.

In addition to the mentioned geometric aspect, two further important points have to be considered during the area pre-selection in order to further increase the probability of a successful matching. First, selected areas should actually contain salient features. Large areas that contain only homogenous structures, e.g. exclusively crops or grass land, are very likely to lead to ambiguities during the matching process, and hence should be excluded. Second, relevant features should be visible in both the optical and the SAR image. Due to a higher level of detail in the optical images it is often the case that features are visible in the optical images but not in the corresponding SAR images. Figure 4.1 shows samples, where many roads and field boarders are visible in the optical but not in the SAR image. Nevertheless, the reverse case occurs when clouds or their shadows cover the image scenes during the acquisition of the optical images. The second aspect is especially important for the development of an automatic process.

Under different circumstances, data sources such as OpenStreetMap (OSM) data could be utilized for the pre-selection of suitable areas by providing additional information. For example, OSM data includes the rough location of the majority of streets, street crossing and roundabout for most areas around the world. However, since only parts of the available information, e.g. the existing road network, provided by OSM is actually visible in the SAR images the direct use of such data is not feasible in our case of application. Nevertheless, we realized the collection of suitable matching areas, which later build the basis for the developed tie point generation approaches (presented in Sections 4.2 and 4.3), through the development and application of a semi-automatic selection procedure, which is introduced in Subsection 4.1.1. Additionally, we provide the concept of a fully automatic scheme for the pre-selection of suitable matching areas in Subsection 4.1.2 in order to facilitate the future development of a fully automatic matching framework.



(a) Optical image samples (b) SAR image samples

**Figure 4.1:** Illustration of the feature visibility in optical and SAR images. Optical images commonly exhibit a higher level of detail compared to SAR images, and hence suitable features for image matching such as roads and field boarders are in many cases not visible in the corresponding SAR images.

### 4.1.1 Semi-automatic Pre-selection of Matching Areas

To get a first indication of areas that contain suitable patterns, such as parts of streets or runways in rural areas, the CORINE land cover [158] from the year 2012 is applied. The utilized CORINE layer includes 44 land cover classes and has a pixel size of 100 m. This data enables the exclusion of unsuitable areas, which contain e.g. cities, industrial areas or woodlands. We choose the following classes for a first pre-selection: airports, non-irrigated arable land, permanently-irrigated land, annual crops associated with permanent crops and complex cultivation patterns, land principally occupied by agriculture, with significant areas of natural vegetation. Figure 4.2 exemplifies a variety of different land cover classes, where the description of the discarded classes are red framed and of retained classes green framed. Note that in this thesis only satellite images acquired over Europe are utilized, and hence only the CORINE land cover is needed. For a similar pre-selection for images outside Europe several existing global land cover maps can be utilized.

Subsequently, the resulting image areas are manually refined to ensure that the above mentioned requirements are fulfilled: 1) all selected areas should exhibit salient features and 2) these features should be visible in the optical and the SAR image. Additionally, due to the relatively large resolution of the CORINE layer some image regions still contain street segments through smaller villages, and hence have to be discarded during the manual selection process. The manual refinement is realized by cropping overlapping image patches

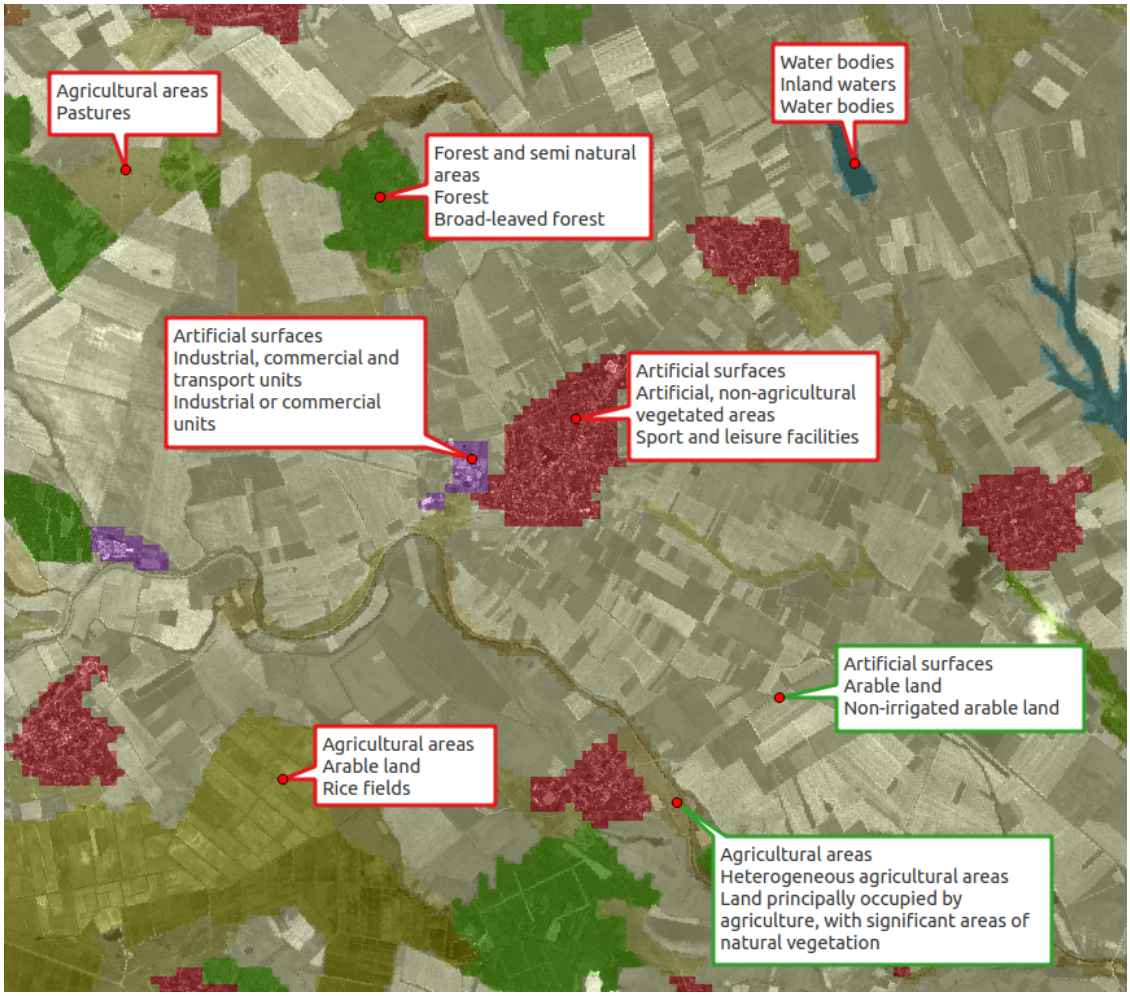



**Figure 4.2:** Illustration of different land cover classes from the CORINE [158] layer. All classes listed in red-bordered text boxes are discarded, while all classes listed in green-bordered test boxes are retained.

from the masked optical and SAR images. The size of the image patches and the offset between them can be adapted to the respective application. In our case we choose a patch size of $201 \times 201$ pixels and an offset between the patches of 20 pixels in easting and northing direction. Note that we generate overlapping patches in order to increase the size of our training dataset. All images patches that violate one of three points, e.g. exhibit only ambiguous structures or exhibit salient features in one patch of each patch pair only, are discarded from the set of matching patches. Without the manual refinement, unsuitable patch pairs could hamper the detection of corresponding features, and hence interfere the later matching process.

Nevertheless, the described semi-automatic process is time consuming and vulnerable to human errors during the manual patch selection. As a consequence, the fast extension to new image pairs is associated with high efforts and a slow realization. To handle these problems and to enable the future development of a fully automatic matching framework, we further propose an automatic matching area selection concept in the following subsection.

### 4.1.2 Automatic Pre-selection of Matching Areas

The pre-selection of matching areas represents an important pre-processing stage for the later introduced deep learning-based tie point generation concepts. Since deep learning techniques require a large set of training data, a concept for the automatic generation of training sets from aligned optical and SAR images have to be provided in order to enable a fast and simple extension to new image pairs. In particular, if the learned model should be applied on new image data acquired from different optical and/or SAR sensors that exhibit different imaging properties compared to the training dataset. For this purpose, we propose the following automatic concept.

Similar to the semi-automatic framework, the use of the CORINE layer is suggested to get a first collections of pre-qualified areas that contain fitting patterns. This step is realized by excluding unsuitable areas such as cities, woodland and industrial areas. This step is followed by the application of OSM data in order to ensure that suitable features are contained in the set of pre-qualified areas. Particularly, we propose the use of the road network provided by OSM data and more specific the derived street crossings as features of interest. Commonly, road networks, and hence street crossings are spread over the whole image scenes and are present in nearly every image. These objects are therefore well suited for the problem of image matching by providing the possibility of generating equally distributed tie points between the images. The advantages of utilizing street crossings instead of streets are the prevention of ambiguities during the matching process. An example of the provided OSM road network and the derived street crossings for a rural area in England is illustrated in Figure 4.4. Here, the road network contains all kind of roads such as highways, country roads and dirt roads.

However, the provided information from OSM data is neither complete nor free of errors (in terms of location accuracy and wrong or outdated information) and, as mentioned earlier, not every street that is visible in optical images is visible in the corresponding SAR images. In order to identify the part of the provided road network, and hence the particular street crossings that are visible in the SAR images the implementation of a further processing step is required. The detection of road in SAR images is a difficult problem in itself (see example in Figure 4.3) and was tackled by different research studies in the past, e.g. [159–161].

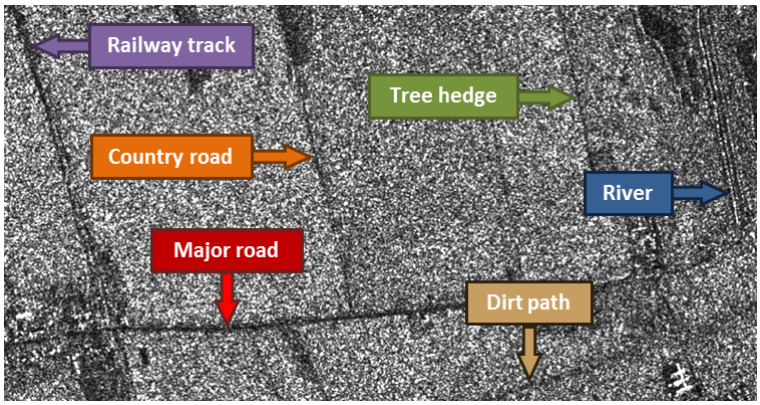
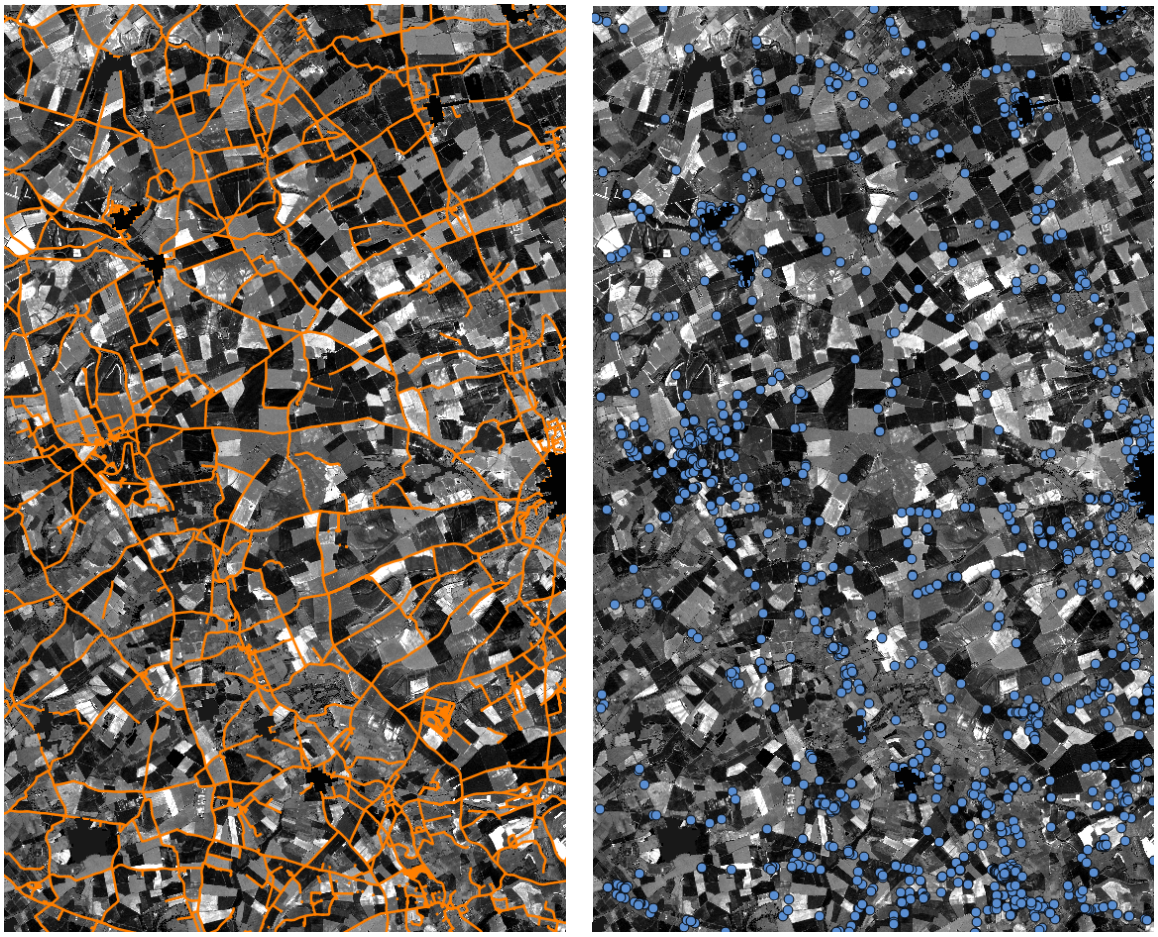



**Figure 4.3:** SAR image subset illustrating objects of different nature which look much alike. A segmentation model must learn to distinguish all kinds of roads from railways, tree hedges and rivers.



(a) OSM road network for a rural area (b) Derived street crossings

**Figure 4.4:** Illustration of the road network and the derived street crossings provided by OSM data for a rural area in England. The orange marked road network in Figure (a) contains all kind of roads provided by OSM, e.g. highways, country roads and dirt roads. The blue marked street crossings in Figure (b) are derived from the road network.

Nevertheless, we developed a novel approach for the road detection in SAR images based on fully-convolutional neural networks (FCNNs) [162]. As this method will not be utilized in later sections we only provide a short summary about the FCNN-based road detection framework for interested readers below and refer to [162] for more implementation aspects and a detailed evaluation of its performance.

In a last step, the information from the OSM data and the detected streets in the SAR images is combined to identify areas that contain street crossings, which are visible in the SAR and optical images. In order to utilize these areas for the training of a neural network or for another matching algorithm patches with a certain size and overlap can be automatic cropped from these areas and utilized for the training. Note that the obtained patches are most likely not complete (in terms of containing all existing visible road crossing). However, for a later matching it is more important that the resulting patches are located across the whole image scene and, actually contain distinct features visible in both the optical and SAR image.

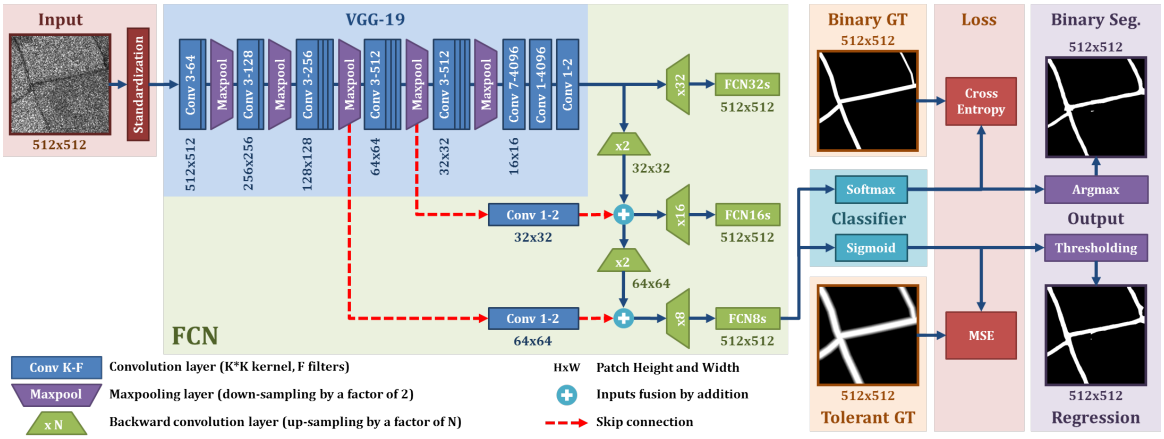



**Figure 4.5:** Illustration of the proposed road segmentation approach, including three specific FCNN architectures: FCN8, FCN16 and FCN32.

**Road Segmentation in SAR Images:** We based our method for a pixel-wise segmentation of roads in SAR images on a FCNN architecture, which have proven to provide accurate and detailed segmentation [163]. A FCNN is commonly composed of two parts: The first part, a DCNN, analyzes the input SAR images and outputs a cluster of predictions by gradually down-sampling to input image through pooling layers and at the same time extracting more and more meaningful features. We choose the VGG-19 [126] architecture as our DCNN. The resulting predictions of the VGG-19 are subsequently processed by the second component of the FCNN, the up-sampling network. It restores the spatial properties of the predictions using backward convolution layers (commonly called deconvolution layers [164]) until the road predictions (output image) share the same size as the input image. In order to assemble both parts, we follow the suggested FCN8s architecture proposed in [163]. This specific version of the FCNN infuses the results from two intermediary layers of VGG-19 into the up-sampling process through skip-connections (see Figure 4.5). These layers have a finer prediction resolution than the DCNN output and help to improve the segmentation accuracy. The FCN8s architecture is commonly preferred over other FCNN versions such as the FCN32s and FCN16s, where the FCN32s directly up-samples the output of the VGG-19 32 times, resulting in a coarse segmentation, while FCN16s fuses only one layer. A comparison between the road predictions obtained by these three FCNN versions is illustrated in Figure 4.6.

**Figure 4.6:** Segmentation result comparison between three the FCN versions (left to right: FCN32s, FCN16s, FCN8s, ground truth)

Instead of treading the problem of road segmentation as a binary classification, where each pixel must be classified as road or as background, we propose the use of regression with an adjustable spatial tolerance during the network training. Therefore, we adapted ideas proposed in [131, 165] and utilize a smooth target distribution $\boldsymbol{Y}_{\text{tol}}$, which is centered around the true ground truth distribution $\boldsymbol{Y}_{\text{bin}}$. The distribution $\boldsymbol{Y}_{\text{bin}}$ contains binary labels for each pixel in the input image depending if the pixels are labeled as road or background. The values of the smooth target distribution in contrast range from a maximal value of 1 (for pixels labeled as roads in $\boldsymbol{Y}_{\text{bin}}$) and linearly decrease to 0 until a fixed distance is reached (e.g. 4 pixel apart from a pixel labeled as road). The advantage of this approach is that incorrect network predictions can be penalized depending on their distance to the ground truth. In other words, incorrect predictions only a few pixels away from the actual road are only slightly penalized during training. In order to train the network we utilize the following mean squared error (MSE) loss

$$\mathcal{L}_{\text{MSE}}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} w_i(y_i - \hat{y}_{\boldsymbol{\theta},i})^2\,, \tag{4.1}$$

where $\boldsymbol{\theta}$ are the network parameters learned during training, $y_i$ the value of the smooth target distribution $\boldsymbol{Y}_{\text{tol}}$, $\hat{y}_i$ the predictions and $w_i$ the loss weighting coefficient for the $i$-th pixel in the output image with an overall number of $N$ pixel. The need for a weighted loss emerges from the fact that roads appear as thin objects in the SAR images and are likely outweighed by the background class, especially outside cities. Therefore, we follow the approach proposed in [166] and reweighting each class during the loss calculation by multiplying the loss associated with pixel $i$ by the corresponding loss weighting coefficient $w_i$, which is defined as follows

$$w_i = \begin{cases} \tau & \text{if pixel } i \text{ is labeled as a road in } \boldsymbol{Y}_{\text{bin}} \\ 1 & \text{otherwise} \end{cases}\,. \tag{4.2}$$

Here, $\tau$ is a fixed value taken from the interval $[1, 1/f_{\text{road}}]$, where $f_{\text{road}}$ is the ratio of road pixels over the total number of pixels $N$. The training of the network is realized by minimizing Equation 4.1 with stochastic gradient descent and the adaptive moment estimation (ADAM) optimizer [63].

Since FCNNs are likely to provide imprecise predictions along object boundaries and as roads are thin and expected to be smooth and continuous, the predictions of our network have to be refined after the training process. Therefore, we follow the common practice and use fully-connected conditional random fields (FCRFs) [167] in order to refine the segmented roads, and hence to further improve the overall prediction quality. FCRFs provide a learnable approach to enhance region boundaries on segmentation maps and have been successfully employed in combination with FCNNs [168]. FCRFs using image wide context instead of local context through a pairwise comparison of pixels. Taking into consideration the predictions and the input image, it aims at improving the border smoothness between side by side areas, by minimizing the following energy function

$$E(\boldsymbol{y}) = \sum_i \phi_i(\hat{y}_i) + \sum_{i,j} \phi_{i,j}(\hat{y}_i, \hat{y}_j)\,, \tag{4.3}$$

where $\hat{y}_i$ is the predicted label for the $i$-th pixel. The energy function is a compound of two potentials. The first one, the unary potential $\phi_i$, penalizes any uncertainty in the prediction. The more confident the network is about the predicted class of a pixel, the lower the weight. The second one, the bilateral pairwise potential $\phi_{i,j}$, contains three terms summed overall pixel pairs. More detailed, $\phi_{i,j}$ compare two pixels and checks for consistency by looking at their predicted class, their corresponding color intensities on the input image and their positions. The more alike they are, the lower their potential. This energy encompasses the whole image, as it interconnects all pixels together, effectively leveraging the full context of the picture. The FCRF must be trained in order to minimize its energy function. FCRFs have an erosion effect on the predictions and since roads are already thin objects, they might be narrowed and even disconnected from each other in the process. Therefore, we apply the FCRFs on the background predictions in order to fill gaps between the roads and helps reconnecting them. For this purpose, we invert the input values given to the FCRFs. The resulting segmentation map is subsequently inverted to obtain the refined road predictions. An example of the resulting raw and refined network predictions is shown in Figure 4.7.

### 4.1.3 Summary

Two frameworks for the pre-selection of suitable matching area have been introduced in this section. The semi-automatic framework can be summarized in the following three steps:

1. Apply the CORINE layer in order to get a first indication of qualified areas by discarding unsuited areas such as cities, industrial areas or woodland.
2. Crop (overlapping) patches with the desired size from the obtained areas in the optical and SAR images.
3. Manually refine the set of patch pairs by discarding patches that do not exhibit salient features in both patches of each optical and SAR patch pair.

The fully automatic framework on the other hand can be summarized in the following four steps:

1. Apply the CORINE layer in order to get a first indication of qualified areas by discarding unsuited areas such as cities, industrial areas or woodland.
2. Apply OSM data in order to derive all available streets crossing from the provided road network in each image scene.
3. Utilize the FCNN-based road detection method in order to identify visible roads in SAR images.
4. Combine the resulting information and crop (overlapping) patches with the desired size from areas that exhibit street crossing in the optical and SAR images.

Both frameworks enable the provision of an optimal initial situation for the tie point generation frameworks presented in the next two sections. More precisely, both matching approaches will benefit due to the following aspects:

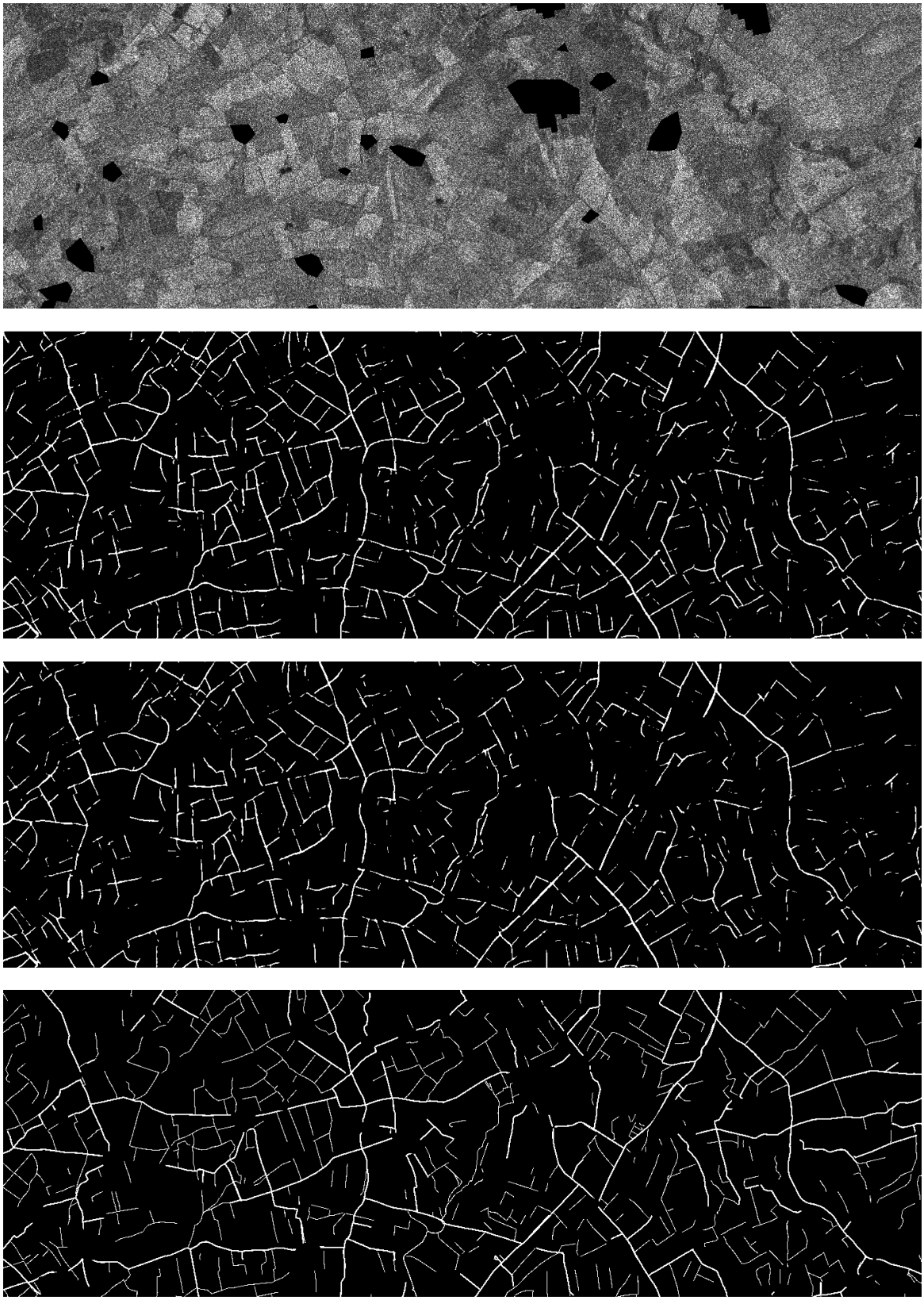


**Figure 4.7:** Illustration of the road segmentation results of a SAR image sample. From top to bottom: The SAR image with masked city areas, the network predictions, the refined predictions using FCRFs and the ground truth road network.

- Due to the elimination of areas containing elevated objects such as cities, the different geometric properties of certain objects in the optical and SAR images (caused by the different sensors) will be reduced to a minimum. As a consequence, the reliability of the obtained tie points with regard to their geo-localization will be increased.
- By ensuring the existence of distinct features in the areas to be matched, the risk for our matching approaches to produce total mismatches will be reduced and the quality of the resulting tie points will be increased.

Note that for the generation of the final training, validation and test dataset, and hence for all experiments presented in Chapter 5 the semi-automatic framework introduced in Subsection 4.1.1 is utilized. Details about the obtained datasets are provided in Subsection 5.1.2. The automatic area pre-selection framework introduced in Section 4.1.2 was developed to provide the option for the future development of a fully automatic matching framework and will not further be utilized, evaluated or discussed in the following chapters.

## 4.2 Conditional Adversarial Networks for Multi-modal Image Matching

Our research objective is the computation of a set of very accurate and reliable tie points between optical and SAR images. Towards this goal we developed two novel frameworks. The first one, presented in this subsection, is based on the application of conditional adversarial networks [169, 170]. The basic idea of this approach is to eliminate (to a certain degree) radiometric differences between the images to be registered, and hence enable the utilization of traditional matching approaches such as NCC, MI, SIFT and BRSIK, which have proven to yield to accurate and reliable results in the case of single-sensor image matching. The proposed approach represents an extension of our previous work [124], where we focused on the generation of artificial roundabout templates and their applicability for the generation of tie points. The general concept of this approach can be summarized in the following three steps: 1) Detect and extract roundabouts in the optical image, 2) generate artificial SAR-like templates out of the extracted roundabout information and, 3) apply an intensity- or feature-based matching approach to match the artificial templates with the SAR images. This approach fulfills important demands of an optical and SAR image matching approach (e.g. features are not directly extracted from SAR images, traditional matching approaches are applied on images patches with similar radiometric properties, geometric differences are almost eliminated through skillful selection of suitable objects) and first results reveal the potential of this concept [18]. However, the approach has some crucial weaknesses, which complicates the extension to more frequent image features such as street crossings.

The first drawback is the need of the precise detection and extraction of feature information to generate high quality feature templates. This requires several processing steps and the result mainly depends on the quality of the detected edges, which makes the approach impractical for images scenes with lower resolutions or for image scenes containing only small objects. Furthermore, in contrast to roundabouts, the shape of features such as street crossings vary in shape between image scenes from urban, suburban, rural areas and between different countries (uniform street blocks in many parts of the United States vs. irregularly shaped street patterns in most parts of Europe). The extensions to street crossing and other more complex features therefore requires the development of an universal concept in order to extract the geometric properties of varying types of features. The second drawback is the template generation step. The hand-crafted generation of artificial templates requires the knowledge of the common radiometric properties of the features and lead to unsatisfied results if the visual appearance of the feature deviates from the norm.

To exploit the advantages of the above described matching approach and to enable the generalization to all kind of optical and SAR image scenes the above mentioned problems have to be tackled. One way of doing this is by utilizing a method which automatically generates artificial templates without the need of extracting geometric information of features. In the field of deep learning the generation of new training samples is crucial for tasks where training data is limited. As described in Subsection 2.2.2 generative adversarial networks (GANs) can be trained for the task of generation artificial images from noise. Based on the concept of GANs, Isola et al. [171] proposed a method, which enables the generation of an artificial image with the texture of a reference image, while keeping the geometric properties of a given input image. This is realized through the training of a conditional generative

adversarial network (cGAN). Inspired by the high potential provided by this approach, we add cGANs to our processing chain. This step removes the feature extraction part, the improvement of the artificial template generation and the generalization to various kinds of input features. Furthermore, due to the automatic image generation framework and the associated omission of the handcrafted feature detection and extraction stage, the generation and utilization of artificially generated optical image patches comes feasible, too.

### 4.2.1 Concept of Optical and SAR Image Matching Based on Conditional Adversarial Networks

The framework for the generation of tie points between optical and SAR images through the use of conditional adversarial networks is composed of the following three steps. The first stage includes the selection of suitable matching areas from optical and SAR images. For the selection of suitable matching areas the semi-automatic method described in Subsection 4.1.1 is utilized and will not be further discussed. The second stage includes the generation of artificial image patches from optical or SAR image patches through a generator network $G$. The training of $G$ is realized through the concept of conditional generative adversarial networks and is outlined in detail in Subsection 4.2.2. The third stage includes the matching of artificially generated patches with the real image patches counterpart through the application of traditional tie point generation concepts and is described in Subsection 4.2.3. An graphical overview of the whole concept is depicted in Figure 4.8.

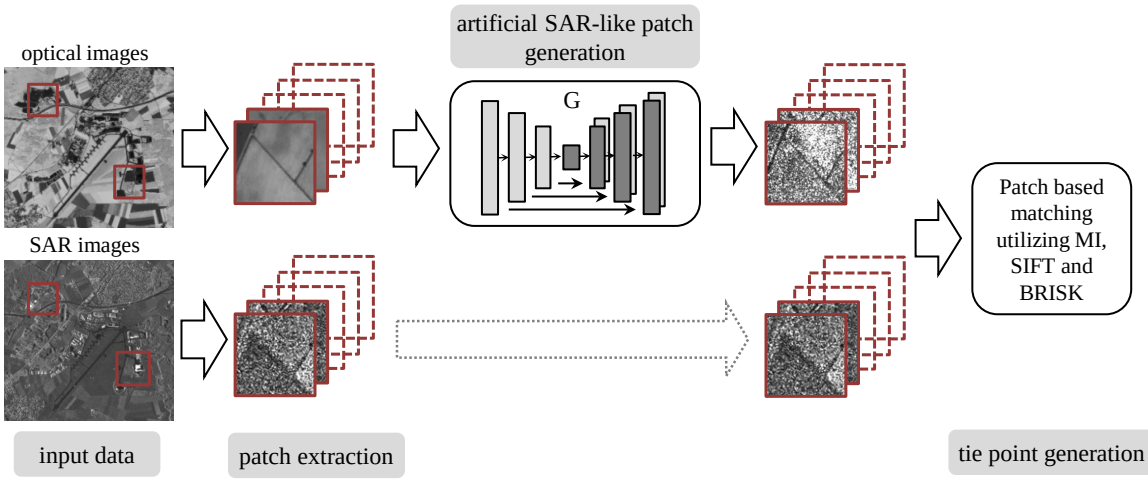



**Figure 4.8:** Graphical overview of the cGAN-based tie point generation framework. Tie points are generated by matching SAR and artificial image patches created by a generator network $G$.

### 4.2.2 Details of the Artificial Image Generation Process

In the case of unconditional GANs (see Subsection 2.2.2) the artificial images are generated from noise and the control of the image generation process, for example through the implementation of certain requirements, is hardly feasible. In our special case of application, the goal is to generate artificial image patches with geometric properties of given input image patches and with radiometric properties of a determined output image. Therefore, we utilize the concept of conditional GANs (cGAN), which enable the use of additional

information or requirements in form of input data such as discrete labels [172, 173], text [70] or images [171, 174]. As a consequence, a more controlled and target-oriented images generation process can be achieved. In our case, the conditioning is based on an optical or SAR image patch as additional input. For reasons of simplification we assume optical patches as input and SAR patches as our target in the following.

Our utilized image cGAN relies next to noise samples $\boldsymbol{z}$ on optical input image patches $\boldsymbol{x}$ and consists, similar to GANs, of a generator network $G$ and a discriminator network $D$. Through the pre-selection of suitable areas (as described in the previous Subsection 4.1) the impact of geometric distortion is reduced to a minimum, and hence we can assume the same geometric properties between the given optical input patches and the corresponding SAR output or target patches. As a consequence, the generator $G$ can directly utilize the geometric structure from the input patches $\boldsymbol{x}$ and only has to learn the radiometric or style transformation from the optical patches to the corresponding SAR patches $\boldsymbol{y}$. Towards this goal, both networks are trained through an adversarial process in order to learn the generation of conditional artificial images. By adding the condition to the GAN loss from Equation 2.22 we obtain the following cGAN loss

$$\mathcal{L}_{\text{cGAN}}(G,D) = \begin{array}{c} \overbrace{E_{\boldsymbol{x},\boldsymbol{y}\sim p_{\text{data}}(\boldsymbol{x},\boldsymbol{y})}[\log D(\boldsymbol{x},\boldsymbol{y})]}^{\text{predicted log probability of } D \text{ that the image pair } (\boldsymbol{x},\boldsymbol{y}) \text{ is real}} \\ + \\ \underbrace{E_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x}),\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1-D(\boldsymbol{x},G(\boldsymbol{x},\boldsymbol{z})))]}_{\text{predicted log probability of } D \text{ that the image pair } (\boldsymbol{x},G(\boldsymbol{z})) \text{ is fake}} \end{array} , \tag{4.4}$$

where $E$ denotes the expected value, $p_{\text{data}}$ the real data distribution, $p_{\boldsymbol{z}}$ a noise distribution, $\boldsymbol{x}$ denotes an optical input patch, $\boldsymbol{y}$ the corresponding SAR output or target patch (the ground truth image patch) and $G(\boldsymbol{x},\boldsymbol{z})$ the artificially generated SAR-like patch.

To further control the image generation process and to force $G$ to produce artificial images patches, which are similar to the ground truth SAR patches $\boldsymbol{y}$ (in the sense of the $L_1$ distance) we follow the idea proposed in [171] and extend Equations 4.4 by the following regularization term

$$\mathcal{L}_{L_1}(G) = E_{\boldsymbol{x},\boldsymbol{y}\sim p_{\text{data}}(\boldsymbol{x},\boldsymbol{y}),\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\|\boldsymbol{y}-G(\boldsymbol{x},\boldsymbol{z})\|_1] , \tag{4.5}$$

which influence only the learning of $G$ and not of $D$. A positive impact on the image generation process through such an additional term was also reported in [175]. Here, the $L_2$ distance was utilized, which leads in contrast to the $L_1$ distance to slightly blurred output images. By combining both losses the final objective can be expressed as

$$G^* = \arg\min_G \max_D \mathcal{L}_{\text{cGAN}}(G,D) + \lambda \mathcal{L}_{L_1}(G) . \tag{4.6}$$

The conceptional training idea of cGAN is the same as for unconditional GANs: $D$, a binary classification network, tries to distinguish as good as possible between real images and images $G(\boldsymbol{z})$ generated by $G$, whereas $G$ tries to produce more and more realistic images to "fool" D as often as possible. The only difference to the GANs training, is the additional input data $\boldsymbol{x}$, which controls the generation process and forces $G$ to produce images that exhibit geometric properties of the input images $\boldsymbol{x}$ and radiometric properties of the target images $\boldsymbol{y}$. The training of both networks takes place simultaneously, where the discriminator network $D$ is alternately trained on two different kinds of training pairs. Half of the training pairs

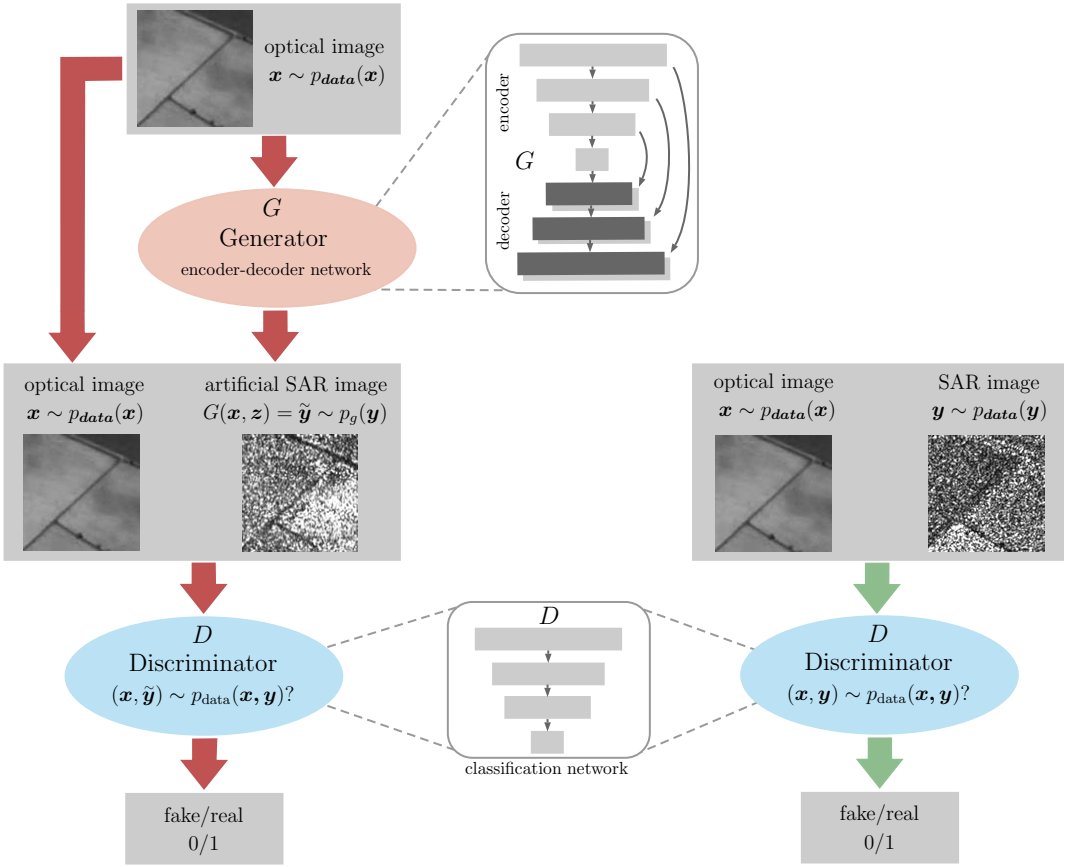



**Figure 4.9:** Overview of cGAN training procedure. On the left side the training setup for "fake" examples (optical and artificially generated SAR patch pairs) as input for the discriminator $D$ and on the right side the training setup for "real" examples (optical and SAR patch pairs) as input for $D$.

are "fake" examples and are composed of optical and artificially generated SAR-like patch pairs. The other half are "real" examples and are composed of optical and SAR patch pairs. An illustration of the two different training setups are shown in Figure 4.9. As proposed in [171] we do not feed noise directly (in form of noise image samples) as additional input into the network. Instead dropout [65] is used inside the generator network as form of noise and regularization in order to prevent the generator from overfitting.

**Network Architectures:** For our application, we utilize the proposed network architectures from [171]. More precisely, the generator $G$ is realized via a U-Net [176], which is an encoder-decoder type of network with skip connections between layer $i$ and layer $L-i$, where $L$ is the total number of layers and $i \in \{1, ..., L\}$. Here, a skip connection between the layers $i$ and $L-i$ means to concatenate all channels of layer $i$ with those of layer $L-i$. U-Nets have proven their high potential for several tasks in the area of image-to-image translation [175, 177, 178] and commonly consist of two parts. The first part, the encoder, consists of $L_e = 8$ convolutional layers and pursues the goal of extraction low to high order features by gradually downsampling the input image with a size $256 \times 256$ pixels to a $1 \times 1 \times 512$ dimensional feature map. The last layer of the encoder is often called the bottleneck. The second part, the decoder, also consists of $L_d = 8$ convolutional layers and pursues the goal of assemble the overall network output from the provided information by gradually upsampling

the encoders output to an image with an size of $256 \times 256$ pixels. The skip connections are additional connections between layers of the en- and decoder and enable the transmission of extracted low level feature information from encoder layers to layers of the decoder, and hence prevent the loss of information by circumvent the bottleneck. An example of such an network architecture is shown in Figure 4.9. Each encoding layer utilizes spatial convolutions of size $4 \times 4$ with a stride of 2, spatial batch normalization [57] and a rectified linear unit (ReLU) as an activation functions. The decoding layers are constructed the same way but utilizes $4 \times 4$ transposed convolutions [164] to realize the upsampling. The number of computed features maps, sometimes called channels, of the encoding layers $i_e \in \{1, ..., 4\}$ are $2^{5+i_e}$ and of the layers $i_e \in \{5, ..., L_e\}$ are 512, whereas the number of feature maps of the decoding layers are $512, 1024, 1024, 1024, 1024, 512, 256$ and 128. A detailed explanation of the usage of spatial batch normalization and the merits of rectified linear units over other activations function can be found in Subsection 4.3.2. For more details about the design choice or architectural details about the utilized U-Net we refer to [171, 176].

The discriminator $D$ on the other hand, is realized via a binary classification network. The classification network consists of five convolutional layers and takes a set of stacked training pairs (either a "real" or a "fake" example) as input. Each of the five layers utilizes spatial convolutions of size $4 \times 4$, spatial batch normalization (except layer one and five) and a rectified linear unit as an activation function (except layer five). The length of the stride in layer one to three is 2 and in layer four and five 1. The feature map size gradually decreases with the depth of the networks and is $64, 128, 256, 512$ and 1 for the five layers, respectively. These five layers incrementally downsample the network input to a $1 \times 30 \times 30$ dimensional output matrix. This is contrary to other methods, since the output of the last layer does not provide a single value for the classification of the input patch pair. Instead, the classification is based on regarding the local structures of the input by provides one classification value for 900 overlapping $70 \times 70$ pixels large sub-patches of input patches. In order to obtain an overall network output the average over all single responses is computed and in a final step, mapped via a Sigmoid function to the interval $[0, 1]$. This enables the consideration of the network output as a form of probability that the given input pair belongs to the class "real" or "fake" (1 or 0). For a more detailed overview we refer to [171, 179].

**Network Training:** The networks are trained with stochastic gradient descent and the adaptive moment estimation (ADAM) optimizer [63]. ADAM is a computationally efficient optimization algorithm, which is developed for machine learning problems and well established in the field of deep learning due to its faster convergence compared to other stochastic optimization methods while providing accurate results [63, 180]. The training of both networks takes place at the same time by alternating the training of $D$ and $G$. More precisely, one gradient descent step of $D$ is followed by one gradient descent step of $G$. As described in Subsection 2.2.2, a frequent problem of a GAN training in combination with the $\mathcal{L}_{\text{GAN}}$ loss from Equation 2.22 are vanishing gradients. For the described cGAN setup with the loss from Equation 4.4 the same problem can occur. Therefore, to limit the problem of vanishing gradients and hence improve the quality of the image generation process, we follow the common practice for the training of $G$, which is to maximize $\log(D(\boldsymbol{x}, G(\boldsymbol{x}, \boldsymbol{z})))$ instead of minimizing $\log(1 - D(\boldsymbol{x}, G(\boldsymbol{x}, \boldsymbol{z})))$. For more details about the cGAN training procedure

see Algorithm 1 and for the selection of the set of suitable hyperparameters we refer to Subsection 5.2.1.

**Image Generation (Network Testing):** After the training process only the generator network is needed for the process of artificial image generation (see the graphical overview of the framework in Figure 4.10). Therefore, a set of optical image patches is feed into the network from which the generator is able to produce within seconds the corresponding set of SAR-like image patches. To enable a fair evaluation of the generators abilities, the set of optical images patches should be unseen, which means that it was not utilized in the training process. During the image generation phase, the weights of $G$ are retained and not modified in any way. Note that through retaining the weights one input patches will always lead to the same artificial output patch.

The cGAN setup described so far has proven to be particularly suitable for the task of image-to-image translation by generating high quality image samples [171]. However, a common problem of (conditional) GANs with an objective function that is based on the negative log-likelihood (see Equation 4.4), is an unstable course of training due to vanishing gradients. Despite the above mentioned change of the objective function (maximize $\log(D(\boldsymbol{x}, G(\boldsymbol{x}, \boldsymbol{z})))$ instead of minimizing $\log(1 - D(\boldsymbol{x}, G(\boldsymbol{x}, \boldsymbol{z})))$) the problem of vanishing gradients cannot be completely avoided. Recent research studies like [181, 182] try to overcome this problem by describing more stable training procedures. Therefore, we introduce two alternative cGAN setups and investigate their influence on the stability of the image generation process and on the quality of the later image matching in order to obtain an optimal tie point generation framework.

The first alternative cGAN setup was proposed in [181] and only requires a change in the loss function $\mathcal{L}_{\text{cGAN}}$ from Equation 4.4. It is based on the following observation: The common (c)GAN losses from Equations 2.22 and 4.4 only slightly penalize generated image samples, which are correctly classified by $D$ but are far away from the real data distribution $p_{\text{data}}$. Therefore, such examples cause almost no error during the training and can lead to vanishing gradients. In order to tackle this problem, Mao et al. [181] proposed the usage of a least squares loss. Such a loss penalizes image samples based on their distance to the decision boundary and is therefore able to penalize samples far away from the real data distribution even though they are correctly classified by $D$. The utilized least square loss $\mathcal{L}_{\text{cLSGAN}}$ is defined as follows

$$\begin{aligned}\mathcal{L}_{\text{cLSGAN}}(G, D) = &E_{x,y\sim p_{\text{data}}(x,y)}[(D(x, y) - 1)^2] \\ &+ E_{x,y\sim p_{\text{data}}(x,y), z\sim p_z(z)}[D(x, G(x, z)))^2]\,.\end{aligned} \tag{4.7}$$

We call the new cGAN setup, where the least square loss is utilized, cLSGAN. Besides the replacement of the cGAN loss from Equation 4.4 with the cLSGAN loss from Equation 4.7 no other changes of the above described method are necessary nor performed, e.g. in the network architecture or during training. Mao et al. [181] showed in an extensive study that the cLSGAN setup is able to generate higher quality images samples and provides a more stable training procedure that is less vulnerability to vanishing gradients in comparison to the common cGAN setup.

The second alternative approach was proposed in [182] and also focuses on the implementation of an improved training objective. The general aim of each (c)GAN training procedure is to learn a model, which is able to generate data samples from the real data distribution $p_{\text{data}}$. In order to reach this goal, the distance between $p_{\text{data}}$ and the learned generator distribution $p_g$ has to be measured during the training. In Subsection 2.2.2 we showed that optimizing the common GAN objective from Equation 2.22 is equivalent to minimizing the Jensen-Shannon (JS) divergence. This means that for the common (c)GAN setups the JS divergence is utilized to measure the distance between $p_{\text{data}}$ and $p_g$ during the training process. However, Arjovsky et al. [182] showed that from an optimization perspective the JS divergence is not optimal since it is not ensured that the gradient is always well-defined during the training process. Additionally, the gradients tend in some situation to be always zero, which causes vanishing gradients, and hence lead to an unstable training process. Therefore, they proposed to use of an alternative distance on the basis of which an improved (c)GAN setup can be build.

The distance utilized in [182] is called the Wasserstein (or Earth Mover distance) and is defined as follows

$$\mathcal{W}(p_{\text{data}}, p_g) = \inf_{\gamma \sim \Pi(p_{\text{data}}, p_g)} E_{(\boldsymbol{x}, G(\boldsymbol{z})) \sim \gamma} \left[ \|\boldsymbol{x} - G(\boldsymbol{z})\| \right] , \tag{4.8}$$

where $\Pi(p_{\text{data}}, p_g)$ denotes the set of all joint probability distributions $\gamma(\boldsymbol{x}, G(\boldsymbol{z}))$. The joint probability distributions $\gamma$ can be interpreted as a transport plan, which describes how much mass has to be moved from $\boldsymbol{x}$ to $G(\boldsymbol{z})$ in order to transform $p_g$ into $p_{\text{data}}$. Consequently, the Wasserstein distance gives the minimal cost or effort that has to be spend to transform $p_g$ into $p_{\text{data}}$ by following the optimal transport plan. The advantages of the Wasserstein distance over the JS divergence is that it provides an novel GAN setup which exhibit a continuous and almost everywhere differentiable loss function. The benefit of such a loss are well-defined gradients during the whole training, which leads to an improved and more stable training process of the generator and discriminator network.

However, since the computation of all possible transport plans, and hence the exact computation of the Wasserstein distance is intractable, an approximation of the Wasserstein distance has to be found. Therefore, the authors utilized the Kantorovich-Rubinstein duality [183] which leads under some assumption to the following approximated Wasserstein distance (for more theoretical details about this derivation see [182])

$$\mathcal{W}(p_{\text{data}}, p_g) \approx \max_{\boldsymbol{\theta}} E_{\boldsymbol{x} \sim p_{\text{data}}} \left[ f_{\boldsymbol{\theta}}(\boldsymbol{x}) \right] - E_{\boldsymbol{z} \sim p_z} \left[ f_{\boldsymbol{\theta}}(G(\boldsymbol{z})) \right] . \tag{4.9}$$

Here, the function $f$ is called the critic (in the common GAN setup the discriminator) and is defined by the set of parameters $\boldsymbol{\theta}$. In contrast to the common (c)GAN setup, the critic does try to classify the input images into "real" or "fake". Instead it tries to compute the Wasserstein distance between $p_g$ and $p_{\text{data}}$. By minimizing Equation 4.9 the generator will learn to produce more and more realistic looking image samples and as a consequence, $p_g$ will get closer to $p_{\text{data}}$. Bringing all together and adapting it to the case of cGANs, lead to a novel cGAN setup called cWGAN, which is based on the following loss

$$\begin{aligned} \mathcal{L}_{\text{cWGAN}}(G, D) = & E_{\boldsymbol{x}, \boldsymbol{y} \sim p_{\text{data}}(\boldsymbol{x}, \boldsymbol{y})}[D(\boldsymbol{x}, \boldsymbol{y})] \\ & - E_{\boldsymbol{x} \sim p_{\text{data}}(\boldsymbol{x}), \boldsymbol{z} \sim p_{\boldsymbol{z}}(\boldsymbol{z})}[D(\boldsymbol{x}, G(\boldsymbol{x}, \boldsymbol{z}))] . \end{aligned} \tag{4.10}$$

In order to ensure the theoretical benefits of minimizing the Wasserstein distance instead of the JS divergence even in the case of utilizing the approximated Wasserstein distance, the previously described cGAN training process has to be adjusted. First, to guarantee a continuous and accurate estimation of the Wasserstein distance during training the discriminator have to be trained to optimality. Therefore, it is recommended in [182] to train the discriminator more often than the generator. We follow the common practice and perform five gradient descent steps of $D$ followed by one gradient descent step of $G$. Second, the weights of the discriminator have to be limit to ensure that the discriminator function is restricted, and hence that an optimal set of parameters can be computed. In practice, the weights are clipped to the interval from $-0.01$ to $0.01$ after each iteration of $D$ during the training process. Third, Arjovsky et al. [182] reported the risk of an unstable training when utilizing the ADAM optimizer for the WGAN training. Therefore, the use of the RMSProp optimizer [62] is recommended, which is another gradient descent optimization algorithm that automatically adapts the learning rate during the training procedure. An overview of the cWGAN training procedure can be found in Algorithm 2 and for more details about the theoretical analysis of the Wasserstein GANs and its characteristics we refer to [182].

**Algorithm 2:** cWGAN training procedure with stochastic gradient descent.

**Input:** A training dataset $\mathcal{D}_{\text{train}}$, a noise distribution $p_{\boldsymbol{z}}$, the learning rate $\lambda$, the batch size $N_b$, the number of training iteration $n_{\text{train}}$, the clipping parameter $t_{\text{clip}}$, the number of discriminator iterations $n_{\text{disc}}$ per one generator iteration

**for** $i = 1, ..., n_{\text{train}}$ **do**

  **for** $i = 1, ..., n_{\text{disc}}$ **do**

    - Sample a mini-batch $\left\{\boldsymbol{z}^{(i)}\right\}_{i=1}^{N_b}$ from the noise distribution $p_{\boldsymbol{z}}$ and a mini-batch $\left\{\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}\right\}_{i=1}^{N_b}$ from the set of real training data $\mathcal{D}_{\text{train}}$ with distribution $p_{\text{data}}$
    - Compute the stochastic gradient $\boldsymbol{g}^{(D)}$ of $D$ w.r.t. its parameters $\boldsymbol{\theta}^{(D)}$:
      $\boldsymbol{g}^{(D)} \leftarrow \nabla_{\boldsymbol{\theta}^{(D)}} \frac{1}{N_b} \sum_{i=1}^{N_b} \left[ D\left(\boldsymbol{y}^{(i)}\right) - D\left(\boldsymbol{x}^{(i)}, G\left(\boldsymbol{x}^{(i)}, \boldsymbol{z}^{(i)}\right)\right)\right]$
    - Update the parameters of $D$ via the optimization algorithm RMSProp:
      $\boldsymbol{\theta}^{(D)} \leftarrow \boldsymbol{\theta}^{(D)} + \lambda \text{ RMSProp}(\boldsymbol{\theta}^{(D)}, \boldsymbol{g}^{(D)})$
      $\boldsymbol{\theta}^{(D)} \leftarrow \text{clip}(\boldsymbol{\theta}^{(D)}, -t_{\text{clip}}, t_{\text{clip}})$

  **end**

  - Sample a mini-batch $\left\{\boldsymbol{z}^{(i)}\right\}_{i=1}^{N_b}$ from the noise distribution $p_{\boldsymbol{z}}$ and a mini-batch $\left\{\boldsymbol{x}^{(i)}\right\}_{i=1}^{N_b}$ from the set of real training data $\mathcal{D}_{\text{train}}$ with distribution $p_{\text{data}}$
  - Compute the stochastic gradient $\boldsymbol{g}^{(G)}$ of $G$ w.r.t. its parameters $\boldsymbol{\theta}^{(G)}$:
    $\boldsymbol{g}^{(G)} \leftarrow \nabla_{\boldsymbol{\theta}^{(G)}} - \frac{1}{N_b} \sum_{i=1}^{N_b} \left[ D\left(\boldsymbol{x}^{(i)}, G\left(\boldsymbol{x}^{(i)}, \boldsymbol{z}^{(i)}\right)\right)\right]$
  - Update the parameters of $G$ via the optimization algorithm RMSProp:
    $\boldsymbol{\theta}^{(G)} \leftarrow \boldsymbol{\theta}^{(G)} - \lambda \text{ RMSProp}(\boldsymbol{\theta}^{(G)}, \boldsymbol{g}^{(G)})$

**end**

### 4.2.3 Tie Point Generation Through Artificial Images Matching

Several approaches exist to realize the matching between artificially generated image patches and the corresponding reference images in order to compute a set of tie points. In our investigations we focus on two intensity-based (NCC and MI) and two feature-based approaches (SIFT and BRISK). These matching approaches have proven their high quality in the case of single-sensor matching, but usually lead to inaccurate results for the matching of optical and SAR images. For our investigation we will evaluate two aspects: 1) Can we improve the conditions for the traditional matching approaches NCC, MI, SIFT and BRISK, and hence improve the quality of the obtained matches and 2) can we obtain accurate and reliable tie points?

Intensity-based approaches measure the similarity between a template **T** and a larger reference image **R** at all locations within the search space. In the later evaluation in Subsection 5.2.3, the template **T** will either be a patch cropped from the optical image or the generated artificial SAR-like patch and **R** a patch cropped from the SAR image. We use a sliding window technique to compute the NCC- or MI-value for every location of **T** within **R** (see Subsection 3.1.1 for details about the NCC- and MI-value computation). The correct matching position is given by the highest NCC-value within the search space. Since we are only interested in reliable and accurate tie points, the raw NCC- and MI-values can be used as a quality measure to detected outliers in the set of tie points. More precisely, by setting a certain threshold unreliable matches can be removed and the quality of the obtained set of tie points can be increased.

In contrast, feature-based approaches are based on the detection of features in both images, called key points, and the measurement of their similarity in the feature space. The two feature detectors utilized for our evaluation are SIFT and BRISK. The idea of both algorithms is to find key points in **T** and **R** and to return a descriptor for every key point. The descriptors of two images are then matched by utilizing the Euclidean distance for SIFT and the Hamming distance for BRISK in combination with a nearest neighbor search (details of the SIFT and BRISK feature descriptor computation can be found in Subsection 3.1.2). To increase the quality and reliability of the detected tie points we remove outliers through RANSAC [184] with an underlying affine model and by setting a distance threshold. More details about the selected threshold for the intensity- and features-based artificial template matching are outlined in Subsection 5.2.1.

### 4.2.4 Summary

The overall tie point generation framework introduced in this section can be summarized in the following five steps:

1. Select suitable matching areas through the framework described in Subsection 4.1.1 and generate a set of optical and SAR training pairs.

2. Train the cGAN, cLSGAN or cWGAN setup on the set of optical and SAR training patch pairs.

3. Apply the trained generator network on a set of unseen optical image patches (these patches should be cropped from the optical images to be registered) and generate the corresponding artificial SAR-like image patches.
4. Apply a NCC-, MI-, SIFT- or BRISK-based image matching on the computed artificial SAR-like patches and the corresponding real SAR image patches.
5. Remove outliers using a RANSAC framework in order to obtain the final set of tie points.

The proposed novel tie point generation method open up the possibility of an automatic and widely applicable tie point generation framework while providing the following benefits:

- The exclusion of ineligible matching areas increases to reliability of the later computed tie points.
- No handcrafted feature detection and extraction algorithms during the artificial image generation process are used nor required.
- Due to the utilization of a cGAN the image generation process is not limited to particular features such as roundabouts.
- Due to the learned image-to-image translation radiometric differences between the artificially generated patches and the corresponding SAR patches can be reduced to a minimum, and hence the applicability of traditional matching approaches such as NCC, MI, SIFT and BRISK for the tie point generation becomes feasible.

An extensive evaluation of this framework is provided in Section 5.2. This evaluation includes the comparison of the three introduced image generation setups (cGAN, cLSGAN and cWGAN) in terms of their ability for the high quality image generation and their applicability for an accurate and precise tie point generation process implemented on the basis of a NCC-, MI-, SIFT- and BRISK-based image matching.

## 4.3 Convolutional Neural Networks for Multi-modal Image Matching

Our second framework for the generation of tie points between optical and SAR images [185] is based on a specific convolutional neural network architecture, called Siamese neural network, and enables the end-to-end learning of an accurate and reliable tie point generation method. As discussed in Subsection 3.2.4, traditional optical and SAR image matching approaches are based on handcrafted feature detection, extraction and matching algorithms. Additionally, these methods are usually tailored to fulfill the needs of certain image features, and hence are not applicable to a wide range of images acquired over different areas or at different times of the year. In the previous sections, we tried to overcome some of these problems by first, eliminating geometric differences by pre-selecting suitable matching areas (see Section 4.1) and second, eliminating radiometric differences through a cGAN-based image-to-image translation framework (see Section 4.2). However, the cGAN-based tie point generation framework still depends on the success of two separated steps: 1) the artificial image generation process and 2) the accurate and precise image matching through handcrafted matching approaches.

In order to obtain an end-to-end tie point generation framework, which does not rely on a single handcrafted processing step, we propose in this section a deep learning-based image matching approach for pre-selected areas, where geometric differences are reduced to a minimum. Inspired by the successful use of Siamese neural networks for the task of image matching (discussed in Subsection 3.3), we base our method on this kind of network architecture. In contrast to the common deep learning-based matching approaches utilized for tasks such as stereo matching or optical flow estimation, our input images are acquired from different sensor types with different radiometric properties and exhibit a lower level of detail. Additionally, due to the speckle in the SAR images, the pre-processing of the images plays in our case an important role for the training process and for the resulting matching accuracy and precision at test time. These circumstances entail the need for a careful adaption of common Siamese neural network architectures, e.g. find the right tradeoff between the number of parameters, the number of layers, and more important the receptive field size. Nevertheless, by implementing a suitable Siamese neural network that fulfills the particular requirements of optical and SAR images and by training it on a large dataset containing images spread over different locations and acquired at different times of the year, the network will learn to handle all kind of image changes, e.g. radiometric or small geometric changes of an object over time or at different locations, and hence will be applicable to a wide range of image scenes.

In the following we will outline the general idea of the Siamese neural network-based tie point generation framework (see Subsection 4.3.1). Subsequently, we will introduce the idea behind the tie point generation process through Siamese neural networks, establish the selection of the final network architectures, give details about the training process and about the final tie point generation through the learned network (Subsection 4.3.2). At last, a summary of the whole process and an outline about theoretical benefits of the proposed method compared to the state-of-the-art is provided in Subsection 4.3.3.

### 4.3.1 Concept of Optical and SAR Image Matching Through Siamese Neural Networks

Our concept for the generation of tie points through the use of Siamese neural networks includes the following three steps: First, to suppress geometrical differences between optical and SAR patches, we focus our training on patches containing flat surfaces like streets or runways in rural areas. Towards this goal, the semi-automatic approach described in Subsection 4.1.1 is utilized to identify suitable matching areas and to extract optical and SAR patch pairs, where the cropped SAR image patches are larger in size compared to the optical patches. Second, a Siamese neural network is trained in order to find the correct location (with the highest similarity) of optical patches within the corresponding larger SAR patches. More precisely, during the training process the Siamese neural network learns to extract important image features from both patches via two independent CNNs. Subsequently, the dot product is utilized to measure the similarity between the extracted features vectors. The resulting network output is a score map for each input pairs and contains a similarity value for every location of a smaller optical patch within the corresponding larger SAR patch. In a last step, tie points are selected with the help of the score maps and confidence scores, which are provided by the network and enable the removal of outliers. The described framework is visualized in Figure 4.10 and will be described in detail in the following subsection.

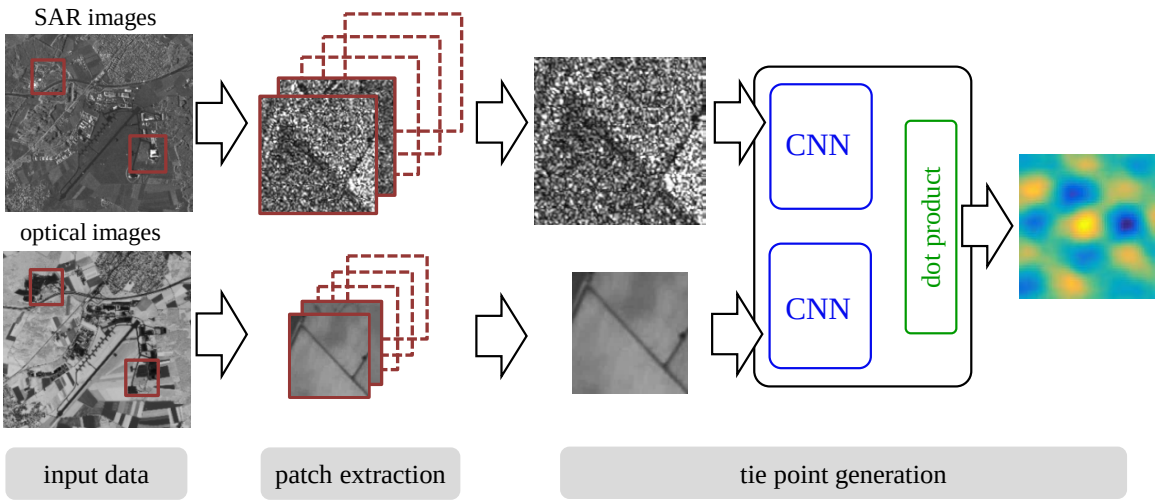



**Figure 4.10:** Graphical overview of the tie point generation framework based on a deep learning-based image matching process. Tie points are generated by training a Siamese neural network, which step by step learns to measure the similarity between optical and SAR patches.

### 4.3.2 Tie Point Generation Through Siamese Neural Networks

In order to learn the automatic computation of tie point between optical and SAR images patches a suitable neural network and a corresponding training concept has been developed. Due to the experiences and insight of previous research studies in the context of deep learning-based image matching (see Subsection 3.3), we base our framework on a Siamese neural network architecture with the proposed matching concept from [131, 132] and adjust it for the case of optical and SAR input images. In general, Siamese neural networks are composed

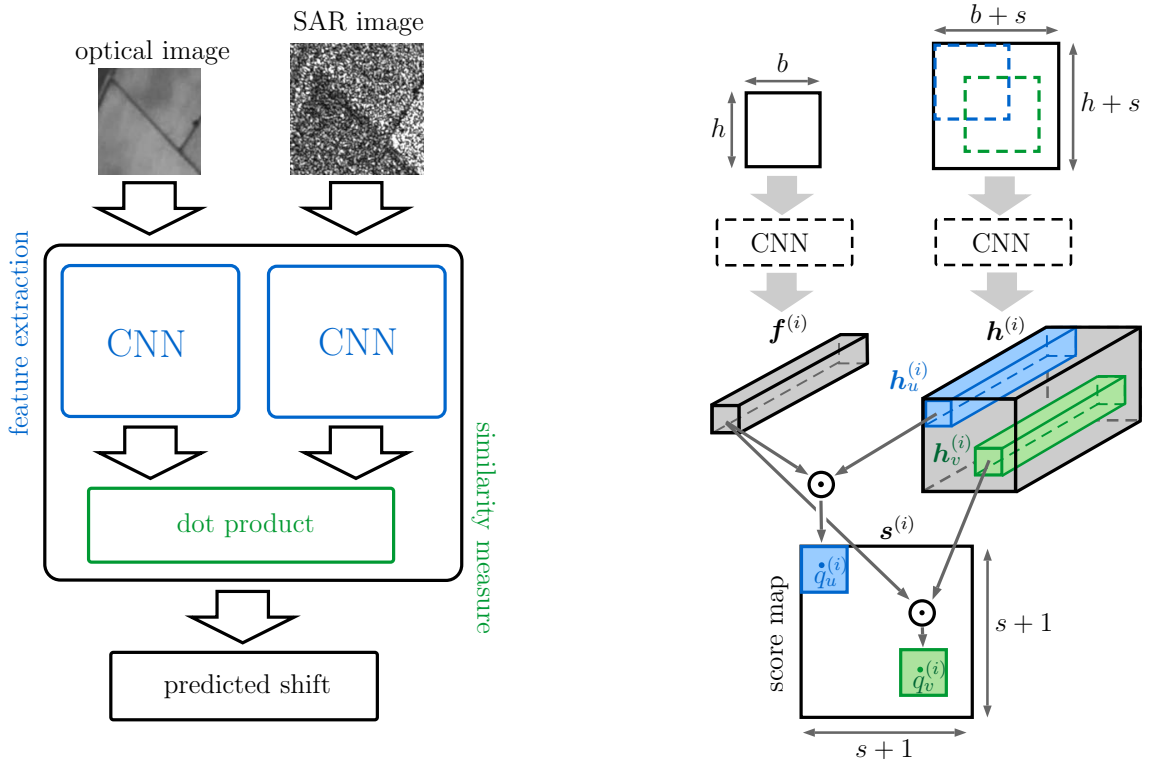



**Figure 4.11:** Illustration of the Siamese neural network architecture (left side) and details of the training concept (right). With the help of two CNNs image features are extracted from the input patch pairs. The resulting outputs of the $i$-th optical patch and of the corresponding SAR patch are a feature vector $\boldsymbol{f}^{(i)}$ and a feature matrix $\boldsymbol{h}^{(i)}$, respectively. The similarity between the extracted features is measured through the use of the dot product, where the value $\boldsymbol{s}_j^{(i)}$ at location $q_j^{(i)}$ of the score map $\boldsymbol{s}^{(i)}$ is given by $\boldsymbol{s}_j^{(i)} = \boldsymbol{f}^{(i)} \cdot \boldsymbol{h}_j^{(i)}$. The correct location (shift) of the optical patch within the larger SAR patch is predicted based on the score map.

of 1) two parallel branches (two sub-neural networks), which pursue the goal of extracting all relevant features from the corresponding input data, and 2) a subsequent part (a fusion or classification network), which pursues the goal of measuring the similarity between the extracted features. The weights of the two branches can be shared (Siamese architecture) or partly-shared (pseudo-Siamese architecture) between each other. A simplified representation of the described Siamese neural network architecture is depicted on the left side of Figure 4.11. In the following we will describe the concept behind the training procedure in detail. For a better understanding we refer to the associated graphical visualization of the training procedure and the utilized terms on the right side of Figure 4.11.

The tie point generation is realized by training a Siamese neural network over a training dataset consisting of optical and SAR image patch pairs, where the smaller optical patches are feed into the left branch and the larger SAR patches into the right branch of the network. In this work, we utilize two CNNs as the two branches of the network, where the used optical training patches have a size of $b \times h$ and the SAR training patches a size of $(b + s) \times (h + s)$. Note that $s$ defines the range of the search space. The search space $S$ has a size of $(s + 1) \times (s + 1)$ and contains all possible location of the optical image within the SAR image with respect to a pixel-wise shift. Given an input image pair, the overall

network output is a two dimensional score map $\boldsymbol{s}$, whose size $(s+1) \times (s+1)$ depends on the size of the search space $S$. The score map $\boldsymbol{s}^{(i)}$ of the $i$-th input image pair contains a similarity score $\boldsymbol{s}_j^{(i)}$ for each location $q_j^{(i)}$ in the search space ($j \in J = \{1, \ldots, |S|\}$, where $|S|$ is the cardinality of $S$). The search space index $J$ is indexing the two dimensional search space, where each position $q_j^{(i)}$ in $S$ corresponds to a specific two dimensional shift of the left optical patch with respect to the larger SAR patch.

The first step to compute a score map for every input image pair of the training set is to apply the two CNNs. The task of the CNNs is to extract all relevant features from the input data and to provide a feature vector $\boldsymbol{f}$ for each optical training patch and a feature matrix $\boldsymbol{h}$ for the corresponding SAR patches. The feature vector $\boldsymbol{f}^{(i)}$ is the output of the left network branch and is a representation of the $i$-th optical training patch. The dimension of $\boldsymbol{f}^{(i)}$ is $c$, where $c$ is number of feature maps (number of applied filters in the last convolutional layer). The feature matrix $\boldsymbol{h}^{(i)}$ is the output of the right network branch and a representation of the $i$-th SAR training patch. The matrix $\boldsymbol{h}^{(i)}$ has a dimension of $|S| \times c$ and is composed of $|S|$ feature vectors $\boldsymbol{h}_j^{(i)}$ (one for each location in the search space). The second step is to compute the similarity between the features vectors $\boldsymbol{f}^{(i)}$ and $\boldsymbol{h}_j^{(i)}$ for every position $q_j^{(i)} \in S$. We are utilizing the dot product in order to measure the similarity between the two vectors $\boldsymbol{f}^{(i)}$ and $\boldsymbol{h}_j^{(i)}$, and hence obtain the similarity scores $\boldsymbol{s}_j^{(i)} = \boldsymbol{f}^{(i)} \cdot \boldsymbol{h}_j^{(i)}$ for all $j \in J$. Note that a high value of $\boldsymbol{s}_j^{(i)}$ indicates a high similarity between the two vectors $\boldsymbol{f}^{(i)}$ and $\boldsymbol{h}_j^{(i)}$ at location $q_j^{(i)}$. In other words, a high similarity score $\boldsymbol{s}_j^{(i)}$ indicates a high similarity between the $i$-th optical patch and the $i$-th SAR patch at location $q_j^{(i)}$ in our search space.

In order to get a calibrated score over all locations within the search space we apply the softmax function at each location $q_j^{(i)} \in S$

$$\tilde{\boldsymbol{s}}_j^{(i)} = \frac{\exp\left(\boldsymbol{s}_j^{(i)}\right)}{\sum\limits_{j \in J} \exp\left(\boldsymbol{s}_j^{(i)}\right)}\,. \tag{4.11}$$

The softmax function provides the probability distribution of $q_j^{(i)}$, or the corresponding shift, being the correct location (shift) over all possible locations (shifts) within $S$, where $\tilde{\boldsymbol{s}}_j^{(i)} \in [0, 1]$ and $\sum \tilde{\boldsymbol{s}}_j^{(i)} = 1$ over all $j \in J$. Therefore, we can interpret the values of the calibrated score maps $\tilde{\boldsymbol{s}}$ as probabilities and the values of the score maps $\boldsymbol{s}$ as confidence scores, which indicate the confidence of the network that a specific location (shift) is correct.

By treating the problem as a multi-class classification problem, where the different classes represent the possible positions (shifts) of an optical patch within a larger SAR patch, we can train our network by minimizing the following cross entropy loss

$$\min_{\boldsymbol{\theta}} \sum_{i \in I, j \in J} p_{\mathrm{gt}}\left(q_j^{(i)}\right) \log p^{(i)}\left(q_j^{(i)}, \boldsymbol{\theta}\right) \tag{4.12}$$

with respect to the Siamese neural network parameters $\boldsymbol{\theta}$. Here, $p^{(i)}\left(q_j^{(i)}, \boldsymbol{\theta}\right)$ is the probability of the training sample $i$ at location $q_j^{(i)}$ in our search space $S$, and $p_{\mathrm{gt}}$ is the corresponding ground truth target distribution. Instead of a delta function with non-zero probability

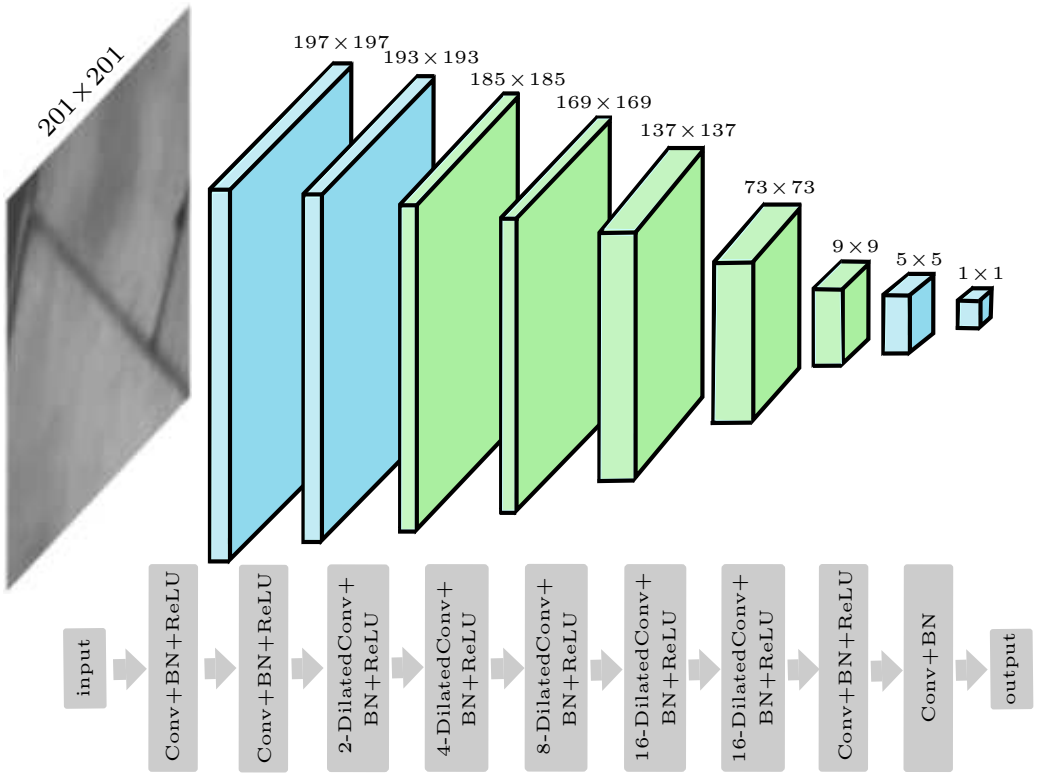



**Figure 4.12:** Detailed overview of the utilized convolutional neural network. The details of the nine layers of the convolutional network are framed in gray and shown on the bottom of the figure. The corresponding example, depicted on the top of the figure, shows the effect (in terms of change in dimensions) of these layers for a given optical input patch. Here, the output of convolutional layers and of dilated convolutional layers are reported in blue and green, respectively. Abbreviations: Convolution (Conv), batch normalization (BN) and rectified linear unit (ReLU).

mass only at the correct location $q_j^{(i)} = q_{\text{gt}}^{(i)}$, we are using a soft ground truth distribution which is centered around the ground truth location. Therefore, we set $p_{\text{gt}}$ to be the discrete approximation of the Gaussian function (with $\sigma = 1$) in an area around $q_{\text{gt}}^{(i)}$

$$p_{\text{gt}}\left(q_j^{(i)}\right) = \begin{cases} \frac{1}{2\pi} \cdot e^{-\frac{\left\|q_j^{(i)} - q_{\text{gt}}^{(i)}\right\|_2^2}{2}} & \text{if} \quad \left\|q_j^{(i)} - q_{\text{gt}}^{(i)}\right\|_2 < 3 \\ 0 & \text{otherwise} \end{cases}, \tag{4.13}$$

where $\|\cdot\|_2$ denotes the $L_2$ (Euclidean) distance. The idea behind utilizing a soft ground truth distribution is to penalize the predictions not only according to their correctness but also to their distance to the correct location. In other words, the soft ground truth enables to put higher penalties on incorrect predictions far away from the correct location while penalizing incorrect predictions that are close to the true location only slightly. Furthermore, without a soft ground truth only one location out of the $|S|$ possible location in our search space $S$ is correct (a positive sample) whereas the other $|S| - 1$ locations are incorrect (negative samples) given one training patch pair. Therefore, the use of a soft ground truth improves the imbalance between positive and negative training samples in our training dataset, and hence improves the quality and speed of your training process.

**Network Architecture:** Our Siamese neural network architecture is based on the architecture proposed in [131] but contains some important adjustments in order to fulfill the particular requirements of optical and SAR image matching (e.g. lower level of detail compared to application in stereo estimation, multi-modal image data). In the following we will describe the selected network architecture and provide reasons for the decisions made, where details a graphically illustrated in Figure 4.12. The basis of our Siamese neural network are the two branches (two CNNs). For the task of single sensor image matching, the weights between the networks are commonly shared in order to utilize learned information from an input image for the extraction of features from the other input image. Furthermore, if the parameters between the two networks are shared, the Siamese architecture provides the advantage of consistent predictions. As both network branches compute the same function, it is ensured that two similar images will be mapped to a similar location in the feature space. However, since the image properties of our input data are quite different we investigate two different setups, one with shared weights (Siamese architecture) the other with partly-shared weights (pseudo-Siamese architecture) between the branches. In the case of our pseudo-Siamese architecture, the weights of the first three layers are different, whereas the remaining layers share their weights.

Both CNNs are composed of nine layers, where each layer consists of three components: 1) a spatial convolution, 2) a spatial batch normalization (BN) [57] and 3) a rectified linear unit (ReLU). The purpose of the convolution layers is to extract spatial features from the input data through trainable filters, where the complexity of the features extracted by the layers increases along with the depth. We employ in all layers convolutions with a filter size of $5 \times 5$ pixels and with a stride of 1 pixel. The number of filters used in layer one to four are 32 and for the others 64. Since our training dataset contains images with a spatial resolution of 2.5 m, the input patches exhibit a lower level of detail in the images compared to the ones used in common matching networks such as in [131, 150–152]. In order to increase the probability of the availability of salient features in the input data, we use optical patches with a size of $201 \times 201$ pixels and SAR patches with a size of $(201 + s) \times (201 + s)$ for training. To achieve that the whole optical input patch and the corresponding area in the larger SAR patch has an impact on our network output, a receptive field size of $201 \times 201$ pixels is desired (the size of the smaller optical input patches). In the context of CNNs, the receptive field refers to the part of the input patches, having an impact on the output of the last convolutional layer. The standard ways to increase the receptive field, such as strided convolutions or pooling layers inside the neural network, always involve a loss of information as these approaches reduce the resolution of the image features. In contrast, dilated convolutions [186] systematically aggregate information through an exponential growth of the receptive without losing resolution. The dilated convolution $*_d$ at a given position $p$ in the image $\boldsymbol{I}$ is defined as

$$(\boldsymbol{I} *_d \boldsymbol{k})(p) = \sum_{m=-r}^{r} \boldsymbol{I}(p - d \cdot m)\boldsymbol{k}(m)\,, \tag{4.14}$$

where $\boldsymbol{k}$ denotes the filter with size $(2r + 1) \times (2r + 1)$ and $d$ the dilation factor. Instead of looking at local $(2r+1) \times (2r+1)$ regions as with standard convolutions, dilated convolutions look at $[d \cdot (2r + 1)] \times [d \cdot (2r + 1)]$ surrounding regions, which lead to an expansion of

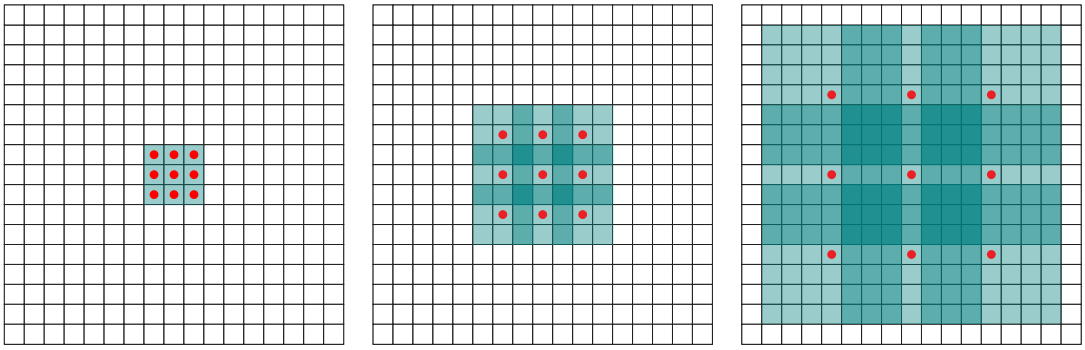


**Figure 4.13:** Illustration of the exponential expansion of the receptive field through dilated convolutions (image source: [186]). The left image shows the output of a standard convolution with a filter size of $3 \times 3$, where each element has a receptive field size of $3 \times 3$. The images in the middle shows the output of a 2-dilated convolution with a filter size of 3 and an obtained receptive field with a size of $7 \times 7$ for each element in the image. The image on the right shows the output of a 4-dilated convolution with a filter size of 3 and an obtained receptive field size of $15 \times 15$ for each element. Note that the number of parameters is the same in all three examples.

the receptive filed size (for an illustration see Figure 4.13). Beyond, dilated convolutions have the same number of network parameters compared to its convolution counterpart. Therefore, we adopt the concept of dilation convolutions to our network architecture in order to overcome the problem of our relatively large input patches and to gradually downsample our $201 \times 201$ dimensional optical input patches to a $1 \times 1 \times 64$ dimensional output and the $(201 + s) \times (201 + s)$ dimensional SAR input patches to a $(s+1) \times (s+1) \times 64$ dimensional output. In practice, we utilize dilated convolutions in the layers three to seven with a dilation factor $d$ of 2, 4, 8, 16 and 16. This setup leads to the desired receptive field size of $201 \times 201$ pixels for each computed feature vector $\boldsymbol{f}$ and each vector $\boldsymbol{h}^{(i)}$ of the feature matrix $\boldsymbol{h}$, and hence ensures that the whole input patches are represented in the network output.

The second component of our layers, batch normalization (BN), is often used as a preprocessing step in order to increase the learning speed and the performance of the network. The idea behind BN is to enable a comparison of the data across all layers by normalize the input of each layer, and hence providing a consistent distribution of each layer input. Additionally, it provides a form of regularization (reduces overfitting) and decreases the dependency of the network performance on the initialization of the weights. Note that next to BN we follow the common practice of normalizing the input data before feeding them into the network, which further enables a comparison between the different input samples. For a detailed overview of BN we refer to [48, 57].

The third component of our layers are non-linear activation functions. Non-linear activation are needed to introduce nonlinearities into the network (otherwise the network can only model linear functions). Therefore, we utilize the most frequently used activation function called ReLU (defined in Subsection 2.2.1). An Advantage of ReLUs compared to other activation function is a more efficient and faster training of the network by decreasing the risk of vanishing gradients. Note that we are not utilizing a ReLU in the last layer of each CNN in order to preserve the information encoded in the negative values. See [48] for more details about ReLUs.

**Network Training:** For the training of the Siamese neural network a training dataset composed of aligned optical and SAR patches is utilized, where the ground truth (correct) location for all training samples is in the center of the larger SAR patches. In order to minimize the loss function from Equation 4.12, and hence train the Siamese neural network to learn the matching between optical and SAR patches, we use stochastic gradient descent with the ADAM optimizer [63]. Note that in the case of shared weights the weights of the two CNNs are identical, and hence only one CNN is actually trained. For more details about the characteristics of the utilized training set see Subsection 5.1.2, for details about the selected set of hyperparameters required for the training process see Subsection 5.3.1 and for an investigation of the effects of shared vs. partly-shared weights see Subsection 5.3.2.

**Tie Point Generation (Network Testing):** After training we keep the learned parameters $\boldsymbol{\theta}$ fixed and decompose the network into two parts: the feature extractors (CNNs) and the similarity measure (dot product layer). As the feature extractor is convolutional, we can apply both CNNs on images with arbitrary size. Thus during test time, we first feed an optical patch as input to the corresponding CNN and compute the feature vector $\boldsymbol{f}$. Then, we feed a larger SAR patch, which covers the desired search space, either to the same CNN (shared weights) or the second CNN (partly-shared weight) and compute the feature matrix $\boldsymbol{h}$. Afterwards, we use the dot product layer to compute the score map $\boldsymbol{s}$ and the calibrated score map $\tilde{\boldsymbol{s}}$ from $\boldsymbol{f}$ and $\boldsymbol{h}$ in the same way as during training. Applying this strategy, we can compute a matching score between optical patches (with arbitrary size) and SAR images over an arbitrary search space. The desired tie points (predicted shifts) are finally computed by selecting for every input pair the points with the highest value (highest similarity between optical and SAR patch) within the corresponding search map. In order to remove outliers we regard the values of raw score map $\boldsymbol{s}$ as the confidence of the network that the provide prediction is correct. For this reason, we set a threshold on the confidence score and remove all tie point from the final set with a confidence score less than the threshold. Note that during the tie point generation process only unseen optical and SAR patches are utilized. Details about the test dataset are provided in Subsection 5.1.2, about the chosen threshold in Subsection 5.3.1 and an investigation about the influence of the outlier removal on the accuracy and precision of the tie points in Subsection 5.3.2.

### 4.3.3 Summary

The overall tie point generation framework introduced in this section can be summarized in the following four steps:

1. Select suitable matching areas through the framework described in Subsection 4.1.1 and generate a set of optical and SAR training pairs.
2. Train the Siamese neural network in order to learn the matching between optical and SAR by extracting relevant features and measuring their similarity.
3. Apply the trained Siamese neural network on a set of previously unseen optical and SAR image pairs (these patches should be cropped from the optical and SAR images to be registered) and generate the corresponding score maps.
4. Select locations with the highest similarity from the score map as tie points and remove outliers through the help of the corresponding confidence scores.

The described optical and SAR image matching framework provides a novel and automatic process for the generation of tie points while providing the following benefits:

- The exclusion of ineligible matching areas increases the reliability of the tie points computed in the subsequent step.
- No handcrafted feature detection, extraction and matching step is required in the tie point generation framework.
- Training the network over a large dataset that contains a variety of different image pairs enables applying our framework to a wide range of images acquired over different cities or at different times of the year.
- Due to the utilized convolutional layers our framework is applicable to different image resolutions and images sizes. In particular, the size of the search space can simply be adjusted according to the assumed offset between the optical and SAR images patches.
- Once the network is trained, tie points can be computed between arbitrary image pairs within seconds.

An evaluation and detailed discussion about the advantages and disadvantages of the proposed approach is provided in Section 5.3. This includes an investigation about the influence of Siamese and Pseudo-Siamese architectures, of outlier removal through the networks confidence and the proposed tie point refinement step on the quality of the final set of tie points. Additionally, the applicability of the obtained tie points for the absolute geo-Localization accuracy enhancement of optical images is discussed in Subsection 5.3.3.

## 4.4 Geo-localization Accuracy Enhancement of Optical Images

The inaccuracy of the absolute geo-localization of the optical satellite data in the geo-referencing process arise mainly from inaccurate measurements of the satellite attitude and thermally affected mounting angles between the optical sensor and the attitude measurement unit. This insufficient pointing knowledge leads to local geometric distortions of orthorectified images caused by the height variations of the Earth surface. To achieve higher geometric accuracy of the optical data, ground control information is needed to adjust the parameters of the physical sensor model. We are following the approach described in [95] to estimate the unknown parameters of the sensor model from GCPs by iterative least squares adjustment. In the following, we will introduce the idea of a physical sensor model (see Subsection 4.4.1) and shortly described the process behind the geo-localization accuracy improvement of optical images (see Subsection 4.4.2).

### 4.4.1 Physical Sensor Model for Direct Georeferencing

In order to set the geometric relation between images and their corresponding ground coordinates, suitable sensor models are required. Physical sensor models, or sometimes called rigorous sensor models, are the most accurate models with respect to an accurate positioning and are built on information such as the type of utilized sensor, the satellite position and the attitude angles [188]. In case of direct georeferencing, the sensor model is based on the collinearity equations and relates a point $\boldsymbol{r}_{\text{object}}^{\text{sensor}}$ in the sensor coordinate system to the corresponding point $\boldsymbol{r}_{\text{object}}^{\text{Earth}}$ in an Earth-bound object coordinate system. The optical images used in this thesis are acquired through the use of a pushbroom scanner system. As mentioned in Subsection 2.1.1, a pushbroom scanner consists of a linear array

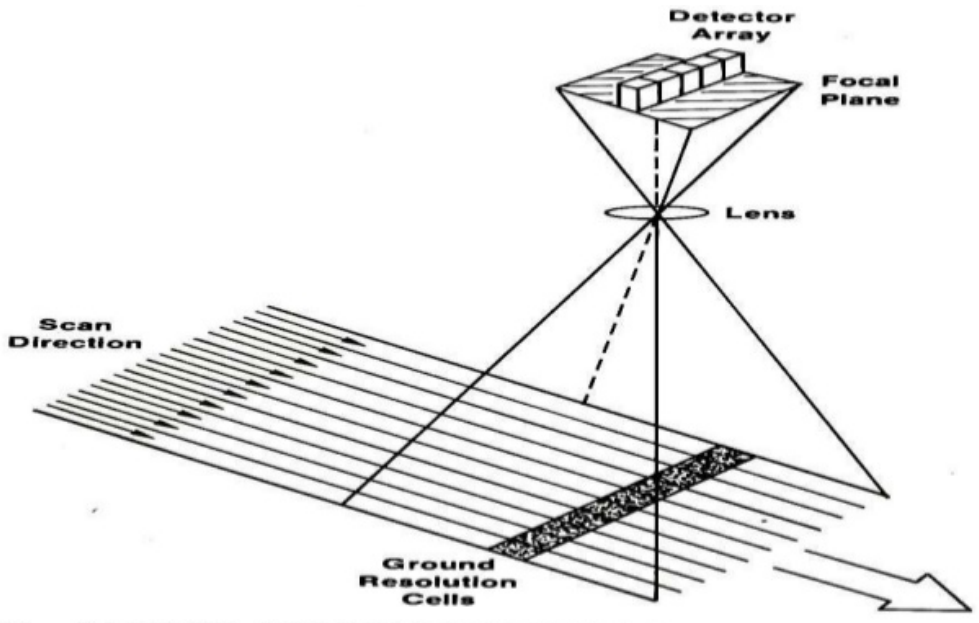



**Figure 4.14:** Illustration of the acquisition principle of a pushbroom scanner system (image source: [187]).

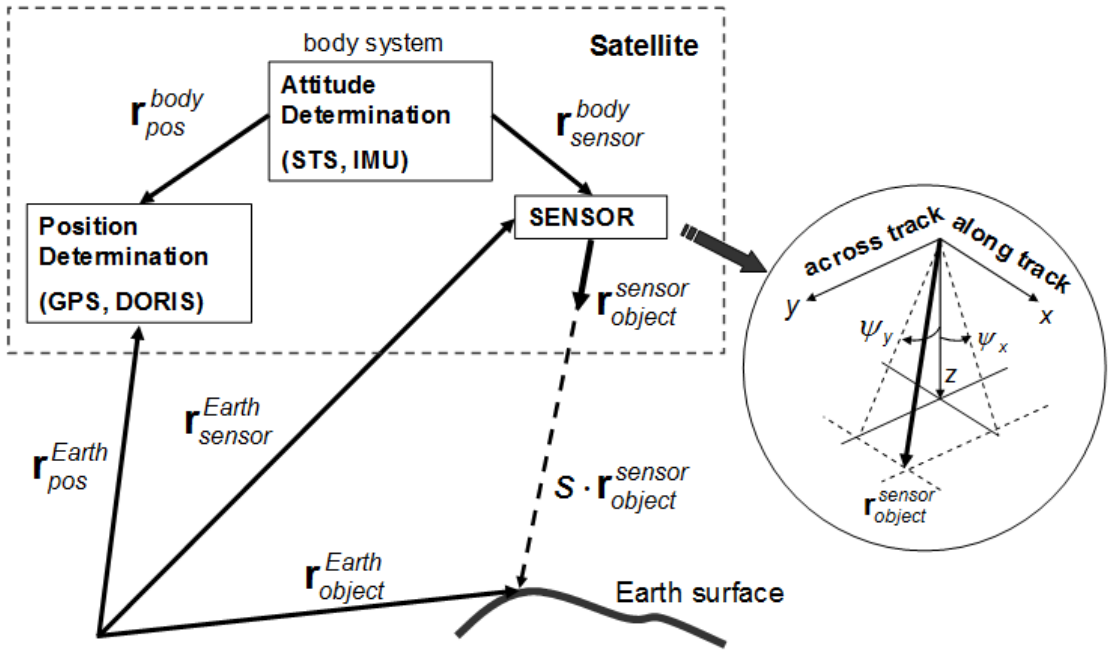



**Figure 4.15:** Illustration of the geometric relation of the utilized physical sensor model (image source: [95]).

of $n$ detectors that are arranged perpendicular to the flight direction of the satellite and simultaneously receive information from the ground [40]. While flying over the ground, the systems records an image line by line, where all pixels in one line are obtained simultaneously from the different detectors (see illustrated in Figure 4.14). As a consequence, the observed image point in the $j$-th scan line is directly related to the recording time $t_j = t_0 + j \cdot \Delta t$, where $t_0$ is the recording time of the first line and $\Delta t$ the sampling time. The geometric relation between an observed image point at location $v_i = i - \frac{n}{2}$ with $i = 0, \ldots, n-1$ and its corresponding ground point can therefore be expressed as

$$\boldsymbol{r}_{\text{object}}^{\text{Earth}}(t_j, v_i) = \boldsymbol{r}_{\text{sensor}}^{\text{Earth}}(t_j) + s_{\text{DEM}}(t_j, v_i) \cdot \boldsymbol{R}_{\text{body}}^{\text{Earth}}(t_j) \cdot \boldsymbol{R}_{\text{sensor}}^{\text{body}}(t_j, v_i) \cdot \boldsymbol{r}_{\text{object}}^{\text{sensor}}(v_i) \ , \quad (4.15)$$

where $s_{\text{DEM}}$ denotes a pixel scaling factor defined by the utilized DEM, $\boldsymbol{R}_{\text{body}}^{\text{Earth}}$ describes the rotation around the three Euler angles $(\omega, \psi, \kappa)$ from the body to the Earth coordinate system (derived from the satellite position and velocity or by orbital parameters) and, $\boldsymbol{R}_{\text{sensor}}^{\text{body}}$ denotes the boresight alignment angles or instrument mounting angles, which describe the rotation around the three Euler angles $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \varepsilon_3)^T$ from the sensor to the body coordinate system (defined by the attitude measurement unit). The term $\boldsymbol{r}_{\text{object}}^{\text{sensor}}$ denotes the mapping of an image point at location $v_i$ to the Cartesian sensor coordinate frame with origin at the sensor projection center and can be stated as $\boldsymbol{r}_{\text{object}}^{\text{sensor}}(v_i) = (\tan \Psi_x(v_i), \tan \Psi_y(v_i), 1)^T$. If the two object sided angles $\Psi_x$ and $\Psi_y$ are measured for a series of focal plane pixels through pre-launched laboratory calibrations. Here, the $x$-axis points along the flight direction and the $y$-axis across track. An illustration of the geometric relation stated in Equation 4.15 and the used terms is provided in Figure 4.15. For more details we refer to [95].

### 4.4.2 Sensor Model Adjustment Through Tie Points

In order to achieve a high positioning accuracy during the geo-referencing process, each parameter of the utilized physical sensor model has to be carefully determined. Commonly, the satellite position and the interior orientation of the camera system can be determined with a high-precision. The relative alignment between the body and the sensor coordinate system on the other hand, causes in most cases pointing errors, mainly due to inaccurate measurements of the satellite attitude and thermally affected mounting angles. As a consequence, additional data in form of well measured GCPs is required in order adjusting the corresponding parameters (the boresight angles) of the physical sensor model. By reformulation Equation 4.15 the following system can be derived

$$\begin{aligned} J_x(\boldsymbol{\varepsilon}) &= \frac{r_{11}\,(x_{oe}-x_{se}) + r_{12}\,(y_{oe}-y_{se}) + r_{13}\,(z_{oe}-z_{se})}{r_{21}\,(x_{oe}-x_{se}) + r_{22}\,(y_{oe}-y_{se}) + r_{23}\,(z_{oe}-z_{se})} - \tan\Psi_x \\ J_y(\boldsymbol{\varepsilon}) &= \frac{r_{21}\,(x_{oe}-x_{se}) + r_{22}\,(y_{oe}-y_{se}) + r_{23}\,(z_{oe}-z_{se})}{r_{31}\,(x_{oe}-x_{se}) + r_{32}\,(y_{oe}-y_{se}) + r_{33}\,(z_{oe}-z_{se})} - \tan\Psi_y\,, \end{aligned} \tag{4.16}$$

where $r_{ij}$ represents an elements of the matrix $\boldsymbol{R}^{\text{sensor}}_{\text{Earth}}(t) = \boldsymbol{R}^{\text{body}}_{\text{sensor}}(\boldsymbol{\varepsilon})^{-1} \cdot \boldsymbol{R}^{\text{Earth}}_{\text{body}}(t)^{-1}$, $\boldsymbol{r}^{\text{Earth}}_{\text{object}} = (x_{oe}, y_{oe}, z_{oe})^T$, $\boldsymbol{r}^{\text{Earth}}_{\text{sensor}} = (x_{se}, y_{se}, z_{se})^T$ and $\boldsymbol{r}^{\text{sensor}}_{\text{object}} = (x_{os}, y_{os}, 1)^T = (\tan\Psi_x, \tan\Psi_y, 1)^T$. The three unknown boresight angles $\boldsymbol{\varepsilon}$ are estimated by minimizing the cost functions $J_x$ and $J_y$ from Equation 4.16 through an iterative least squares adjustment. In order to remove outliers from the given set of GCPs, and hence estimate the unknown angles $\boldsymbol{\varepsilon}$ as precise as possible, an iterative blunder detection is integrated into the least squares adjustment. Here, outliers are defined as GCPs with a residual greater than a certain threshold (usually 1 to 2 pixels), where the residuals are the 2D deviation at the GCPs in image space. A detailed description of the blunder detection step is provided [95].

After estimating the boresight angles and adjusting the sensor model parameters, the improved model and a corresponding DEM are utilized for the orthorectification of given optical images (level-1 products). Through this procedure new orthorectified optical images with an improved absolute geo-localization accuracy can be achieved. Note that in contrast to [95], where the GCPs are generated from optical images, we are using tie points generated by the methods described in the Subsections 4.2 and 4.3. The results of the described sensor model adjustment procedure applied on a set of optical test images and automatic generated tie points are evaluated and discussed in Subsection 5.3.3 and 5.2.4.

## 4.5 Summary

In this section we presented a novel and automatic optical and SAR satellite image registration framework and the associated absolute geo-localization accuracy enhancement of optical images. The three main components of the framework are:

1. Selection of suitable matching areas in order to eliminate geometric differences between optical and SAR images through a semi-automatic process.
2. Generation of a reliable and accurate set of tie points through a deep learning-based matching of optical and SAR image patches cropped from pre-selected areas.
3. Adjustment of the physical sensor model parameters through the generated tie points in order to register optical and SAR images and therefore enhance the absolute geo-localization accuracy of the corresponding optical images.

In contrast to traditional approaches our developed framework provides the following theoretical benefits:

- Through the pre-selection of suitable areas the existence of salient features can be guaranteed and on the other hand areas containing elevated objects and therefore exhibit different geometric properties in optical and SAR images can be eliminated. As a consequence, the risk for our matching approaches to produce total mismatches is reduced and the quality and reliability of the obtained tie points with regard to their geo-localization is increased.
- If the cGAN-based matching approach is utilized for the tie point generation, radiometric differences between arbitrary optical and SAR image pairs can be reduced to a minimum through the generation of artificial SAR-like patches. As a consequence, the application of traditional matching approaches for the tie point generation becomes feasible. In addition, the image generation process is independent of handcrafted feature detection and extraction algorithms and not limited to particular features. This circumstance enables its applicably to a wide range of image scenes.
- If the Siamese neural network-based matching approach is utilized for the tie point generation, no handcrafted feature detection, extraction and matching algorithms are required for a single step and new tie points can be generated within seconds. Furthermore, the end-to-end training over a large dataset and the particular design of our network enable the application to a wide range of images acquired over different scenes, at different times of the year, with different resolutions and image sizes.

In order to assess the proposed framework, we will perform an excessive evaluation of the tie point generation methods and their abilities for a geo-localization accuracy improvement for a set of optical test images in the Sections 5.2 and 5.3 of the following Chapter. Beforehand, the specifics of the utilized optical and SAR images will be presented and the training, validation and test dataset derived from our semi-automatic area selection process will be described in Section 5.1.1. In a final step, the two tie point generation concepts of our registration frameworks will be compared and their strength, weaknesses and potential for future developments will be discussed in Section 5.4.

# 5

# Results and Discussion

*In this chapter the proposed concept for the registration of optical and SAR images through tie points, automatically generated over pre-selected image regions, is tested and evaluated on several image pairs spread across Europe. The main focus of our investigation lies on the evaluation of the two novel tie point generation methods and their ability to generate reliable and accurate tie points. Therefore, the experimental setup with the image characteristics, pre-processing steps and the final datasets for the training, validation and testing of our deep learning based approaches is introduced and an overview of the utilized statistical measures and baseline methods is provided. Then, both tie point generation approaches are consecutively tested on the same test set and compared with state-of the art approaches with regard to their potential for an accurate and precise tie point generation and for an absolute geo-localization accuracy enhancement of optical images. At last, a detailed comparison of the advantages, disadvantages, strength and limitations of both methods is carried out.*

## Contents

## 5.1 Experimental Setup

This section forms the basis for the evaluation of both tie point generation methods (outlined in Chapter 4) and their abilities for the geo-localization accuracy enhancement of optical images. Therefore, the image specifications and pre-processing steps of the utilized optical and SAR image pairs are described in Subsection 5.1.1. Then, the final training, validation and test sets obtained from the semi-automatic matching area selection procedure described in Subsection 4.1.1 are presented in Subsection 5.1.2. Finally, a description of the statistical measures and the baseline methods on which our evaluation are based on is provided in Subsection 5.1.3 and 5.1.4, respectively.

### 5.1.1 Image Specifications and Pre-processing

To perform our experiments we generated training, validation and test datasets out of 46 orthorectified optical (PRISM[i]) and radar (TerraSAR-X[ii] acquired in stripmap mode) satellite image pairs acquired over 13 cities in Europe (see Figure 5.1 for an illustration of the image distribution across Europe). The images cover greater urban zones including suburban, industrial and rural areas with a total coverage of around 20.000 $km^2$. The spatial resolution of the optical images is 2.5 m and the pixel spacing of the of the SAR images is 1.25 m. To have a consistent pixel spacing within the image pairs we downsampled the SAR images to 2.5 m using bilinear interpolation. To enable the possibility to generate a larger

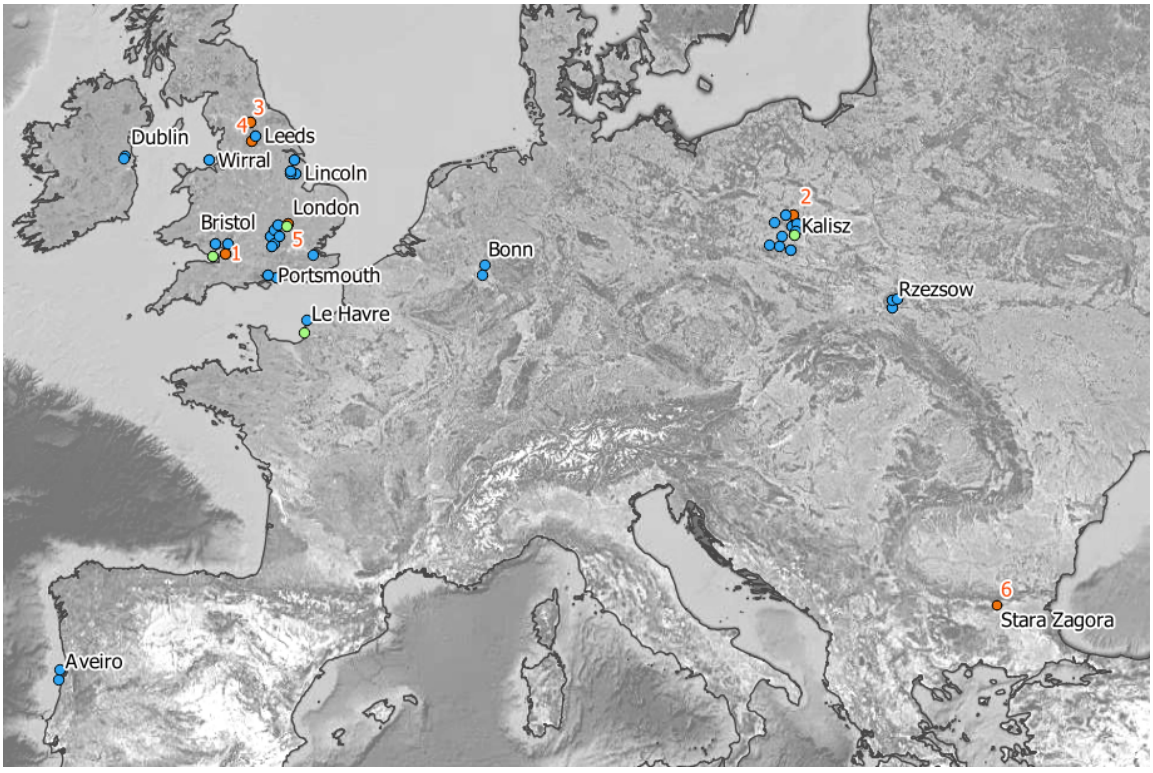
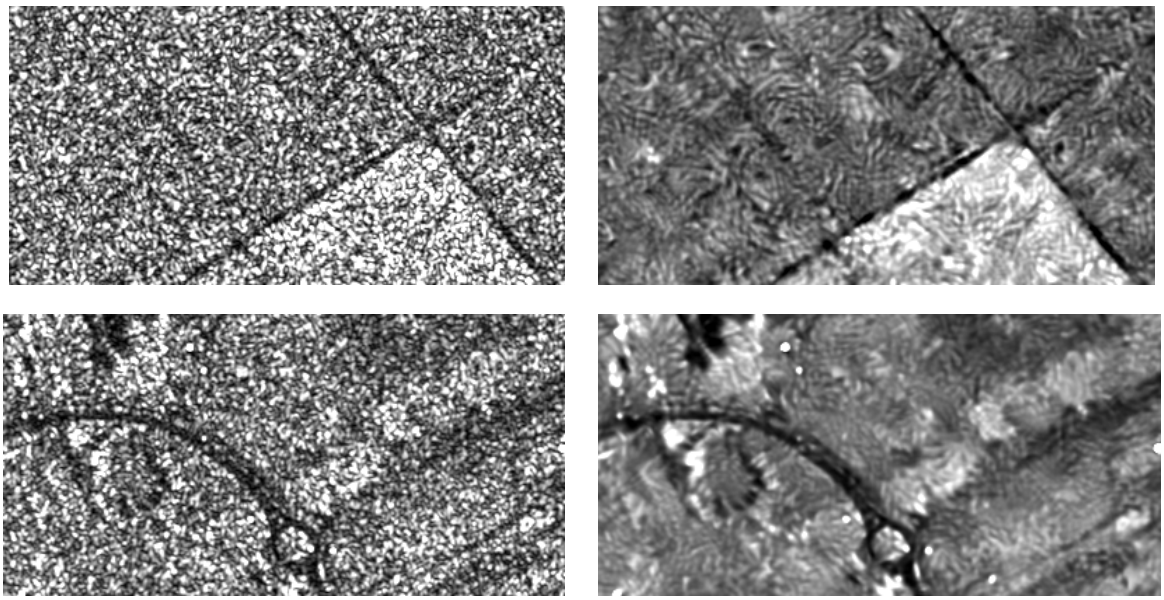



**Figure 5.1:** Overview of the training (blue), validation (green) and test (red) set image locations (image source: [189]).

[i]PRISM: high resolution panchromatic sensor mounted on the satellite ALOS of the Japanese Space Agency
[ii]TerraSAR-X: high-resolution SAR satellite of the German Aerospace Center and EADS Astrium



(a) SAR image samples (b) Despeckled SAR image samples

**Figure 5.2:** Visual comparison between SAR and despeckled SAR image samples by applying the probabilistic patch-based (PPB) filter from [191].

training dataset and to train the networks with multi-resolution data, we further use bilinear interpolation to downsample all optical and SAR images to a pixel spacing of 3.75 m.

All optical and SAR image pairs are aligned and hence, serve as our ground truth (assumed true matching location). The alignment was carried out in the Urban Atlas project [190] where all optical images were manually co-registered to the corresponding SAR images. In order to achieve this, several hundred tie points were manually selected between every image pair. Subsequently, the selected tie points were refined and used to improve the sensor model related to the optical images. For this step, the same procedure as described in Section 4.4 was utilized. By using the improved sensor models to orthorectify the optical images a second time, the global alignment error could be reduced from up to 23 m to around 3 m in this project. Note that the following evaluation of our results (see Sections 5.2 and 5.3) must always be set in relation to this accuracy.

**SAR Image Filtering:** In order to investigate the influence of the SAR image despeckling on the quality of the obtained results we applied the probabilistic patch-based (PPB) filter proposed in [191] for the generation of a filtered SAR image dataset. This filter is developed to suppress speckle in SAR images by adapting the non-local mean filter by Buades et al. [192] to SAR images. The idea of the non-local mean filter is to estimate the filtered pixel value as the weighted average over all pixels in the image. The weights are measuring the similarity between the pixel values of a patch $\Delta_s$ centered around $s$ and the pixel values of a patch $\Delta_t$ centered around $t$. The similarity between two patches is measured with respect to the Euclidean distance. In [191] the noise distribution is modeled using the weighted maximum likelihood estimator. Here, the weights are expressing the probability that two patches centered around the pixels $s$ and $t$ have the same noise distribution under a given image. A comparison between SAR and despeckled SAR patches are shown Figure 5.2.

### 5.1.2 Training, Validation and Test Datasets

As mentioned in Section 2.2, three independent datasets are needed in order to train the networks (training set), find the best set of hyperparameters (validation set) and to evaluate the performance of the networks (test set). The training, validation and test datasets are generated by randomly splitting the 46 image pairs into 36 image pairs for training, 4 for validating and 6 for testing of our methods. An overview of the different image location of our three sets within Europe is provided in Figure 5.1. Through the image pre-processing (described in Subsection 4.1.1), each of the image pairs is available in four different versions: with a resolution of 2.5 m, with a resolution of 3.75 m, with a resolution of 2.5 m and despeckeled SAR images and, with a resolution of 3.75 m and despeckeled SAR images. To minimize the impact of the different acquisition modes of PRISM and TerraSAR-X, we focus on flat surfaces where primarily the radiometry between the optical and SAR images is different. Note that this is not a strong restriction of our approaches since these kind of condition frequently appear in nearly every satellite image. The pre-selection of the images is carried out through the semi-automatic pre-selection process described in Subsection 4.1.1. Subsequently, the training, validation and test patches are cropped from the pre-selected areas of the images of the corresponding sets. The cropped optical patches have a size of $201 \times 201$ pixels and the corresponding SAR patches a size of $221 \times 221$ pixels. Note that the alignment error between the SAR and the optical image is expected to be not larger than

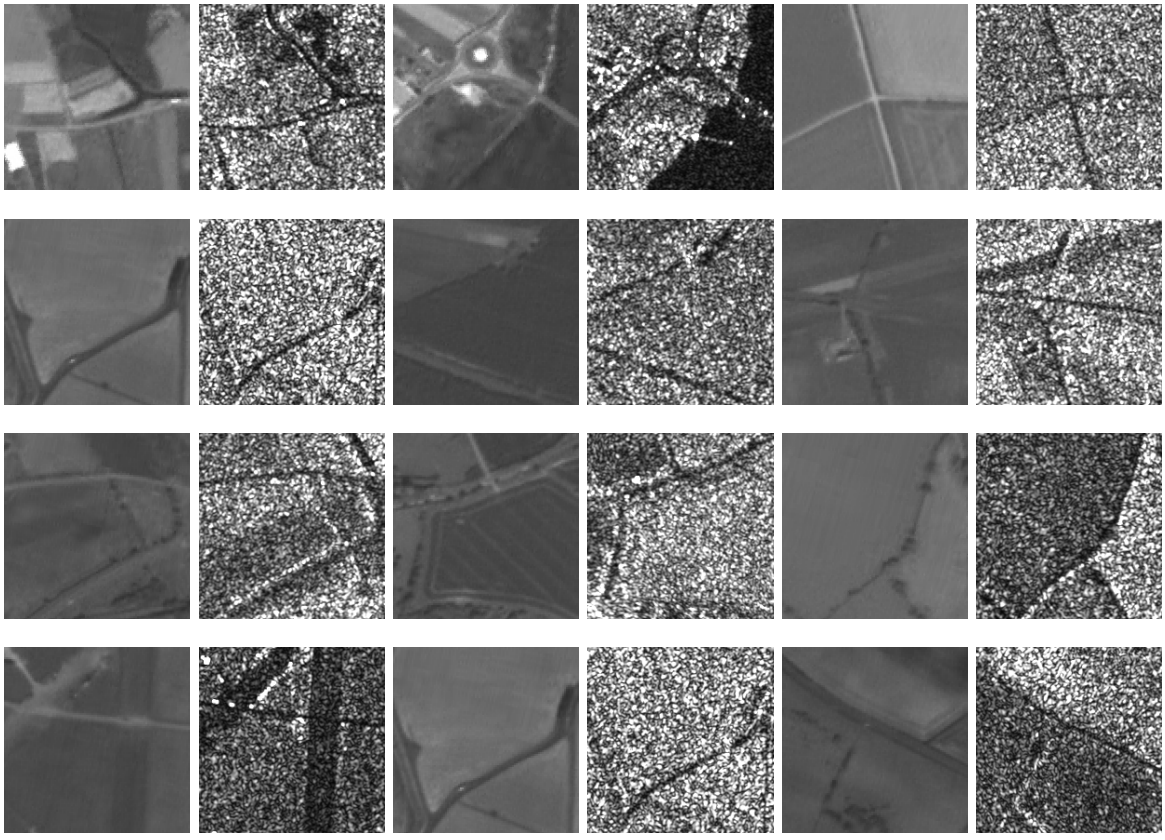


**Figure 5.3:** Samples of Optical and SAR patch pairs with a size of $201 \times 201$ pixels and a resolution of 2.5 m cropped from the pre-selected matching areas (in three columns).

| | Datasets | pixel spacing | speckle filter | # of patch pairs |
|---|---|---|---|---|
| **training** | 1 | 2.5 m | - | 69,900 |
| | 2 | 2.5 m | ✓ | 69,900 |
| | 3 | 2.5 m+3.75 m | - | 137,450 |
| | 4 | 2.5 m+3.75 m | ✓ | 137,450 |
| **val.** | 1 | 2.5 m | - | 5,000 |
| | 2 | 2.5 m | ✓ | 5,000 |
| **test** | 1 | 2.5 m | - | 14,401 |
| | 2 | 2.5 m | ✓ | 14,401 |

**Table 5.1:** Details of the different training, validation (val.) and test datasets.

32 m (the mean alignment error of the 6 test images is provided in Table 5.2). Therefore a $21 \times 21$ pixel search space with a pixel spacing of 2.5 m and a total number of search locations of 441 is assumed to be large enough in order to simulate a real world scenarios. Samples of resulting optical and SAR image pairs are shown in Figure 5.3 (here both with a size of $201 \times 201$ pixels).

Overall, we generate four different training datasets and two different validation and test sets, respectively. The two smaller training sets contain 69,900 patch pairs cropped from the optical and SAR image pairs with a pixel spacing of 2.5 m, where either the SAR or the filtered SAR images are used (single-resolution; with or without filtered SAR patches). The two larger training sets contain 137,450 patches pairs cropped from the image pairs with a pixel spacing of 2.5 m and 3.75 m, where either the SAR or the filtered SAR images are used (multi-resolution; with or without filtered SAR patches). Note that the patches with 3.75 m resolution are centered around the same location as the 2.5 m resolution patches but contain bigger areas and only exists in the dataset if the patches do not exceed the image boundaries. The larger training dataset is deployed to enlarge the number of training samples and to investigate the influence of different image resolutions on the quality of the network trainings. Since the matching should be as precise as possible, the validation and test sets contain only patches with a resolution of 2.5 m. The two validation sets contain each 5,000 patch pairs (with or without filtered SAR patches) and the two test sets contain each 14,401 patch pairs (with or without filtered SAR patches). Note that patches extracted from one image are either used for the training, the validation or the test dataset. An overview of all dataset is provided in Table 5.1.

| **Test images** | city | size [pixel] | # of patch pairs | mean error [pixel] |
|---|---|---|---|---|
| 1 | Bristol | $7,014 \times 10,083$ | 705 | 3.57 |
| 2 | Kalisz | $3,877 \times 7,653$ | 343 | 12.93 |
| 3 | Leeds | $8,003 \times 9,318$ | 1065 | 7.22 |
| 4 | Leeds | $5,623 \times 7,790$ | 356 | 8.14 |
| 5 | London | $8,569 \times 14,095$ | 6054 | 9.17 |
| 6 | Stara Zagora | $7,554 \times 13,865$ | 5878 | 7.95 |

**Table 5.2:** Details of the six test image pairs with a pixel spacing of 2.5 m.

The training and validation sets are utilized for the computation of the network parameters and for finding the optimal set of hyperparameters, respectively. The test set is not involved in these stages and is only utilized for the later performance analysis of the learned networks. Here, we will evaluate the performance over the whole test set but also on an image level. For this purpose, the details of the six images that form the test set are provided in Table 5.2.

### 5.1.3 Statistical Measures

In order to assess the performance of our methods and to be able to compare different setups and methods with each other, the quality of the generated sets of tie point has to be measured. Towards this goal, we evaluate the quality of our tie points regarding two aspects: the accuracy and precision. The accuracy of a set of tie points indicates how far away each predicted point is from its true location (correctness). The precision on the other hand, indicates how much the error of the predicted locations differs between tie points (consistency). In order to measure the accuracy and precision of a set of tie points in the subsequent sections, we apply the following measures.

**Accuracy Measure:** To measure how close the predicted tie point locations are to the true locations, a metric has to be defined. A metric, sometimes called a distance function, is a function that defines a distance between any two points of a given set [193]. Independent of the type of metric, the distances between two points lies in the range of $[0, \infty[$ and is equal to zero if and only if the two points are equivalent. Our applied accuracy measure $\mu$ is based on the $L_2$ (Euclidean) distance $d$ and simply measures the accuracy by computing the average over the distances between the tie points and the corresponding ground truth locations

$$\mu = \frac{1}{N}\sum_{i=1}^{N} d^{(i)} = \frac{1}{N}\sum_{i=1}^{N} \left\| \boldsymbol{p}_{\mathrm{gt}}^{(i)} - \boldsymbol{p}^{(i)} \right\|_2 . \tag{5.1}$$

Here, $d^{(i)}$ denotes the Euclidean distance between the true location $\boldsymbol{p}_{\mathrm{gt}}^{(i)}$ and the predicted location $\boldsymbol{p}^{(i)}$ of the $i$-th tie point and $N$ the total number of tie points.

**Precision Measure:** The corresponding precision of the tie points is measured through the standard deviation $\sigma$ of the distances $d^{(i)}$. The standard deviation is defined as the square root of the variance and, in our case, given as

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (d_i - \mu)^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \left( \left\| \boldsymbol{p}_{\mathrm{gt}}^{(i)} - \boldsymbol{p}^{(i)} \right\|_2 - \mu \right)^2}, \tag{5.2}$$

where $d^{(i)}$ denotes the Euclidean distance between the true location $\boldsymbol{p}_{\mathrm{gt}}^{(i)}$ and the predicted location $\boldsymbol{p}^{(i)}$ of the $i$-th tie point, $\mu$ the mean distance between the predicted and the true locations and $N$ the total number of tie points.

### 5.1.4 Baseline Description

To set the results of our developed tie point generation methods in relation to traditional and state-of-the-art methods we utilize five baseline methods. To enable a fair comparison, all baselines are applied on the same test data as our methods, the accuracy and precision of the corresponding set of tie points are computed as described in Subsection 5.1.3 and we implemented specific outlier removal strategies adapted to each baseline. Three of the applied baselines belong to the class of area-based approaches and two to the class of feature-based approaches and are described in the following:

**Area-based Baseline Methods:** From the large set of area-based matching approaches we choose NCC, MI and CAMRI [21] in order to compute tie points between the patch pairs from our test datasets. The computational details of a NCC- and MI-based matching and tie point generation procedure are provided in Subsection 3.1.1 and will not further be discussed. Since we are only interested in reliable and accurate tie points, we use the corresponding NCC- and MI-values of each tie point as a quality measure to detect and remove outliers from the set of tie points. More precisely, we remove all tie points with a NCC-value of less than 0.4 and a MI-value of less than 1.12. CAMRI on the other hand is a MI-based registration framework tailored to the problem of optical and SAR matching. CAMRI is a fully automatic framework with an integrated speckle filter and outlier removal procedure. Therefore, we apply CAMRI not on the test dataset containing despeckled images. For more details about CAMRI we refer to [17, 21].

**Feature-based Baselines Methods:** From the set of feature-based approaches we utilize a SIFT- [85] and BRISK- [86] based matching and tie point generation process. Computational details about both methods were described in Subsection 3.1.2 and will not further be discussed. To increase the quality and reliability of the detected tie points through the SIFT- and BRISK-based matching, we remove outliers through RANSAC [184] with an underlying affine model and with a distance threshold of 5 pixels. Note that we applied the SIFT and BRISK based matching in combination with RANSAC on the patches of the six test image scenes separately.

Note that besides CAMRI, none of the baseline method is particular developed for the problem of optical and SAR image matching. Nevertheless, we apply these methods due to the following reasons: 1) To investigate the performance of these methods as they often form the basis of single-sensor and traditional optical and SAR matching frameworks but are said to poorly perform without any adaptions to optical and SAR imagery. 2) To investigate our claim that the cGAN-based image-to-image translation scheme described in Subsection 4.2.2 improved the performance of these methods and enables the use of traditional matching approaches for an accurate and reliable tie point generation process between optical and SAR images. 3) To set the results of methods, specifically developed to the problem of optical and SAR image matching such as CAMRI and our two developed frameworks, in relation to such that are not particularly adapted to this problem.

## 5.2 Optical and SAR Image Registration Through Artificial Image Matching

The first approach for the registration of optical and SAR images is based on tie points generated through the usage of cGANs. In order to find the best model for this task, several cGAN configurations are trained and evaluated in the following. Therefore, an overview of the utilized configurations and the associated training parameters are provided in Subsection 5.2.1. An analysis and discussion about the quality and characteristics of the resulting artificial optical and SAR image patches in relation to the different cGAN setups are provided in Subsection 5.2.2. Subsequently, the ability of the different cGANs for an accurate and reliable tie point generation through a NCC-, MI-, SIFT- and BRISK-based image matching between the artificial image patches and the reference images is provided in Subsection 5.2.3. Additionally, the quality of the resulting tie points is compared to the state-of-the-art method CAMRI [21]. In Subsection 5.2.4, the potential of the generated tie points for the registration of optical and SAR images, and hence for the absolute geo-localization accuracy enhancement of optical images is discussed. Finally, the results of the proposed framework are summarized and its limitations and strengths discussed in Subsection 5.2.5.

### 5.2.1 Training Setups and Parameter Settings

In order to find the best cGAN for an accurate and reliable tie point generation between optical and SAR images, the three different cGAN setups (cGAN, cLSGAN, cWGAN) from Subsection 4.2.2 are trained with several configurations on the four training sets described in Subsection 5.1.2. Note that all training datasets contain larger SAR patches (for the later matching) but for the artificial image generation and hence for the training of the cGAN, equally large optical and SAR patches are required. Therefore, the SAR patches are cropped around the center to have the same size as the optical patches. Since the optical and SAR images used for the dataset generation are aligned, the cropped SAR patches show now the same image regions as the optical patches. The final training dataset consists therefore of optical and SAR image patches with a size of $201 \times 201$ pixels.

We investigated several configurations for the generation of artificial image patches. These include the generation of (despeckled) SAR-like and optical-like image patches at varying scales (pixel spacing: 2.5 m and 3.75 m), the training of the networks through different losses (cGAN, cLSGAN and cWGAN), the training with different batch sizes ($1, 4$ and $40$) and the training with despeckled SAR, SAR images and optical images as reference. Here, the batch size refers to the number of training instances used in one iteration of the training procedure. An overview of the different training configurations can be seen in Table 5.3.

| **Setup** | dataset | batch size | filter | direction |
|---|---|---|---|---|
| cGAN | 2.5 m / 2.5 m+3.75 m | 1/4/40 | yes / no | SAR→Opt / Opt→SAR |
| cLSGAN | 2.5 m / 2.5 m+3.75 m | 1/4/40 | yes / no | SAR→Opt / Opt→SAR |
| cWGAN | 2.5 m / 2.5 m+3.75 m | 1/4/40 | yes / no | SAR→Opt / Opt→SAR |

**Table 5.3:** Overview of the different cGAN training configurations.

As mentioned is Subsection 4.2.2, each network is trained using stochastic gradient descent with the ADAM optimizer [63] and an initial learning rate of 0.01 for the cGAN and cLSGAN setups and with the RMSProp optimizer [62] and an initial learning rate of 0.0002 for the cWGAN setup. For all setups the generator $G$ and discriminator $D$ networks are trained at the same time by alternating the training of $D$ and $G$ (one gradient descent step of $D$ is followed by one gradient descent step of $G$ in the cGAN and cLSGAN setups and five gradient descent steps of $D$ are followed by one gradient descent step of $G$ in the cWGAN setup). For each setup the corresponding cGAN is trained over 200 epochs (one epoch refers to one whole cycle through the entire training set) on a single NVIDIA GeForce GTX Titan X GPU. The training time varies from several days to several weeks depending on the batch size, the size of the training dataset and the chosen cGAN setup.

### 5.2.2 Artificial Image Generation

In this subsection, we provide a quantitative evaluation of the quality of the artificial image generation process. A qualitative analysis of the artificial patches with regard to their usability for the generation of accurate and reliable tie points through traditional matching approaches is provided in Subsection 5.2.3. All artificially generated images patches shown in this and the following subsections are obtained from the set of test image patches, and hence have never been shown to the different generator networks during the training process.

**SAR image generation:** We first investigate the generation of artificial SAR and despeckled SAR image patches from optical images. Figure 5.4 shows examples of (despeckled) SAR patches with a pixel spacing of 2.5 m generated by two different generators. The first generator, utilized for the SAR image generation, was trained with the cWGAN loss, a batch size of one and on the smaller training dataset. The second generator, utilized for the despeckled SAR image generation, was trained with the cGAN loss, a batch size of 40 and on the smaller dataset, where the filtered SAR images were used as reference. These two configurations led, from a visual point of view, to the most realistic-looking SAR and despeckled SAR image patches. The illustrated examples show that the geometric structures of streets extracted from optical images are preserved in the generated patches, while the radiometric properties are adapted to SAR or despeckled SAR images. Through the training process the generators learned that in contrast to optical images, streets normally appear with a lower intensity in SAR images. Furthermore, the generators try to represent the characteristics of speckle or the resulting pattern from the speckle filter. The development of the learning process of both generators trained with the two different configurations over the training time are exemplified in Figure 5.5. The longer we trained the networks, the better become the generators in generating realistic looking (despeckled) SAR images from optical image patches.

Despite the good visual appearance of the imitated texture of the speckle and the speckle filter, it is important to note that it is randomly generated and independent from the real image objects or their properties. Furthermore, the (despeckled) SAR image generation is not free of errors and in some situations the generators produces unsatisfying results (see Figure 5.6). A possible reason for the difficulties of the networks in generation image patches

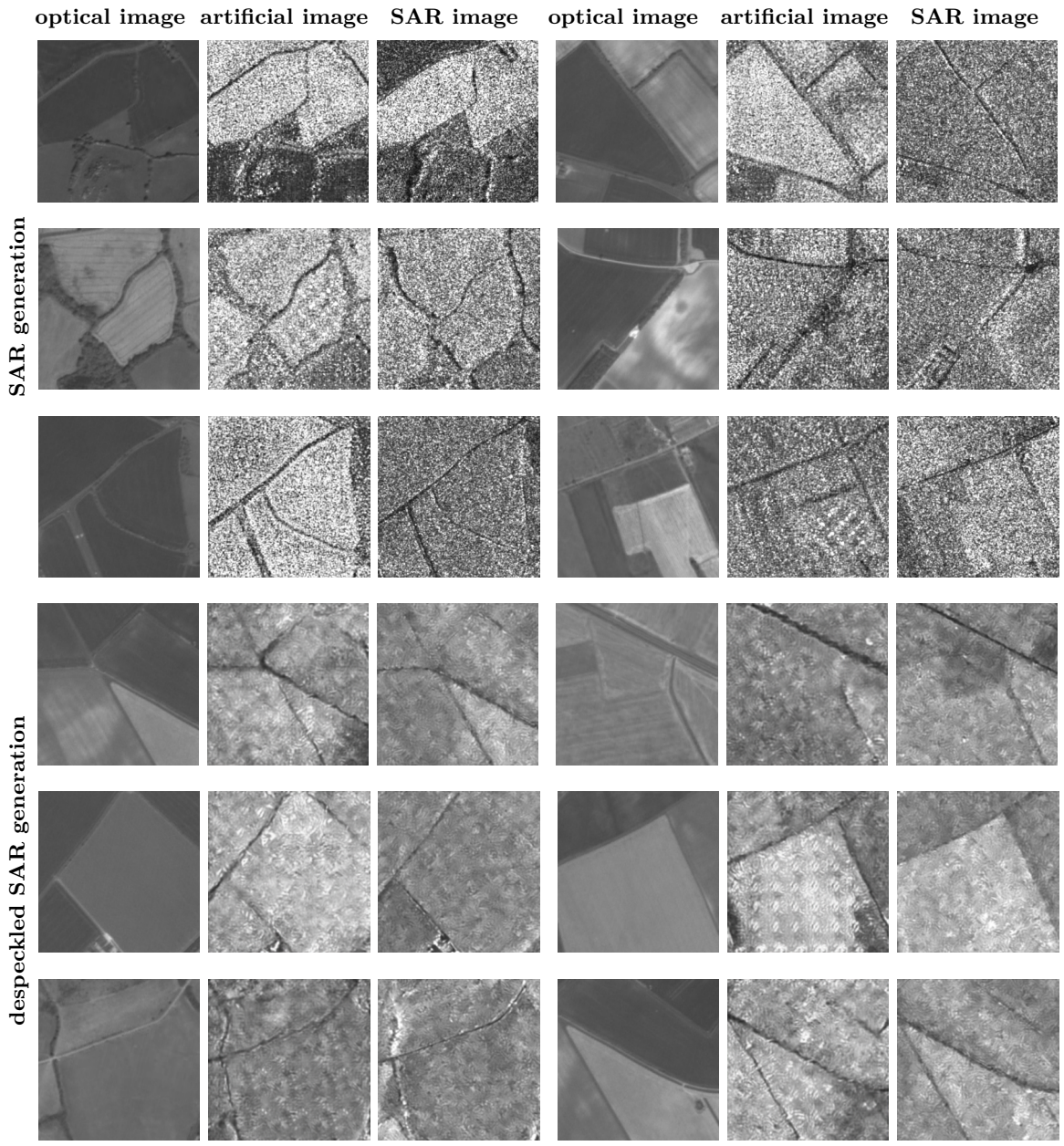



**Figure 5.4:** Side by side comparison between optical, artificial (despeckled) SAR and real (despeckled) SAR image patches with a pixel spacing of 2.5 m in two columns. SAR generation: The generator used to generate the artificial SAR images was trained with the cWGAN loss, a batch size of one and on the smaller training dataset. Despeckled SAR generation: The generator used to generate the artificial despeckled SAR images was trained with the cGAN loss, a batch size of 40 and on the smaller dataset with filtered SAR images as reference.

for the runway example (first row and column) is the small amount of runway patches in the training dataset. This problem could be solved with a larger runway training dataset or a separated training of street and runway patches. In the other three cases it can be seen that some features are present in the optical images but are missing in the generated images. Since optical images exhibit a higher level of detail than the SAR images the network learns during the training to ignore some of the features/objects for the generation of (despeckled) SAR images. However, for our later application it is essential that features that are valuable

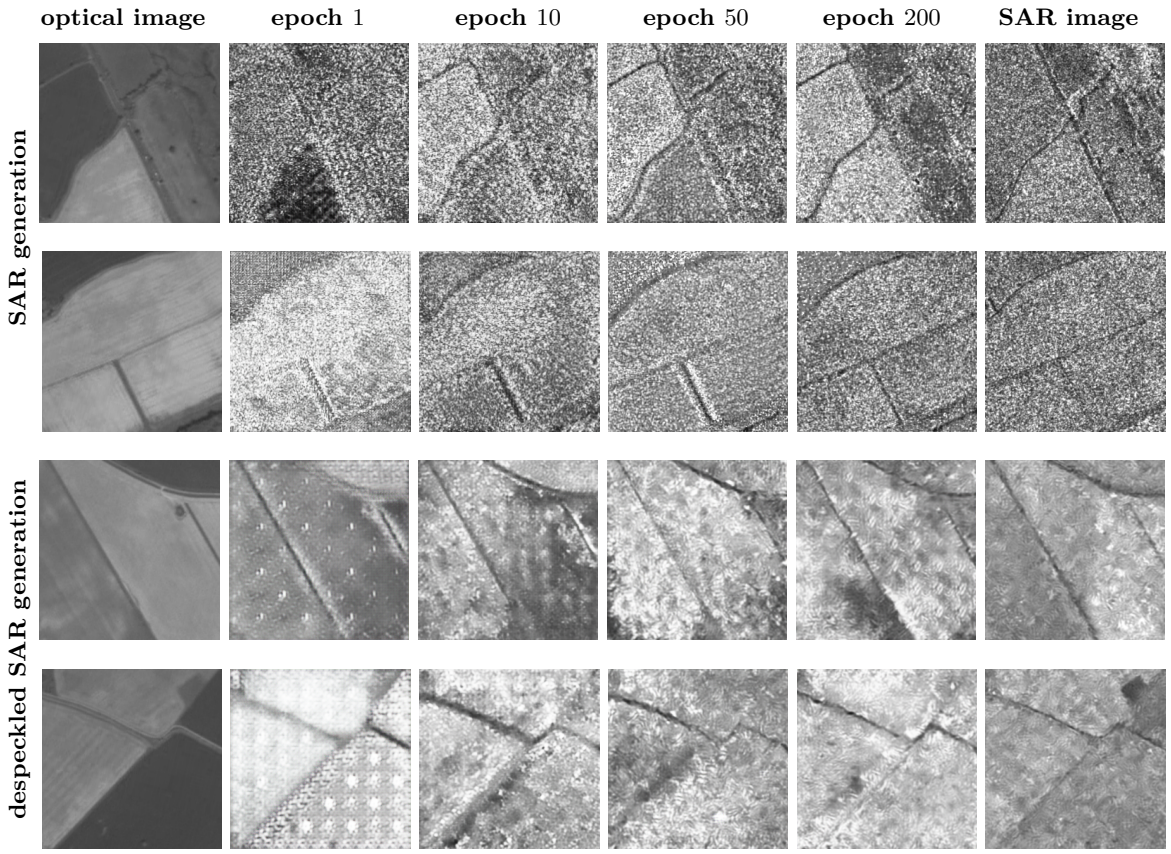



**Figure 5.5:** Development of the generator over training. From left to right: optical input patches, the artificially generated patches at epoch 1, 10, 50, 200 and the (despeckled) SAR target patches. The first two rows show the development of a generator trained for the generation of SAR patches by using the cWGAN loss, a batch size of 1 and the smaller training dataset. The third and fourth rows show the development of a generator trained for the generation of despeckled SAR patches by using the cLSGAN loss, a batch size of 4 and the larger training dataset with the filtered SAR images.

for the image matching, e.g. street and street crossings, are still present in the generated images. A possible solution for this problem could be to adjust the training procedure by adding the actual problem, the matching between the generated images and the reference image, into to training objective. Thereby the generator would learn, which features are crucial for the matching process, and hence which features should be kept in the artificial image generation process.

**Optical image generation:** We further investigated the reversed process and therefore trained networks in order to generate artificial optical images out of SAR images. Examples of such artificially produced optical images are shown in Figure 5.7. The corresponding generator was trained using the cGAN loss, a patch size of 4 and over the large training dataset. This configuration led (from a visual point of view) to the best and most realistic looking artificial optical images. Similar to the (despeckled) SAR image generation, the generator learned to keep the geometric structures of objects such as streets from the SAR images, while adapting the radiometric properties to optical images. Figure 5.8 shows two samples that illustrate the learning process of the generator over the training time.

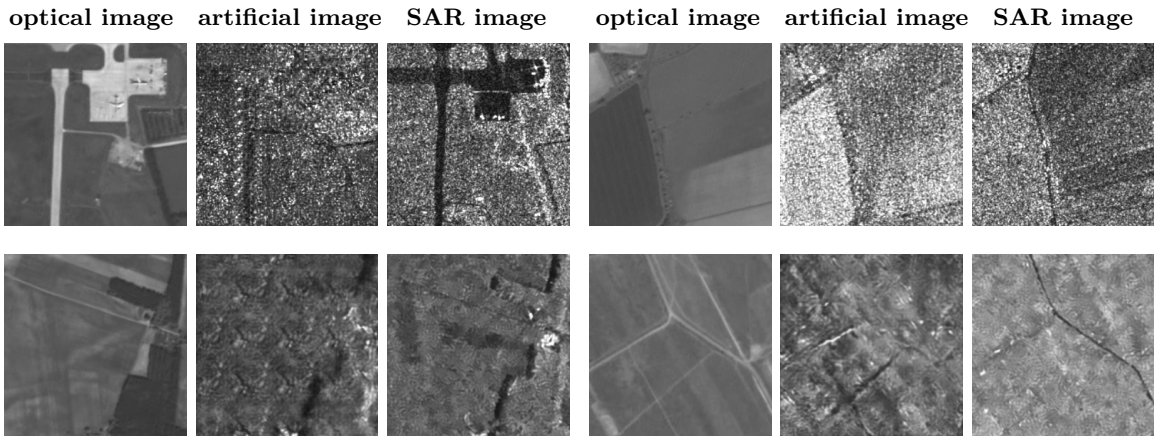



**Figure 5.6:** Comparison of failure cases of artificially generated SAR images with optical and real (despeckled) SAR image patches. The first row shows low quality artificial SAR images, and the second row low quality artificial despeckled SAR images.

Like for the SAR image generation, the learned generator model is not perfect and provides for some input images optical images of low quality (see Figure 5.9). Due to the lower level of detail in SAR images and the speckle it is more difficult to extract and recreate features from SAR than from optical images. Most of the details are missing in the SAR images and a realistic recreation is therefore almost impossible for the networks. As a consequence, the network tries to come as close as possible to real optical images by adding additional structure to the artificially generated images. These created structures might look realistic but is not derived from the input images.

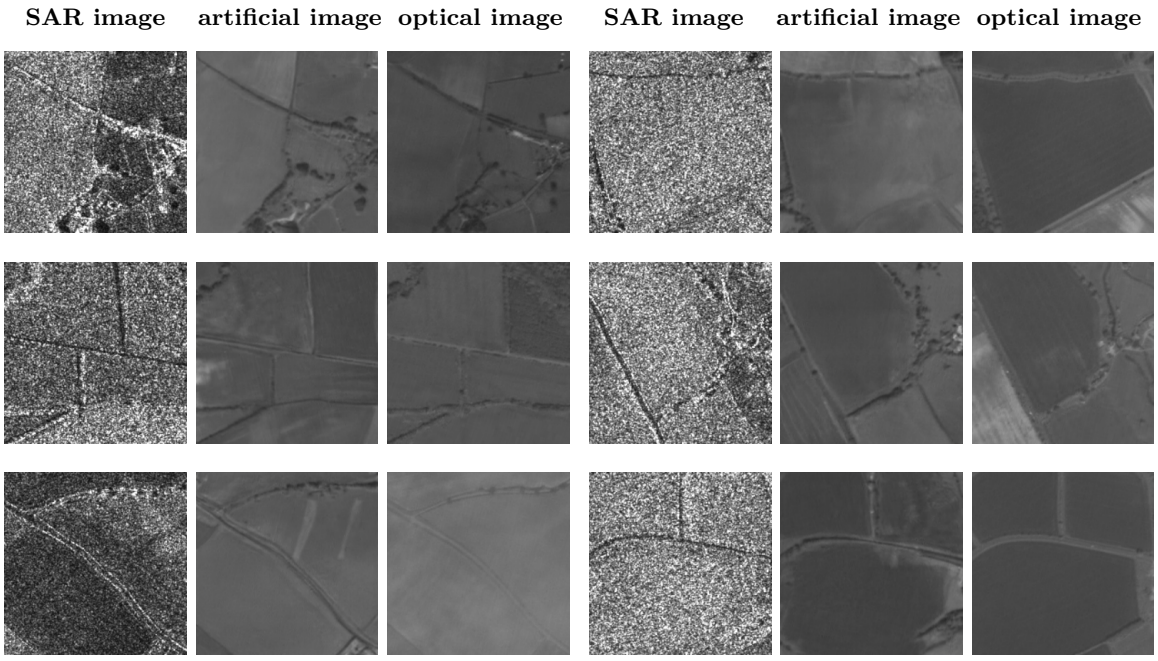



**Figure 5.7:** Comparison between SAR, artificial optical and real optical image patches. The generator used to generate the artificial optical images was trained with the cGAN loss, a batch size of 4 and on the larger training dataset.

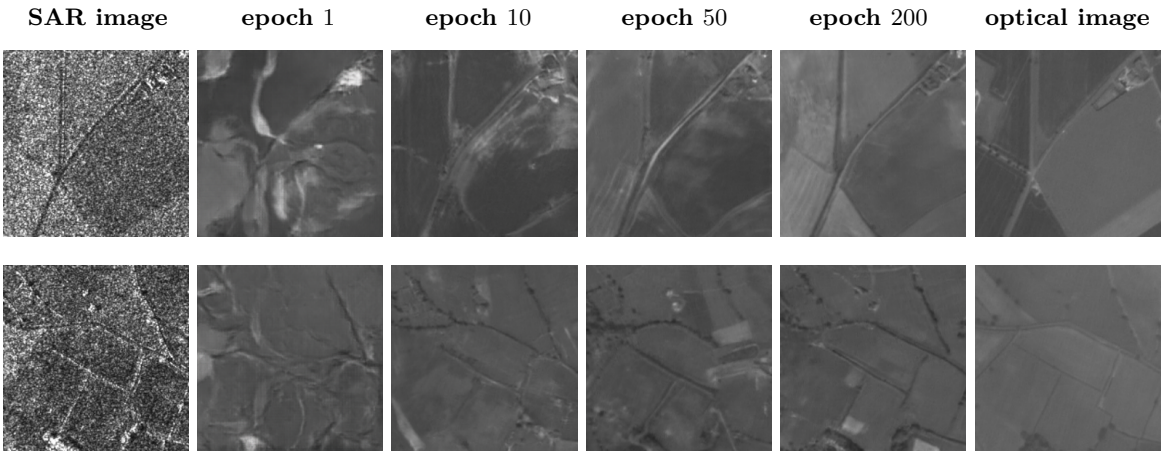



**Figure 5.8:** Development of the generator over training. From left to right: the SAR input patches, the artificially generated patches at epoch 1, 10, 50, 200 and the optical target patch. The generator used to generate the artificial optical images was trained with the cGAN loss, a batch size of 4 and on the larger training dataset.

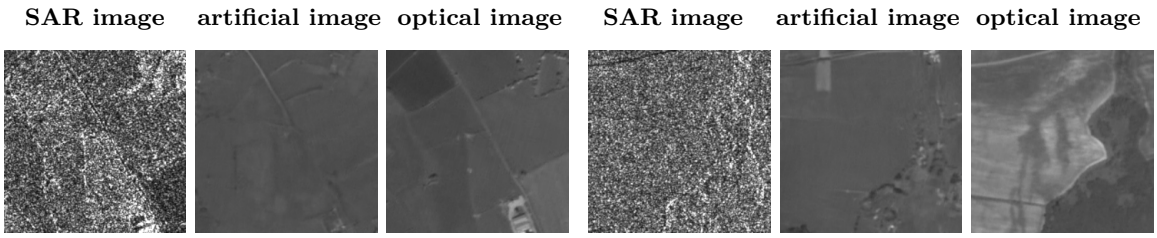



**Figure 5.9:** Comparison of failure cases of artificially generated optical images with real optical and SAR image patches.

In general, the concept of cGANs (introduced in Subsection 4.2) enables the translation from SAR to optical images and vice versa. In both directions, realistic looking optical and SAR images respectively can be generated. In practice, the cGAN and cWGAN loss led to more realistic looking images compared to the cLSGAN loss (see Figure 5.10 for a comparison). However, for our pursued application it is not important that the obtained images look real. The important aspect is that the artificially generated patches improve the quality of the matching between optical and SAR images. Therefore, a detailed investigation and discussion about the effects of the different configuration, e.g. the three losses and the kind of input and reference data (SAR, despeckled or optical), on the quality of the generated tie points follows in the next subsection.

### 5.2.3 Tie Point Generation

In this subsection, we investigate the influence of the artificially generated image patches on the matching quality of a NCC-, MI-, SIFT- and BRISK-based matching between optical and SAR images. More precisely, the matching quality will be assessed with respect to an accurate and reliable tie point generation between the artificial and real image patches and compared to the matching results between the real optical and SAR patches. Towards this goal, several aspects of the image generation process will be examined such as the influence of a speckle filter, of the matching directions (optical to SAR or vice versa) and of the training loss on the quality of the resulting tie points. In the following, the matching accuracy is measured, next to method described in Subsection 5.1.3, as the percentage of tie points having a $L_2$ distance with less than 3 pixels to the ground truth location.

**Influence of the Speckle Filter:** The application of a speckle filter is an important pre-processing step for many matching methods applied on SAR images. We exploited therefore two application cases of the speckle filter and provide the corresponding results in Table 5.4. First, we investigated the influence of the filter on the NCC-, MI-, SIFT- and BRISK-based matching between optical and despeckled SAR images (without the use of cGANs). Here, the use of the speckle filter led only in the case of a BRISK-based matching to an improvement of the matching quality. Second, we investigated the generation of SAR-like despeckled patches from optical patches via cGANs and their influence on the NCC-, SIFT- and BRISK-based matching. Utilizing the artificially generated despeckled SAR patches for an NCC- and MI-based matching with the real SAR image patches led to a deterioration of the accuracy and precision of the resulting tie points, whereas the application of a SIFT- and BRISK-based matching lead to an improvement of the results. Overall, the second application case of the speckle filter led to better results and provides an first indication of the usefulness of the artificially generated patches for the problem of optical and SAR image matching. On the other hand, Table 5.5 reveals that the best overall results are achieved without the usage of a speckle filter (see the last block of four). More precisely, the best overall results are achieved by using artificial SAR-like patches generated from a generator who was trained through the use of the cLSGAN loss. For more details about the best training configuration see the later paragraph "Influence of the Loss Function". A possible reason for the lower performance of despeckled artificial image patches could be a modification of the shape or boundaries of the objects through the speckle filter. Since despeckled SAR images are utilized during the training as reference, the generator learns to simulate these effects while creating the artificial images. Even if the changes on the extracted objects are small, they can interfere an accurate and precise matching.

**Influence of the Matching Direction:** We further considered to reverse the whole process and utilize artificially generated optical-like image patches for the matching with the real optical image patches. Despite the reasonable visual appearance (see Figure 5.7) the artificial optical images could only slightly improve the matching quality of the four applied matching approaches (see Table 5.4). We attribute this to the fact that optical images reveal a higher level of detail as SAR images and that the extraction and recreation of features from SAR images is more difficult than from optical images. It is therefore more difficult to preserve all

| Methods | matching accuracy $< 3$ pixels | matching accuracy $\mu$ [pixel] | matching precision $\sigma$ [pixel] |
|---|---|---|---|
| NCC | 35.55% | 5.50 | 4.76 |
| MI | 64.47% | 3.09 | 4.69 |
| SIFT[85] | 31.10% | 5.61 | 1.64 |
| BRISK[86] | 39.58% | 3.61 | 1.70 |
| $NCC_{fil}$ | 19.75% | 6.91 | 4.79 |
| $MI_{fil}$ | 29.40% | 4.89 | 3.60 |
| $SIFT_{fil}$ | 26.37% | 6.09 | 1.84 |
| $BRISK_{fil}$ | 52.21% | 2.98 | 1.37 |
| $NCC_{cLSGAN,fil}$ | 37.59% | 5.93 | 5.11 |
| $MI_{cLSGAN,fil}$ | 33.12% | 5.44 | 4.34 |
| $SIFT_{cLSGAN,fil}$ | 62.80% | 2.62 | 1.23 |
| $BRISK_{cLSGAN,fil}$ | 68.93% | 2.38 | 1.12 |
| $NCC_{cGAN,opt}$ | 20.05% | 8.05 | 4.55 |
| $MI_{cGAN,opt}$ | 39.35% | 6.90 | 3.96 |
| $SIFT_{cGAN,opt}$ | 46.39% | 4.44 | 1.45 |
| $BRISK_{cGAN,opt}$ | 60.22% | 2.62 | 1.08 |
| $NCC_{cLSGAN}$ | 75.48% | 2.94 | 5.79 |
| $MI_{cLSGAN}$ | 65.60% | 3.19 | 4.67 |
| $SIFT_{cLSGAN}$ | 68.85% | 2.40 | 1.05 |
| $BRISK_{cLSGAN}$ | 75.21% | 2.22 | 1.10 |
| CAMRI[23] | 57.06% | 2.80 | 2.86 |

**Table 5.4:** Influence of the artificially generated templates on the matching accuracy and precision of a NCC-, MI-, SIFT-[85] and BRISK-[86] based image matching, and comparison with baseline method (CAMRI[23]). The matching accuracy is measured as the percentage of tie points having $L_2$ distance to the ground truth location smaller than 3 pixels, and as the average over the $L_2$ distances between the predicted tie points and the ground truth locations $\mu$. The matching precision is represented by the standard deviation $\sigma$.

image features that are important for a reliable and accurate matching. Nevertheless, this direction provides several possibilities for future developments, which will be discussed in detail in Chapter 6.

**Influence of the Loss Function:** To identify the best training configuration for our application we investigated the influence of the three different loss functions introduced in Subsection 4.2.2 and their dependency on the batch size and the dataset size for the case of SAR image generation. An overview of the tested configuration is provided in Table 5.3. We achieved the best matching results, with respect to the utilized cGAN, cLSGAN and cWGAN training losses, with the artificial patches whose corresponding generators where trained over the larger dataset and with a batch size of 4, 4 and 1, respectively. An overview of the best results with respect to the three training losses is shown in Table 5.5. In contrast to Subsection 5.2.2, where the best results were obtained through a training based on the cGAN and cWGAN loss, here the best results were achieved through a training based on the cLSGAN loss. This circumstance can be explained by regarding the corresponding artificially generated image patches (shown in Figure 5.10) and by taking the different requirements of

| Methods | matching accuracy $< 3$ pixels | matching accuracy $\mu$ [pixel] | matching precision $\sigma$ [pixel] |
|---|---|---|---|
| $NCC_{cGAN}$ | 63.70% | 3.59 | 5.56 |
| $MI_{cGAN}$ | 65.30% | 3.14 | 4.72 |
| $SIFT_{cGAN}$ | 65.87% | 2.52 | 1.15 |
| $BRISK_{cGAN}$ | 74.82% | 2.24 | 1.08 |
| $NCC_{cLSGAN}$ | 75.48% | 2.94 | 5.79 |
| $MI_{cLSGAN}$ | 65.60% | 3.19 | 4.67 |
| $SIFT_{cLSGAN}$ | 68.85% | 2.40 | 1.05 |
| $BRISK_{cLSGAN}$ | 75.21% | 2.22 | 1.10 |
| $NCC_{cWGAN}$ | 49.47% | 3.96 | 4.61 |
| $MI_{cWGAN}$ | 23.31% | 6.54 | 3.72 |
| $SIFT_{cWGAN}$ | 56.51% | 2.89 | 1.31 |
| $BRISK_{cWGAN}$ | 61.71% | 2.61 | 1.18 |

**Table 5.5:** Influence of loss function on the matching accuracy and precision of a NCC-, MI-, SIFT-[85] and BRISK-[86] based image matching between artificially generated SAR-like and SAR image patches. The matching accuracy is measured as the percentage of tie points having $L_2$ distance to the ground truth location smaller than 3 pixels, and as the average over the $L_2$ distances between the predicted tie points and the ground truth locations $\mu$. The matching precision is represented by the standard deviation $\sigma$.

both tasks into account (realistic looking image generation vs. accurate image matching). The cGAN and cWGAN loss enables the creation of speckle and resulting patterns of the speckle filter, whereas the cLSGAN loss causes a blurry appearance of objects such as fields without any speckle patterns but with sharper boarders between objects. This effect of the cLSGAN could be a possible explanation for the better matching performance since the applied matching methods (NCC, MI, SIFT and BRISK) normally suffer from speckle in the image patches. Moreover, since the "real" speckle of the SAR images and the "real" pattern from the speckle filter cannot be derived from the optical patches, it cannot be learned by the generator. As a consequence, the generator network will produce patches, which contain random speckle or simulated patterns from the speckle filter that look real enough to "fool" the discriminator network but negatively influence the matching quality. Overall, the occurrence of artificially generated patterns in the generated patches makes the matching more difficult, and hence less accurate.

**Influence of the Matching Method:** We realized the matching between the test image pairs through four traditional matching approaches. Without the use of the artificially generated image patches none of the methods could provide accurate and precise tie points between the optical and SAR image patches (see Table 5.4). Through the use of generated SAR-like image patches the matching performance of NCC, SIFT and BRSIK can be improved significantly. Only in the case of a MI-based matching the results could not be improve and the application of the artificial patches led to a slight deterioration. Overall, the feature-based approaches SIFT and BRISK performed better and led next to an accurate also to a very precise set of tie points, whereas the points obtained through an intensity-based matching through NCC and MI show a lack of precision.

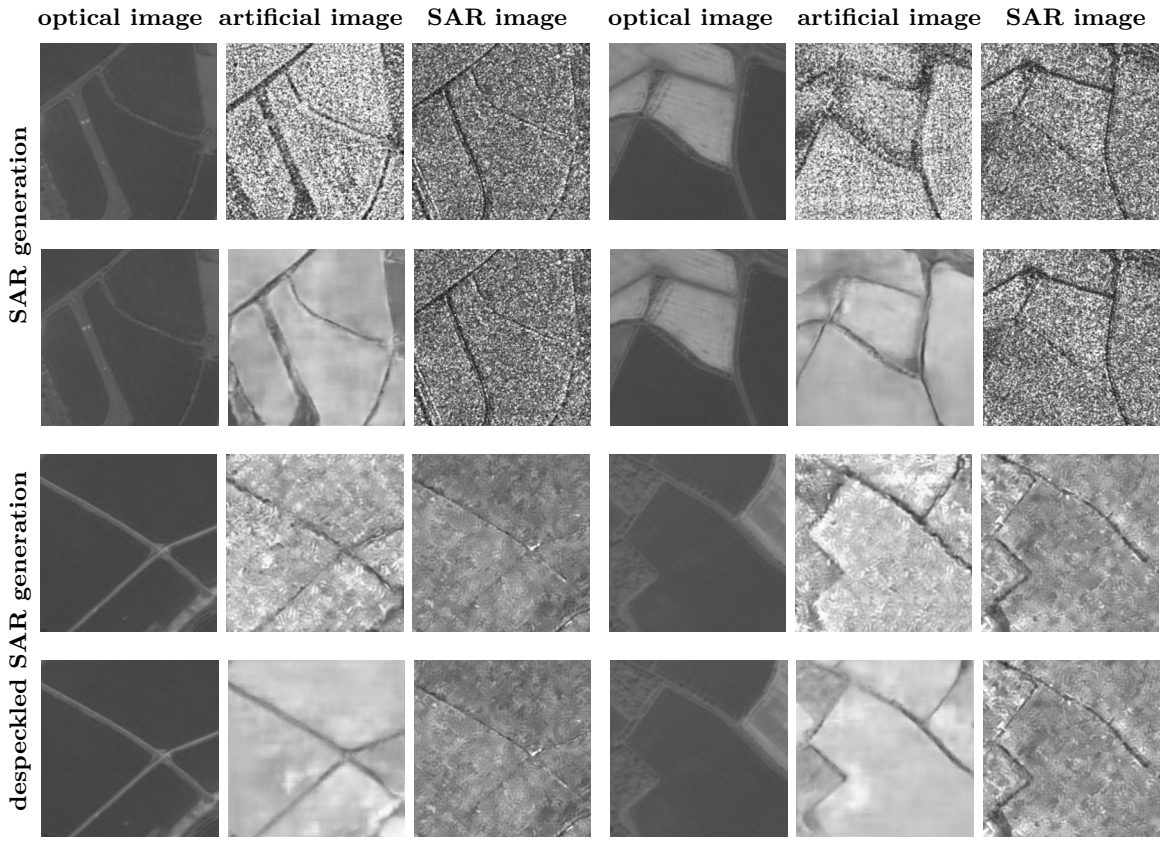



**Figure 5.10:** Comparison between the most realistic looking artificial images patches and the best artificial patches for the task of image matching. The first two rows show a comparison between the most realistic looking artificial SAR patches (training: cWGAN loss, a batch size of 1 and on the smaller training dataset) and the artificial SAR patches leading to the best matching results (training: cLSGAN loss, a batch size of 4 and on the larger training dataset). The last two rows show a comparison between the most realistic looking artificial despeckled SAR patches (training: cGAN loss, a batch size of 40 and on the smaller training dataset) and the artificially generated despeckled SAR patches leading to the best matching results (training: cLSGAN loss, a batch size of 4 and on the larger training dataset). All depict patches have a pixel spacing of 2.5 m.

**Comparison to Baseline Method:** For a better assessment of the quality of the resulting tie points a comparison with the state-of-the-art approaches CAMRI [23] is carried out. By performing a SIFT- and BRISK-based matching between the artificially generated SAR and real SAR image patches, we can achieve better results than CAMRI [23] (applied on the optical and SAR image patches) with respect to the matching accuracy and precision of the obtained tie points (see Table 5.4). This fact underlies the high potential provided by the cGAN-based image matching as an accurate and reliable tie point generation framework. Whether the obtained tie points are suitable for an accurate optical and SAR image registration, and hence for an absolute geo-localization improvement of optical images is further investigated in Subsection 5.2.4. Additionally, a comparison between the cGAN and the Siamese neural network-based tie point generation approaches is performed in Subsection 5.3.2.

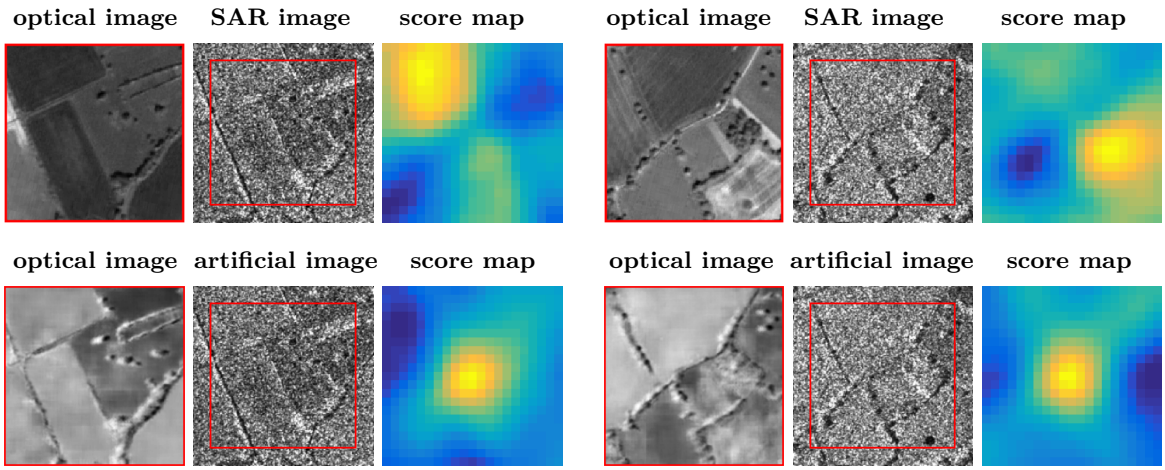



**Figure 5.11:** Two comparisons (top/bottom) of the score maps between the NCC-based matching of the optical image and the SAR image (left), and between the artificially generated images and the despeckled SAR image (right).

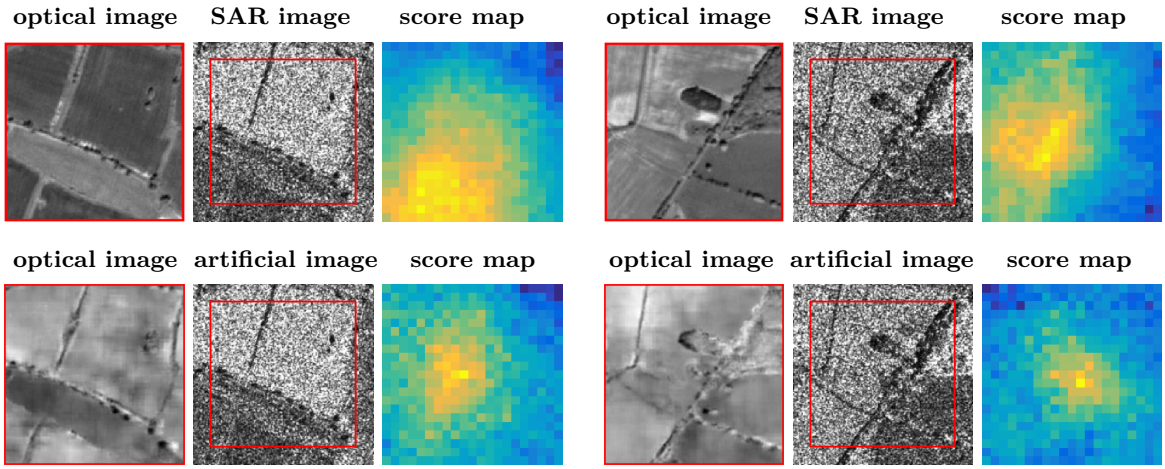



**Figure 5.12:** Two comparisons (top/bottom) of the score maps between the MI-based matching of the optical image and the SAR image (left), and between the artificially generated images and the despeckled SAR image (right).

**Qualitative Results of NCC and MI:** At last, Figure 5.11 and Figure 5.12 show a qualitative comparison of the NCC- and MI-based matching between optical and SAR patches and, between artificially generated SAR and SAR patches, respectively. Note that CAMRI [23] does not provide a score map as output. Here, the optical and artificial SAR image patches have a size of $201 \times 201$ pixels and the SAR patches a size of $221 \times 221$ pixels. The search space is $\Delta_x = \Delta_y = 20$ pixels in each direction around the center position. All artificial patches are generated with the same generator network, which have proven to provide the best patches for an accurate matching (for details see the paragraph "Influence of the Loss Function"). For all examples, the correct matching position is in the center of the SAR patches (red framed area). The brighter the color of the score map, the higher is the NCC- or MI-value at the corresponding location. The examples emphasize that the generated SAR-like image patches can improve the matching between optical and SAR images through a NCC- and MI-based matching.

### 5.2.4 Geo-localization Accuracy Enhancement Though Tie Points

In order to archive precisely registered optical and SAR images it is important to have a set of reliable and accurate tie points. The set of tie points does not have to be too large but the points have to exhibit a high accuracy and precision and, to a certain degree, have to be equally spread over the whole image scene. So far, we only evaluated the first aspect and investigated the accuracy and precision of the resulting tie points in Subsection 5.2.3. We showed that the usage of artificial image patches significantly improved the matching accuracy and precision of a NCC-, SIFT- and BRISK-based matching (see Table 5.4). Additionally, we showed that in the case of a SIFT- and BRISK-based image matching the accuracy and precision of the obtained tie points are better compared to the state-of-the-art approach CAMRI [23]. Nevertheless, several important aspects have not yet been examined: How many tie points does the proposed method provide per optical and SAR test image pair? Do we obtain enough tie points per image scene? How accurate and precise are these tie points? Are the tie points spread over the whole image or accumulated at one location?

In Table 5.6 an overview of the obtained numbers of tie points per test image scene and the corresponding accuracies and precisions (with and without the use of artificial patches) for each of the four matching approaches is given. Note that here and in the rest of this subsection the utilized artificial patches are generated with the same generator network. This generator was trained utilizing the cLSGAN loss, a batch size of 4, on the larger training set and has proven to provide the most suitable artificial patches for the task of optical and SAR image matching (see Subsection 5.2.3). The achieved results indicate once again that a NCC- and MI-based matching is not capable to generate a usable set of tie points whether the artificial images are used or not. A SIFT- and BRISK-based matching in combination with the artificial images on the other hand, provides a large set of accurate and especially precise tie points for every image scene. In all cases, the usage of the artificial image patches increased the number of computed tie point while increasing their accuracy and precision. Additionally, the performance of the SIFT- and BRISK-based matching is nearly equal between the different image scenes even though they are spread across Europe. This is particularly remarkable for test image six, which is located far away from any training images (see images Figure 5.1). In the following only the BRISK- and SIFT-based matching is further considered due to the better performance.

In order to ensure that the obtained tie points are not all located at the same location within the image scenes we utilize a further post-processing step. More precisely, we set an empirical distance threshold to 50 m to ensure that the final set of tie points contain only points with at least a spatial Euclidean distance of 50 m to each other. The resulting numbers of tie points and the corresponding accuracies and precisions per test images scene before and after applying the distance threshold are provided in Table 5.7. In most of the cases the distance threshold led to a minor deterioration of the accuracy and precision of the tie points and to a reduction in the number of tie points. Comparing the results of the SIFT- and BRISK-based matching underlies the better performance of BRISK in all aspects examined. Overall, a sufficient number of tie points was achieved for all image, although their accuracy and precision still leave room for improvement. An example of the tie point distribution within an images scene is exemplified for the first test image in Figure 5.13.

| | Test image | # of patch pairs | # tie points without | # tie points with cGAN | accuracy $\mu$ [pixel] without | accuracy $\mu$ [pixel] with cGAN | precision $\sigma$ [pixel] without | precision $\sigma$ [pixel] with cGAN |
|---|---|---|---|---|---|---|---|---|
| **NCC** | 1 | 705 | 65 | 64 | 3.86 | 1.67 | 5.62 | 4.63 |
| | 2 | 343 | 0 | 2 | - | 2.53 | - | 3.47 |
| | 3 | 1065 | 60 | 22 | 8.44 | 3.42 | 3.42 | 6.98 |
| | 4 | 356 | 0 | 18 | - | 2.32 | - | 6.83 |
| | 5 | 6054 | 140 | 33 | 5.64 | 3.62 | 4.64 | 4.99 |
| | 6 | 5977 | 81 | 16 | 4.45 | 6.71 | 5.21 | 4.50 |
| **MI** | 1 | 705 | 24 | 97 | 3.10 | 2.21 | 4.18 | 4.67 |
| | 2 | 343 | 6 | 1 | 1.72 | 2.24 | 3.78 | 3.51 |
| | 3 | 1065 | 58 | 35 | 6.21 | 4.96 | 3.71 | 3.96 |
| | 4 | 356 | 18 | 230 | 2.29 | 3.05 | 5.79 | 5.28 |
| | 5 | 6054 | 85 | 54 | 3.41 | 3.29 | 4.14 | 4.20 |
| | 6 | 5977 | 97 | 13 | 2.90 | 5.71 | 5.32 | 3.89 |
| **SIFT** | 1 | 705 | 84 | 235 | 3.27 | 2.04 | 1.85 | 0.97 |
| | 2 | 343 | 7 | 120 | 25.53 | 2.52 | 0.30 | 1.23 |
| | 3 | 1065 | 10 | 70 | 19.28 | 2.61 | 1.83 | 0.99 |
| | 4 | 356 | 9 | 27 | 21.19 | 2.48 | 0.43 | 1.21 |
| | 5 | 6054 | 55 | 363 | 3.11 | 2.49 | 1.03 | 1.08 |
| | 6 | 5977 | 110 | 286 | 4.86 | 2.49 | 1.95 | 1.07 |
| **BRISK** | 1 | 705 | 460 | 697 | 3.49 | 2.14 | 1.43 | 1.04 |
| | 2 | 343 | 53 | 393 | 3.02 | 2.12 | 1.35 | 1.12 |
| | 3 | 1065 | 592 | 520 | 3.99 | 2.04 | 1.78 | 1.06 |
| | 4 | 356 | 101 | 164 | 3.25 | 2.11 | 1.72 | 1.03 |
| | 5 | 6054 | 1409 | 2834 | 3.83 | 2.26 | 1.88 | 1.13 |
| | 6 | 5977 | 687 | 1052 | 3.00 | 2.30 | 1.48 | 1.07 |

**Table 5.6:** Influence of the artificially generated patches on the numbers of tie points and their accuracies and precisions obtained through a NCC-, MI-, SIFT- and BRISK-based matching on the six test image pairs. Here, the term "without" indicates the tie point generation through the matching between real optical and SAR image patches, while "with cGAN" through the matching between artificial SAR-like and real SAR patches from the set of test image pairs.

Here, the distribution of the tie points across the whole image is clearly visible. In the following the final set of tie point will be used to register the corresponding optical and SAR image, and hence to increase the absolute geo-localization accuracy of the optical images.

In a last step, the final sets of tie points (provided by the BRSIK-based matching) are utilized to improve the parameters of the corresponding sensor models and, hence, to improve the geo-localization accuracy of the optical images as described in Section 4.4. The unknown parameters of each sensor model are estimated from the corresponding set tie points by iterative least squares adjustment. During this process, a blunder detection is used to further remove outliers from the set of tie points (details provided in Section 4.4). For our six test images 5-10% percent of the tie points were removed during this step. At the end, we used the improved sensor model to generate new orthorectified optical image, which show an improved absolute geo-localization accuracy.

| | Image | # of patch pairs | # tie points before | # tie points after | accuracy $\mu$ [pixel] before | accuracy $\mu$ [pixel] after | precision $\sigma$ [pixel] before | precision $\sigma$ [pixel] after |
|---|---|---|---|---|---|---|---|---|
| **SIFT**$_{\text{cLSGAN}}$ | 1 | 705 | 235 | 29 | 2.04 | 2.22 | 0.97 | 1.10 |
| | 2 | 343 | 120 | 20 | 2.52 | 2.87 | 1.23 | 1.11 |
| | 3 | 1065 | 70 | 18 | 2.61 | 2.76 | 0.99 | 0.93 |
| | 4 | 356 | 27 | 8 | 2.48 | 2.87 | 1.21 | 0.94 |
| | 5 | 6054 | 363 | 68 | 2.49 | 2.63 | 1.08 | 1.04 |
| | 6 | 5977 | 286 | 65 | 2.49 | 2.44 | 1.07 | 1.02 |
| **BRISK**$_{\text{cLSGAN}}$ | 1 | 705 | 697 | 42 | 2.14 | 2.20 | 1.04 | 1.04 |
| | 2 | 343 | 393 | 27 | 2.12 | 2.36 | 1.12 | 1.04 |
| | 3 | 1065 | 520 | 40 | 2.04 | 2.47 | 1.06 | 1.03 |
| | 4 | 356 | 164 | 28 | 2.11 | 2.28 | 1.03 | 1.06 |
| | 5 | 6054 | 2834 | 101 | 2.26 | 2.35 | 1.13 | 1.04 |
| | 6 | 5977 | 1052 | 94 | 2.30 | 2.30 | 1.07 | 1.10 |

**Table 5.7:** Influence of the empirical distance threshold on the numbers of tie points and their accuracies and precisions obtained through a SIFT- and BRISK-based matching between the artificial SAR-like and SAR image patches with respect to the six optical and SAR test image pairs.

A quantitative analysis of the results of the image registration process is provided in Figure 5.14 and Figure 5.15. Here, the checkerboard overlays of two different optical and SAR image pairs from test image one are shown. The overlays in Figure 5.14(a) and Figure 5.15(a) show the uncorrected optical and SAR image, where the residual alignment error between the images is clearly visible in northing direction along the runway and roads. The overlays in Figure 5.14(b) and Figure 5.15(b) on the other hand show the corrected optical with the same SAR images. In contrast to Figure 5.14(a) and Figure 5.15(a) the images here seem to be aligned.

### 5.2.5 Summary

We proposed a new concept for the problem of multi-modal image matching, based on conditional generative adversarial networks (cGANs). Different cGANs setups were trained for the task of generating SAR-like image patches from optical images and for the reversed task. We showed the ability of cGAN for the task of realistic looking (despeckled) optical and SAR image generation. Beyond that, we showed the feasibility to improve the matching accuracy and precision of a NCC-, SIFT- and BRISK-based matching between optical and SAR image patches through the use of artificially generated patches. By performing a BRISK-based matching between SAR and artificial SAR-like patches we achieved tie points with an average Euclidean distance to the ground truth locations of 2.22 pixels and a precision (standard deviation) of 1.10 pixels over six test image scenes. Furthermore, the quality of the tie points is stable across the different scene even though they are spread across Europe. Finally, the obtained tie points were successfully used to register the corresponding optical and SAR image pairs, improving the absolute geo-localization accuracy of the optical images. Thereby, the overall alignment error could be reduced from up to 23 m to around 5 m. In the following paragraphs, the main drawbacks and advantages of the approach are summarized and a brief outlook for potential future developments is provided.

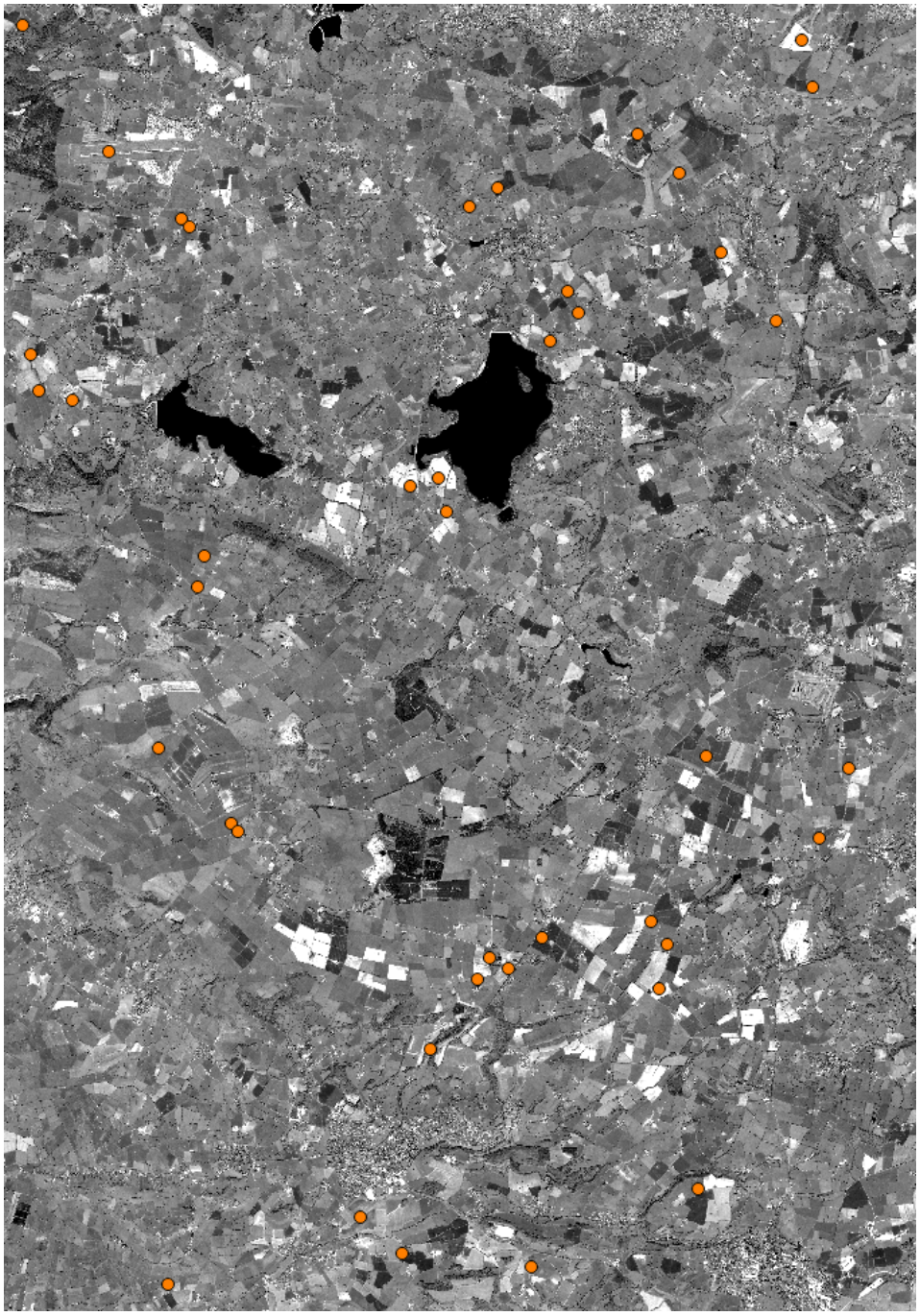
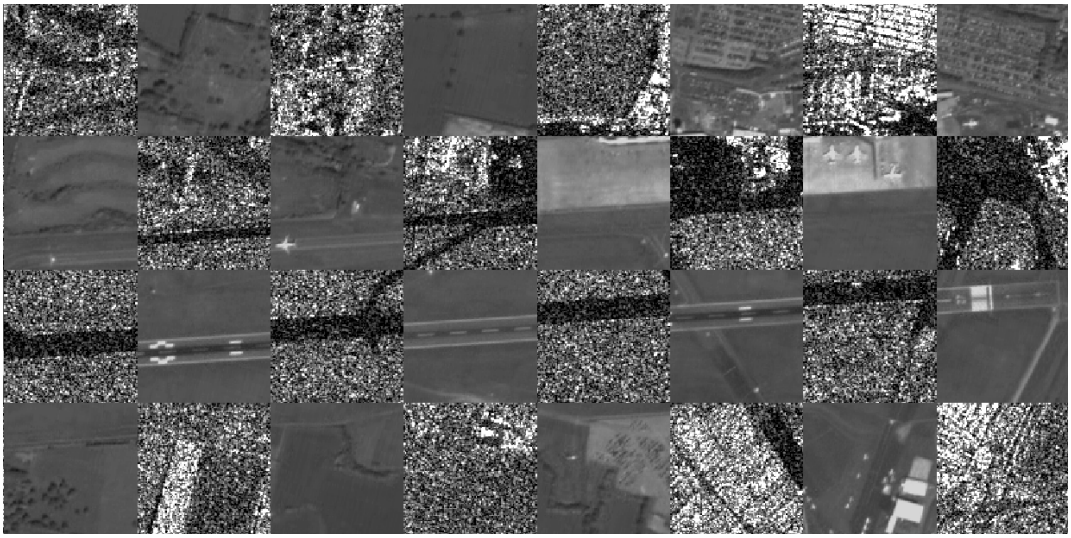
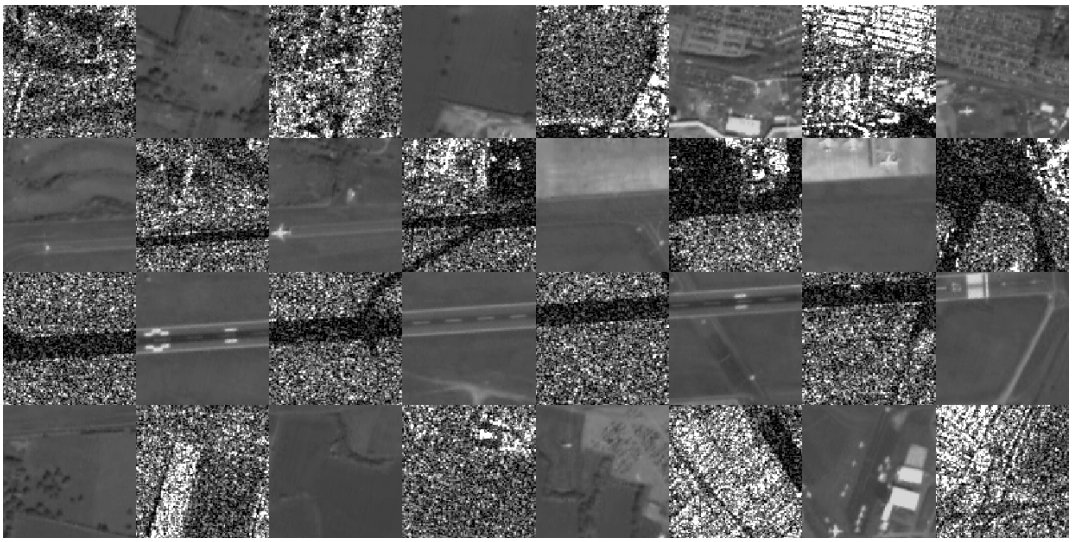


**Figure 5.13:** Illustration of the final set of tie points (marked orange) of the first test image superimposed on the corresponding optical image. The optical and SAR image pair of test scene one cover an area close to the city of Bristol, England.

**Limitations:** A general problem of our training and test data is the existing global alignment error of around 3 m. This error restricts the assessment of our method. In order to determine the actual quality of our proposed method, a set of independent reference points with an absolute geometric accuracy in the range of a few centimeter, e.g. measured with a GPS, would be needed. A problem of (conditional) GANs is the difficult validation of the training success. In contrast to other machine learning architectures, where a loss function or different



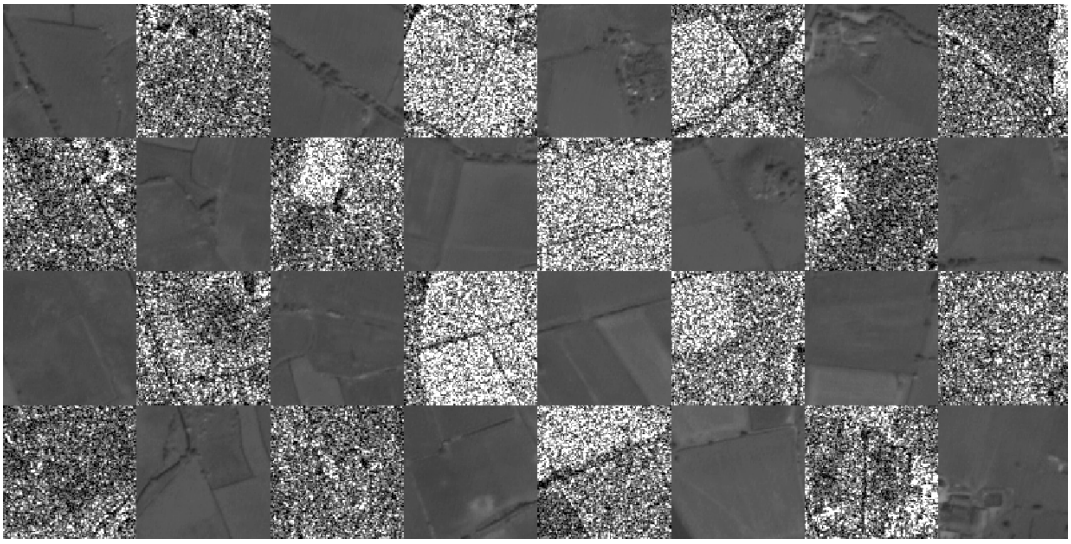
(a) Before the geo-localization enhancement of the optical image.



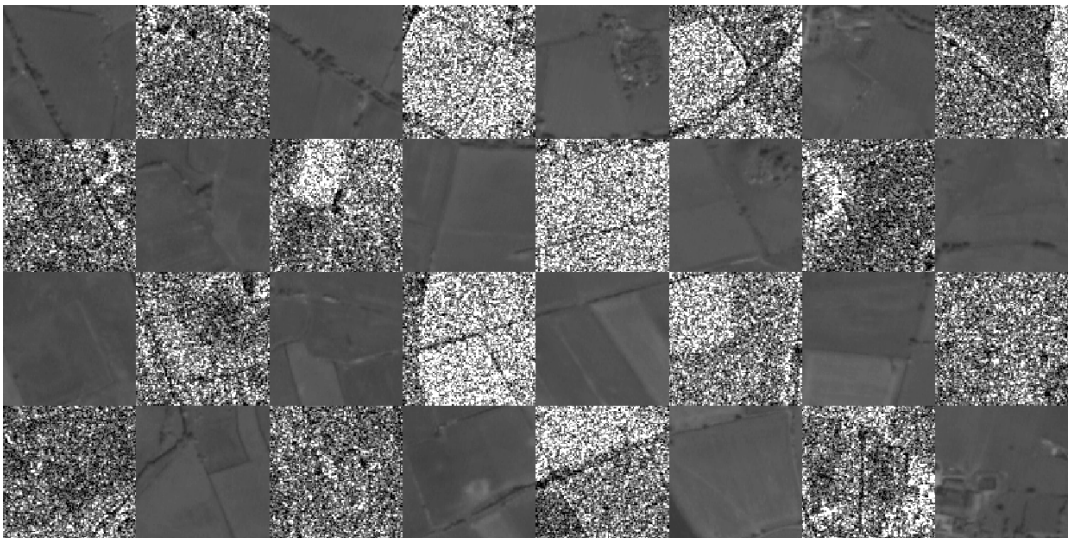
(b) After the geo-localization enhancement of the optical image.

**Figure 5.14:** Checkerboard overlays of two optical and one SAR image with a pixel spacing of 2.5 m. The image tiles have a size of 100 m. Figure 5.14(a) and Figure 5.14(b) show the optical image before after the sensor model adjustment (geo-localization enhancement) through the generated tie points, respectively.

measures can be used to evaluate the quality of the training process over a validation set, GANs require mainly a visual assessment of the generated images or (in our case) the evaluation of the matching results. This is time consuming, since every setup has to be trained till the end to find the best one. Additionally, time consuming task is the training of the cGANs, which can takes up to several weeks. Besides the high computational cost of the network training and data quality evaluation, the experiments revealed that it is important to generate patches which retain the geometric structures of the optical patches instead of generating patches which visually look like real SAR images. Therefore, not every loss and cGAN setup is applicable for the problem of optical and SAR image matching.



(a) Before the geo-localization enhancement of the optical image.



(b) After the geo-localization enhancement of the optical image.

**Figure 5.15:** Checkerboard overlays of two optical and one SAR image with a pixel spacing of 2.5 m. The image tiles have a size of 100 m. Figure 5.15(a) and Figure 5.15(b) show the optical image before and after the sensor model adjustment (geo-localization enhancement) through the generated tie points, respectively.

**Strengths:** An advantage of the proposed method is that it enables the application of well know matching techniques (NCC, BRISK and SIFT) for the matching of optical and SAR images. These methods provided high quality matchings for images acquired from the same sensor (e.g. NCC for SAR to SAR matching [194], and SIFT and BRISK for matching optical images [195]) respectively, but normally fail in the case of optical and SAR images. The evaluation of the results and the comparison with state-of-the-art matching approaches such as CAMRI [23] revealed the potential of the proposed method, and the possibility to apply it for the problem of absolute geo-localization accuracy improvement of optical images. A further benefit is the fast applicability of the method to new image scenes once the generator is trained. In such cases, artificial SAR-like patches can be generated within minutes from

optical patches. Furthermore, the variety of scenes in our training dataset, containing images acquired at different times of the year and over different locations in Europe, ensures the applicability of the method to a wide range of images acquired over different landscapes.

Overall, the proposed method opens up new possibilities for future developments towards the goal of matching optical and SAR images. The provided results validate the potential of the proposed approach in comparison to a state-of-the-art method but also reveal the need for further enhancements of the image generation process. More specific, the necessity for a generator network, which reliably and precisely retains the geometric structures of the optical images, should be the main focus of further investigations. The combination of a generator network with a deep learning-based matching approach represents thereby a promising future extension to generate more suitable artificial images patches, and hence to further improve the quality of the image matching.

## 5.3 Optical and SAR Image Registration Through Siamese Neural Networks

The second approach for the registration of optical and SAR images is based on tie points generated through Siamese neural networks. In order to find the best model for this task, several network configurations are trained and evaluated in the following. Therefore, an overview of the utilized configurations and the associated training parameters are provided in Subsection 5.3.1. An analysis and discussion about the networks qualities for an accurate and reliable tie point generation between optical and SAR image patches and a comparison of these results with traditional approaches, the state-of-the-art method CAMRI [21] and our cGAN-based matching framework is presented in Subsection 5.3.2. In Subsection 5.3.3, the potential of the generated tie points for the registration of optical and SAR images, and hence for the absolute geo-localization accuracy enhancement of optical images is discussed. Finally, the results of the proposed framework are summarized and its limitations and strengths discussed in Subsection 5.3.4.

### 5.3.1 Training Setups and Parameter Settings

Several training configurations are tested and evaluated in Subsection 5.3.2 in order to find the best Siamese neural network for an accurate and reliable tie point generation between optical and SAR image patches. The training of all models was performed over the two larger training datasets, which contain patches with a pixel spacing of 2.5 m and 3.75 m and SAR and despeckled SAR patches, respectively (for details see Table 5.1 and Subsection 5.1.2). Note, that first experiments showed that the smaller training datasets do not provide enough data to realize a successful training. Besides the different training data, we investigated the influence of two network architectures: A Siamese and a pseudo-Siamese architecture. For the Siamese architecture the weights between the two branches are shared and for the pseudo-Siamese architecture the weights of the first three layers are different, whereas the remaining layers share their weights. For more architectural details see Subsection 4.3.2.

As further mentioned in Subsection 4.3.2, each network is trained using stochastic gradient descent with the ADAM optimizer [63] and an initial learning rate of 0.01. The learning rate is reduced by a factor of 5 at iteration 60 and 80. The training with each configuration is performed 100 rounds, where each round takes 200 iterations over a single batch. We trained all networks in parallel on 2 Titan X GPUs using a batch size of 100. Note that given a batch size of 100 we actually perform 15 epochs per training (one epoch refers to one whole cycle through the entire training set). The weights of the network are initialized with the scheme described in [54], which particularly considers the rectifier nonlinearities. The whole training process takes around 30 hours.

### 5.3.2 Tie Point Generation

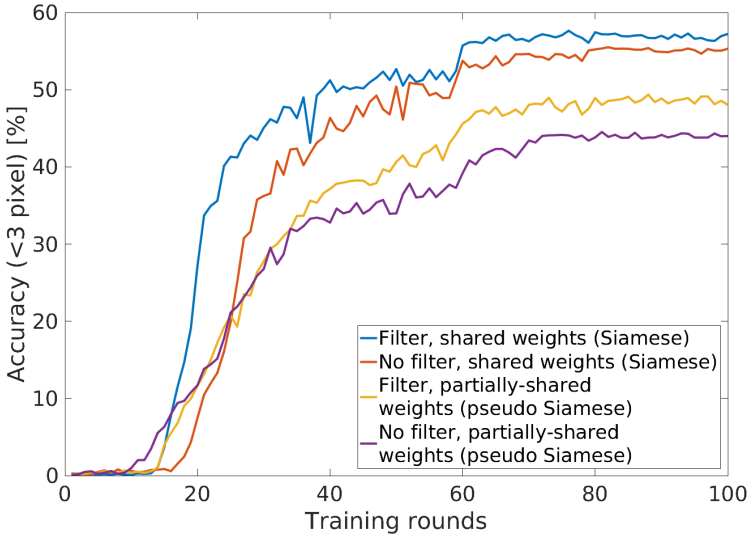
In this subsection, the quality of the tie points generated from various Siamese neural networks is evaluated. In detail, the effects of the different training configurations (no filter vs. speckle filter and Siamese vs. pseudo-Siamese architecture) on the accuracies and precisions of the corresponding sets of tie points is examined. Additionally, a comparison between the

best performing Siamese neural network and the state-of-the-art approach CAMRI [23] and our proposed cGAN-based tie point generation approach (see Subsection 5.2.3) is carried out. Note that all results shown in this and the following subsections that are related to a performance analysis of the networks during training time are obtained from the validation sets, whereas the results related to the computed set of tie points are obtained from the test sets. Neither the validation nor the test set image patches have been shown to the different Siamese neural networks during the training process. In the following, the matching accuracy is measured (next to method described in Subsection 5.1.3) as the percentage of tie points having a $L_2$ distance with less than 3 pixels to the ground truth location.

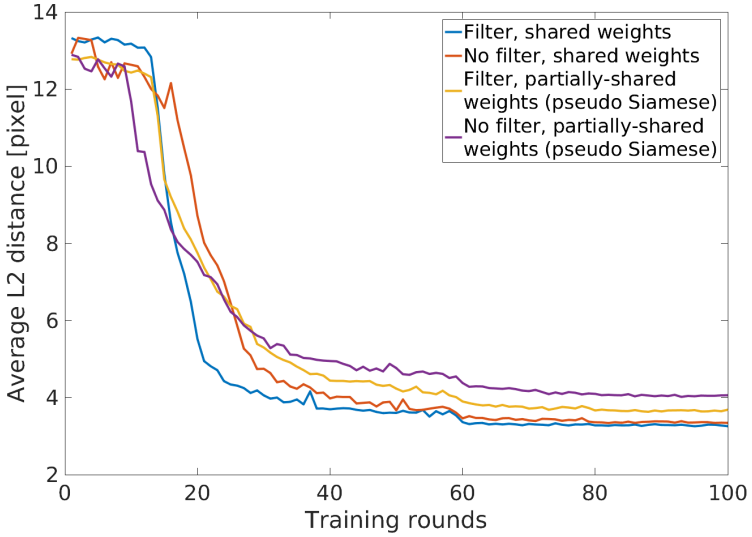
**Influence of Speckle Filter:** In order to find the optimal training configuration we first investigated the influence of a speckle filter on the learning process of the networks and on the quality of the resulting tie points. Therefore, Figure 5.16 shows the accuracies of the tie points computed between the optical and SAR and, the optical and despeckled SAR patches of the validation set during training time. In Figure 5.16(a) the matching accuracy is measured as the percentage of tie points, where the Euclidean distance to the ground truth location is less than 3 pixels, whereas in Figure 5.16(b) the matching accuracy is measured as the average over the Euclidean distances between the computed tie points and the ground truth locations. Both images reveal that regardless of the network architectures, using a speckle filtering helps the network to better learn the similarities between optical and SAR patches and thus, to improve the accuracy of the generated tie points.

**Comparison of Network Architectures:** We further investigated the influence of partially-shared (pseudo-Siamese architecture) and shared weights (Siamese architecture) between the two network branches during training. Note that in the case of the pseudo-Siamese architectures, the weights of the first three layers are different whereas the remaining layers share their weights and in the case of the Siamese architectures, all weights are shared. Figure 5.16 shows a comparison of the matching accuracy between the results of the Siamese and pseudo-Siamese architecture over the validation set. It can be seen that a Siamese architectures learns slightly faster and achieve higher matching accuracies in the end. As a consequence, the following evaluations are carried out only for the best training configuration: despeckeled SAR images as reference in combination with a Siamese neural network architecture.

**Outlier Removal:** So far, we used the normalized score (after applying the soft-max function) and we selected the locations with the highest value (highest probability) within each search area as the predicted tie point in the SAR image patches. Another possibility, which was presented in Subsection 4.3.2, is to use the raw score (before soft-max) as an indicator of the confidence of the predictions. In theory, this additional quality measure should enable the detection of outliers, and hence should lead to a higher overall matching performances. In Figure 5.17 we investigated the influence of the raw score as a threshold. As shown in the single images, a higher threshold on the raw score lead to a better accuracy in terms of correct predictions as well as a smaller Euclidean distance between the predicted tie points and the ground truth locations. Note that the rough shape at the right side of the




(a)




(b)

**Figure 5.16:** Influence of a speckle filter and of different network architectures on the matching accuracy during training time. All results are generated from the validation set. Figure 5.16(a) shows the percentage of tie points, where the $L_2$ distance to the ground truth location is less than or equal to 3 pixels. Figure 5.16(b) shows the average $L_2$ distance between the tie points and the ground truth location.

curves in Figure 5.17(b) and Figure 5.17(c) is the result of an outlier. Here, an outlier has a strong influence since these numbers are computed from less than 20 test patches (tie points). By using only the first 1000 matches with the highest raw score, the average over the $L_2$ distances between the tie points and the ground truth location can be reduced from 3.91 pixels (using all matches) to 1.91 pixels, and the standard deviation (matching precision) from 3.37 to 1.14 pixels (see Table 5.4). Note that a higher threshold results in a smaller number of valid tie points, which are more reliable (in terms of the $L_2$ distance). For a later application a threshold does not have to be specified. Depending on the number of tie points $x$ needed for an image pair, the best $x$ tie points can be chosen, based on the raw score.

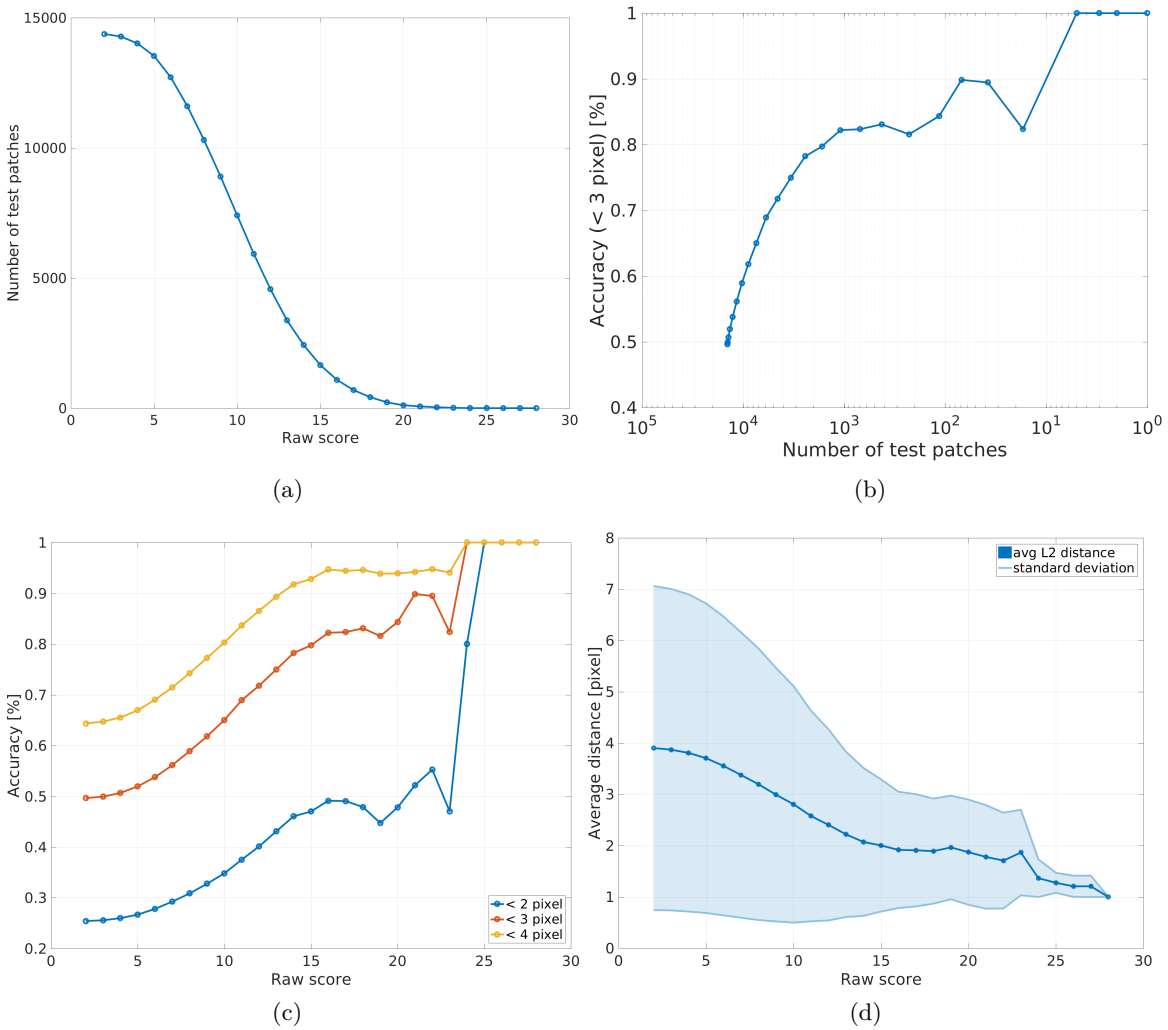



**Figure 5.17:** Influence of the raw score as a threshold. Figures 5.17(a-d) show respectively the relations between: (a) predicted score and number of patches, (b) number of patches and matching accuracy, (c) predicted score and matching accuracy, and (d) predicted score and average distance ($L_2$) between the predicted tie points and the ground truth locations. The matching accuracy in Figure 5.17(b) is measured as the percentage of tie points, where the $L_2$ distance to the ground truth location is less than 3 pixels and in Figure 5.17(c) less than 2, 3 and 4 pixels.

**Comparison to Baseline Methods:** For a better evaluation of our results, we compare our method with several baseline methods: a similarity-based matching through NCC [196] and MI [77], a feature-based matching through SIFT and BRISK, the MI-based state-of-the-art method CAMRI [23] and our own cGAN matching based approach introduced in Section 4.2. Since the evaluation presented in Subsections 5.2.3 has shown that the use of a speckle filter deteriorated the results of all utilized baseline methods, except CAMRI [23], we apply these methods on the optical and SAR image patches. In case of CAMRI, a slightly different speckle filter is implemented internally and therefore it is also applied on the optical and SAR image patches. Table 5.8 shows the comparison of our method with the baseline methods. Here, the expression "DeepMatch" denotes our Siamese-based tie point generation method, where we used a threshold to detect outliers and to generate more precise and reliable tie points (detailed explanation in the previous paragraph "Outlier Removal"). "DeepMatch" achieves

| Methods | matching accuracy < 3 pixels | matching accuracy $\mu$ [pixel] | matching precision $\sigma$ [pixel] |
|---|---|---|---|
| NCC | 35.55% | 5.50 | 4.76 |
| MI | 64.47% | 3.09 | 4.69 |
| SIFT [85] | 31.10% | 5.61 | 1.64 |
| BRISK [86] | 39.58% | 3.61 | 1.70 |
| CAMRI [23] | 57.06% | 2.80 | 2.86 |
| $\text{BRISK}_{\text{cLSGAN}}$ | 75.21% | 2.22 | 1.10 |
| DeepMatch | **82.80%** | **1.91** | **1.14** |

**Table 5.8:** Comparison of matching accuracy and precision of our method with NCC-, MI- SIFT-, BRISK-based matchings, the state-of-the-art approach CAMRI [23] and our cGAN-based matching framework over the test set. The matching accuracy is measured as the percentage of tie points, having a $L_2$ distance to the ground truth location smaller than a specific number of pixels, and as the average over the $L_2$ distances between the predicted tie points and the ground truth locations. The matching precision is represented by the standard deviation $\sigma$.

higher matching accuracy and precision than all utilized baseline methods. Furthermore, the comparison of the matching precisions reveals that our tie points with a standard deviation $\sigma$ of 1.14 pixels are, next to the BRISK-based matching in combination with the cLSGAN, the most reliable ones. The running time of our method during test time is 3.3 minutes for all 14,000 test patches on a single GPU. The baseline methods are running to a large extent on a single CPU, which makes a fair comparison difficult. Nevertheless, CAMRI [23] requires around 3 days for the computation of the tie points between the test set image patches. In Subsection 5.3.3 we will further assess the quality of the computed tie points by investigating important aspects for an accurate optical and SAR image registration, and hence for an absolute geo-localization improvement of optical images.

**Qualitative Results:** The last examination of this subsection is shown in Figure 5.18, where a side by side comparison of the score maps of the proposed approach with two similarity-based matching methods NCC and MI for several sample image patches is illustrated. Therefore, we perform our search over a search space with size $51 \times 51$ pixels, where the used patches have a resolution of 2.5 m. The images in the first column are optical image patches and the images in the last column the despeckled SAR image patches. To generate the images in column 2 to 4 we perform the matching between the corresponding image pairs using NCC, MI and our method. Yellow indicates a higher score and blue indicates a lower score. The ground truth location is in the center of each patch. Our approach performs consistently better than the corresponding baseline methods. More precisely, the score maps generated with our approach shows one high peak at the correct position, except for the last example. Here, two peaks are visible along a line which corresponds to a street in the SAR patch. In contrast, both baseline methods show a relatively large area with a constantly high score at wrong positions for most examples.

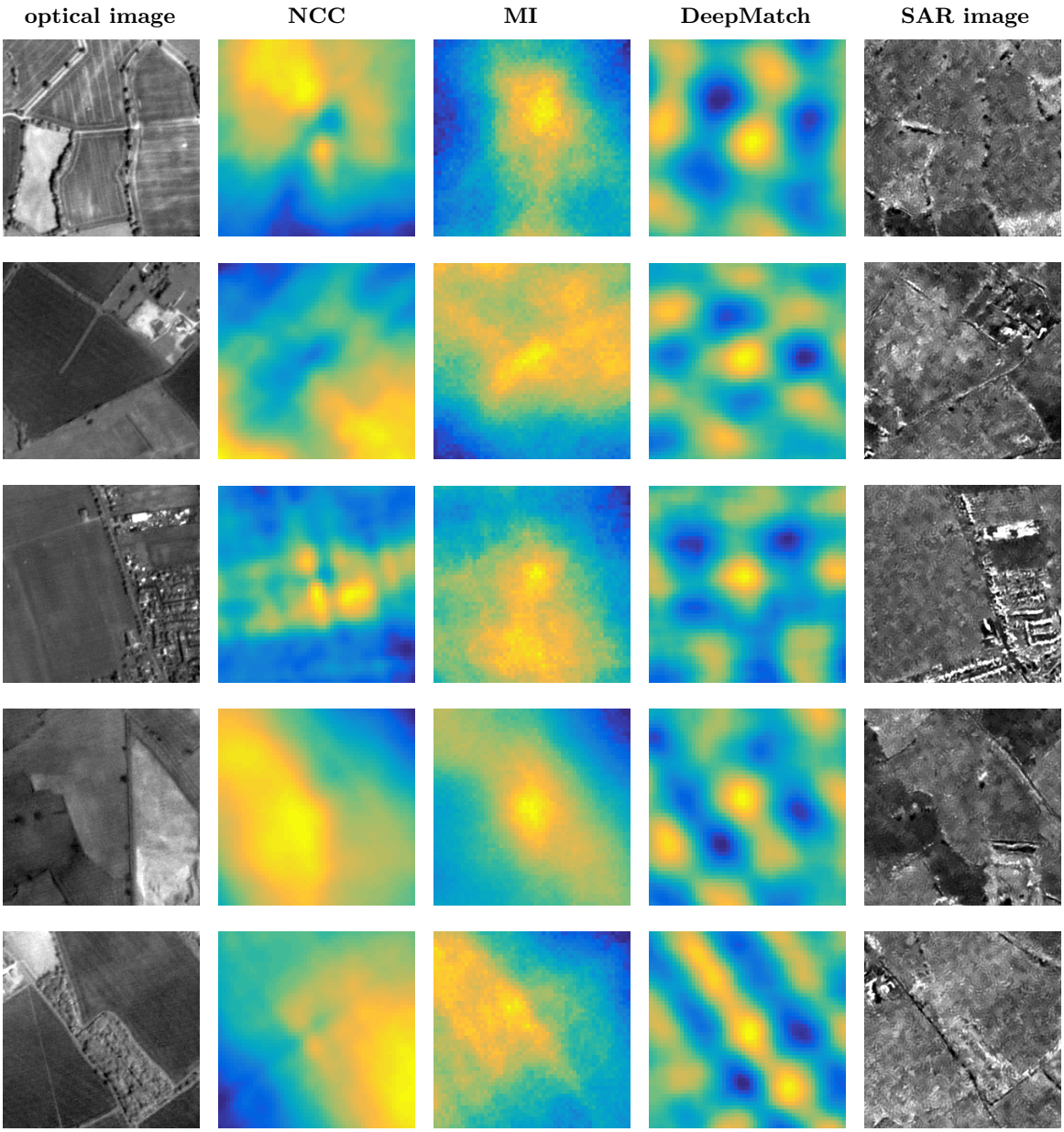



**Figure 5.18:** Side by side comparison between optical patches (201 × 201 pixels), the resulting score maps of NCC, MI and our method (51 × 51 pixels), and the despeckled SAR reference patches (251 × 251 pixels).

### 5.3.3 Geo-localization Accuracy Enhancement Though Tie Points

As mentioned in Subsection 5.2.4, a precise registration of optical and SAR images requires a reliable and accurate set of tie points. Additionally, the tie point have to be spread across the whole image scene to handle local distortions. In the previous subsection, we showed the potential of the Siamese-based optical and SAR matching framework for an accurate and precise tie point generation. Furthermore, the evaluations revealed a better performance of this methods in comparison to the state-of-the-art approach CAMRI [23] and our cGAN-based matching approach. Nevertheless, the numbers presented in Subsection 5.2.3 were obtained by using the first 1000 tie points with the highest confidence score. In the

| | Image | # of patch pairs | # tie points before | # tie points after | accuracy $\mu$ [pixel] before | accuracy $\mu$ [pixel] after | precision $\sigma$ [pixel] before | precision $\sigma$ [pixel] after |
|---|---|---|---|---|---|---|---|---|
| DeepMatch | 1 | 705 | 112 | 50 | 2.45 | 1.99 | 1.82 | 1.00 |
| | 2 | 343 | 71 | 25 | 2.60 | 2.14 | 1.87 | 1.39 |
| | 3 | 1065 | 129 | 50 | 2.14 | 1.90 | 1.31 | 0.97 |
| | 4 | 356 | 34 | 25 | 2.08 | 1.80 | 1.45 | 0.78 |
| | 5 | 6054 | 350 | 50 | 2.47 | 1.91 | 2.14 | 1.05 |
| | 6 | 5977 | 339 | 50 | 3.61 | 2.28 | 2.72 | 1.95 |

**Table 5.9:** Influence of the confidence score on the numbers of tie points and their accuracies and precisions for the sic test images. The tie points are generated through our Siamese-based matching approach DeepMatch and the application of the empirical distance threshold.

worst case, these 1000 points could all be obtained from one of the six test image pairs and located nearly at the same location. In this section, we therefore investigate whether the proposed methods is able to generate accurate and reliable tie points for each of the test image scenes, and hence is applicable for the registration of optical and SAR image patches. For this reason, we investigated the following aspects: How many tie points does the proposed method provide per optical and SAR test image pair? Do we obtain enough tie points per image scene? How accurate and precise are tie points? Are the tie points spread over the whole image or accumulated at one location?

As is Subsection 5.2.4 we set an empirical distance threshold to 50 m to ensure that the obtained tie points are spread across the image scenes and not accumulated around one location. Subsequently, we use the confidence score of the network to find the best set of tie points for each image scene. More precisely, we utilized the confidence score to find the best 50 tie points for the test images $1, 3, 5$ and 6 and, due to the lower number of test patches and resulting tie point after the distance threshold, the best 25 tie points for the test images 2 and 4. Table 5.9 contains the number of obtained tie points and their matching accuracies and precisions for each test image scene after applying the empirical distance threshold (before applying the confidence score) and after applying the confidence score. In all of the cases the use of the confidence score led to a significant improvement of the accuracy and precision of the tie point. Additionally, the confidence score allows the selection of a suitable number of tie points that can be adapted to the respective application. For a better insight of the tie point distribution within the image scene, Figure 5.19 depicts the final set of tie points of test image five. Here, the points are superimposed on the corresponding optical image. The accuracy and precision of the computed tie points varies around half a pixel between the six image scenes. It is noticeable that for test image six the most inaccurate and unprecise tie point were obtained. A possible cause is that this test image was acquired around the city of Stara Zagora, which is located far away from the training data. Additionally, for test image two, which contains the smallest amount of extracted patches, less accurate and precise tie points were computed. Both problems could be avoided in the future by applying a better and automatic matching area selection, such as described in subsection, in order to expand the training and test datasets. In the following, the obtained points are utilized to register the corresponding optical and SAR image and an absolute geo-localization accuracy of the optical images will be investigated.

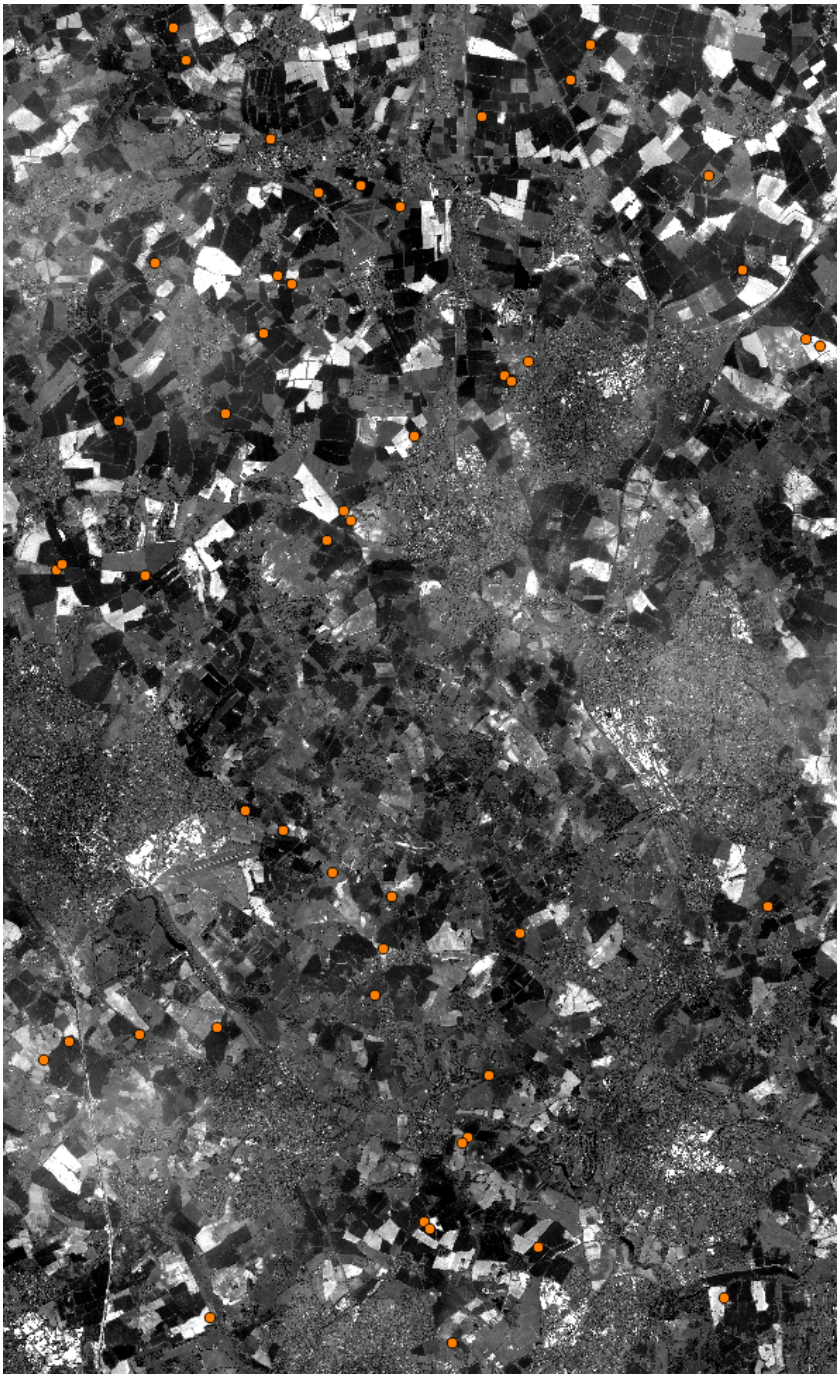
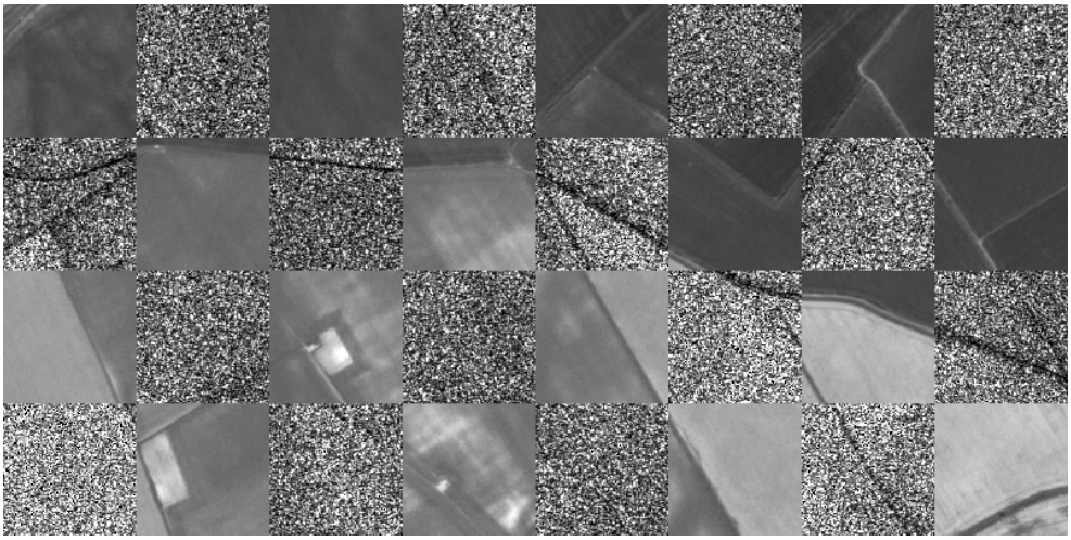
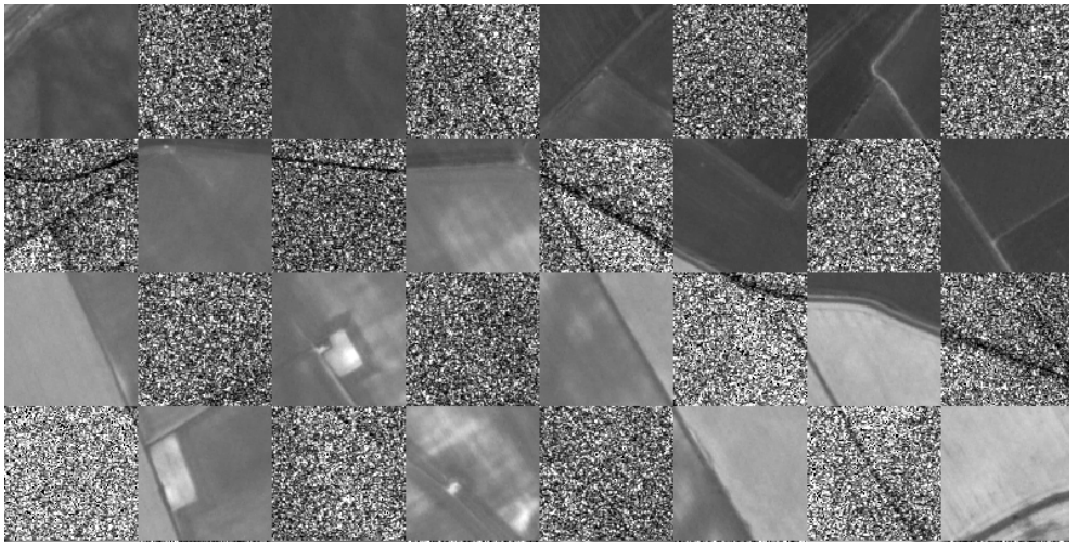


**Figure 5.19:** Illustration of the final set of tie points (marked in orange) of the fifth test image overlaid on the corresponding optical image. The optical and SAR image pair of the test scene five covers an area close to the city of London, England.

In order the enhance the absolute geo-localization of the optical test image, the final sets of tie points are utilized to improve the parameters of the corresponding sensor models. The unknown parameters of each sensor model are estimated from the corresponding set tie points by iterative least squares adjustment. During this process, a blunder detection removed around 5-10% tie point from the final sets. For details about this process we refer to Section 4.4. At the end, we used the improved sensor model to compute new orthorectified



(a) Before the geo-localization enhancement of the optical image.

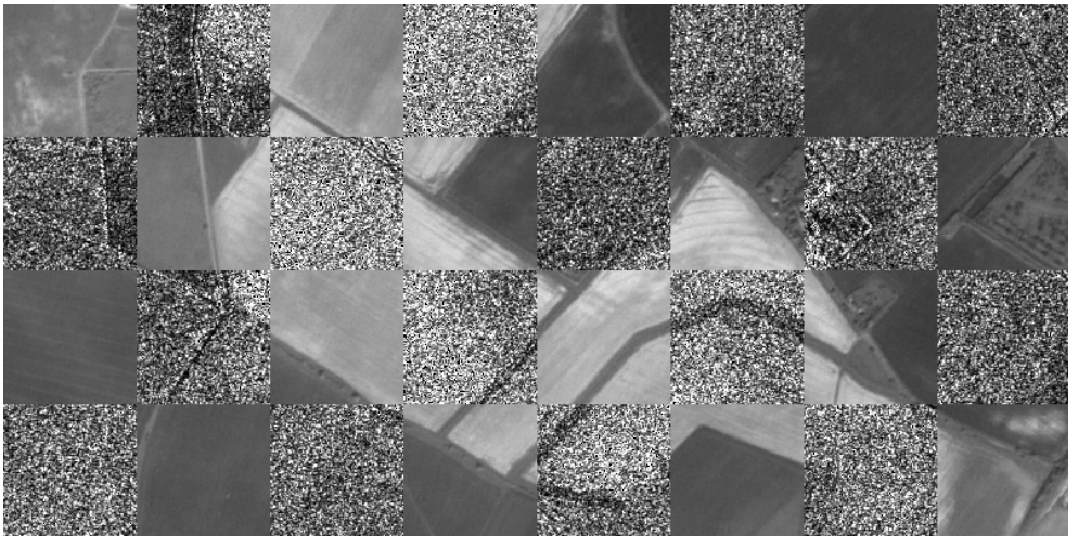


(b) After the geo-localization enhancement of the optical image.

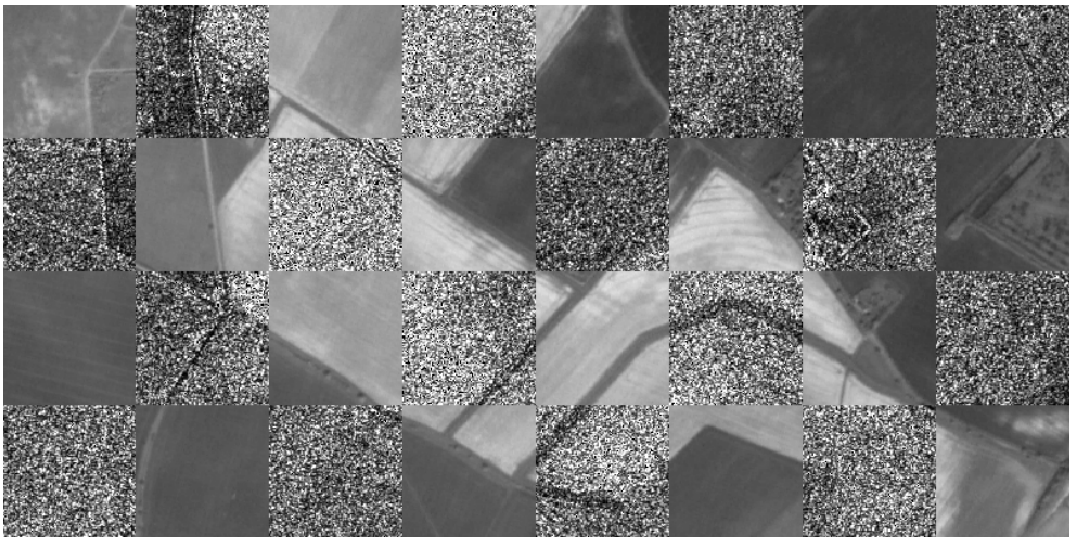
**Figure 5.20:** Checkerboard overlays of two optical and one SAR image with a pixel spacing of 2.5 m. The image tiles have a size of 100 m. Figure 5.20(a) and Figure 5.20(b) show the optical image before and after the sensor model adjustment (geo-localization enhancement) through the generated tie points, respectively.

optical image, which exhibit an improved absolute geo-localization accuracy in contrast to the optical image orthorectified with the original sensor model.

A quantitative analysis of the results of the image registration process five is provided in Figure 5.20 and Figure 5.21. Here, the checkerboard overlays of two different optical and SAR image pairs from test image five are shown. The overlays in Figure 5.20(a) and Figure 5.21(a) show the uncorrected optical and SAR images, where the residual alignment error between the images is clearly visible in easting direction along the roads. The overlays in Figure 5.20(b) Figure 5.21(b) show the corrected optical images and the identical SAR images that seems, in contrast to Figure 5.20(a) and Figure 5.21(a), to be aligned.



(a) Before the geo-localization enhancement of the optical image.



(b) After the geo-localization enhancement of the optical image.

**Figure 5.21:** Checkerboard overlays of two optical and one SAR image with a pixel spacing of 2.5 m. The image tiles have a size of 100 m. Figure 5.21(a) and Figure 5.21(b) show the optical image before and after the sensor model adjustment (geo-localization enhancement) through the generated tie points, respectively.

### 5.3.4 Summary

Our second optical and SAR image registration framework on based in the automatic generation of tie points through a Siamese neural network. In detail, a Siamese neural network has been trained to learn the similarity between optical and SAR images patches, and hence to spatially shift patches. The network is composed of a feature extraction part (Siamese neural network) and a similarity measure part (dot product layer). The network was tested on $14,000$ pairs of patches cropped from optical and SAR satellite image pairs acquired over 6 urban areas spread across Europe. Our results proved an effective generation of accurate and reliable tie points between optical and SAR images patches, outperforming

state-of-the-art matching approaches, such as CAMRI [23] and our cGAN-based matching framework (see Subsection 5.2.3). In particular, tie points can be achieved with an average $L_2$ distance from known locations of 1.91 pixels and a precision (standard deviation) of 1.14 pixels. Furthermore, by utilizing the resulting improved sensor model for the geo-referencing and orthorectification processes, we achieve an enhancement of the geo-localization accuracy of the optical images. Regarding the observed accuracies along the tie points, our method is able to reduce the overall alignment error from 23 m to under 5 m.

**Limitations:** Our training, validation and test dataset has a global alignment error of 3 m. In addition to the resulting assessment problems of the quality of our method (as described in the limitations of the cGAN approach in Subsection 5.2.5), the alignment error causes here another problem. As the training process of the Siamese neural network is directly based on the tie point generation, we could penalize the network for predictions that are actually more precise then the ground truth. This could hamper the learning process and limit the quality of the generated tie points. A drawback of the current network architecture is the restriction to input patches of size $201 \times 201$ pixels for the left branch of the network. If we would use the full resolution of the SAR images and upsample the optical images to 1.25 m, our training and test dataset would contain a large amount of image patches containing just one straight line (street segment). These patches are ambiguous for our two dimensional search, and hence not suitable for the training process. As a consequence, we need larger image patches to reduce the amount of ambiguities. Therefore, we downsampled the optical and SAR images. Due to memory limits of our available GPUs, it was not possible to increase the input patch size and simultaneously keep a proper batch size. Possible solutions could include the investigation of a new network architectures, enabling the use of larger input patches, or a better selection process of the patches, e.g. only patches containing street crossings. The processing chain for the generation of our dataset and the relative small amount of training data represent the main current weaknesses. The selection of the image patches for the dataset was mainly done manually and is limited to one optical and SAR satellite sensor (PRISM and TerraSAR-X). Through the usage of OpenStreetMap and/or a road segmentation network, the generation of the dataset could be done automatically and our datasets could be promptly extended with new image patches. A larger dataset would help to deal with the problem of overfitting during training, and further improve the network performance. Additionally, the success of our approach depends on the existence of salient features in the image scene. To generate reliable tie points, these features have to exhibit the same geometric properties in the optical and SAR image, e.g. street-crossings. Therefore, the proposed method is not trained to work on images without such features, e.g. images covering only woodlands, mountainous areas or deserts.

**Strengths:** The results prove the potential of our method for the task of geo-localization improvement of optical images through SAR reference data. By interpreting the raw network output as the confidence for predicted tie points (predicted shifts) between optical and SAR patches, we are able to generate tie points with high matching accuracy and precision. Furthermore, the high quality of the tie points does not increase the computation time. After training, we can compute new tie points between arbitrary optical and SAR image pairs within seconds. In contrast, a MI-based approach like CAMRI [23] needs up to several days

to compute the tie points between the same image patches, yielding less accurate results. In contrast to other deep learning-based matching approaches, our network is able to match multi-modal images with different radiometric properties, is extendable to other optical or radar sensors with little effort, and is applicable to multi-resolution images. In contrast to other feature-based matching approaches, our method is based on reliable (in terms of equal geometric properties in the optical and SAR image patches) features, e.g. streets and street crossings, which frequently appear in many satellite images. Furthermore, the variety in our training image pairs makes our method applicable to a wide range of images acquired over different landscapes or at different times of the year.

Overall, the proposed method has proven its qualities for the task of optical and SAR image registration and represents a promising basis for further developments. However, for a better assessment of the Siamese tie point generation approach, but also of the cGAN-based approach, we provide an extensive comparison of both methods in the following subsection.

## 5.4 Comparison of the Image Registration Frameworks

In this thesis, we presented two novel deep learning-based approaches for the registration of optical and SAR images through automatic generated tie points and performed an extensive evaluation. Thereby, both methods have shown their potential for an accurate and reliable tie point generation, and hence for an accurate and precise registration of optical SAR image pairs. However, each approach has specific advantages and disadvantage that have to be taken into account for further developments. For a better assessment of the individual methods, we will compare them on the basis of various aspects.

**Training Time and Handling:** Both methods utilize deep learning techniques, and hence require a sufficient amount of training data. Due to the specific concept of cGAN, the involved networks are less vulnerable to the problem of overfitting in comparison with Siamese neural networks. As a consequence, the cGAN matching framework requires less training data, which is an asset for a fast and efficient extension to new optical and SAR image pairs, e.g. acquired from different sensors and with another pixel spacing. On the other hand, the training of cGANs is very time consuming and the monitoring of the training process difficult. So far, our cGAN losses are not based on the actually problem of image matching and only penalize patches that visually look less like real SAR or optical images. Whether the final generator is able to provide useful patches for the task of image matching or not does not become clear before performing and evaluating the SIFT- or BRSIK-based patch matching. This circumstance has an effect on the training time, as we do not know when the best time to finish the training has come. For the Siamese neural network we can monitor the quality of the generated tie points over a validation set and stop the training when there is no sign of improvement after several training iteration.

**Test Time and Handling:** With respect to the test time (the duration of tie point computation after the training) the performance of the Siamese-based method is much higher than of the cGAN-based one. In detail, the Siamese neural network can be used directly to generate tie points for a series of image patch pairs after the training, whereby the entire process takes only a few seconds. In contrast, the trained generator network is only able to generate artificial images from given input patches. The actually tie point generation has to be performed subsequently and, depending on the method, requires up to several hours. Therefore, the handling of the Siamese-based approach is easier since it provides an end-to-end solution for the problem of tie point generation between image patches. Additionally, the Siamese-based matching approach provides an included quality measure through the raw network output. This raw network output can be treated as an confidence score, which enables the removal of doubtful tie points, while providing the possibility to keep a desired amount of tie points. For the cGAN methods on the other hand, additional methods such as RANSAC have to be utilized to remove outliers from the resulting set of tie points. Here, the number of final tie points can only be controlled indirectly through different input parameters.

**Quality of the Results:** The final results of both methods are summarized in Table 5.10. For all six test areas, our Siamese neural network based method DeepMatch provides more accurate and precise set of tie points even though this method is not able to perform the

| | Image | # of patch pairs | # tie points | accuracy $\mu$ [pixel] | precision $\sigma$ [pixel] | error reduction [%] |
|---|---|---|---|---|---|---|
| **BRISK**$_{\text{cLSGAN}}$ | 1 | 705 | 42 | 2.20 | 1.04 | 40.06 |
| | 2 | 343 | 27 | 2.36 | 1.04 | 83.45 |
| | 3 | 1065 | 40 | 2.47 | 1.03 | 71.75 |
| | 4 | 356 | 28 | 2.28 | 1.06 | 74.08 |
| | 5 | 6054 | 101 | 2.35 | 1.04 | 75.35 |
| | 6 | 5977 | 94 | 2.30 | 1.10 | 71.07 |
| **DeepMatch** | 1 | 705 | 50 | 1.99 | 1.00 | 44.26 |
| | 2 | 343 | 25 | 2.14 | 1.39 | 83.45 |
| | 3 | 1065 | 50 | 1.90 | 0.97 | 73.68 |
| | 4 | 356 | 25 | 1.80 | 0.78 | 77.52 |
| | 5 | 6054 | 50 | 1.91 | 1.05 | 79.17 |
| | 6 | 5977 | 50 | 2.28 | 1.95 | 71.32 |

**Table 5.10:** Comparison between the cGAN- and Siamese-based matching frameworks with regard to the quality and quantity of the obtained tie points for each of the six test image scenes.

matching on a sub-pixel accuracy level. For test image number two and four, both methods provide less tie points, which is probably caused through the smaller set of test images patches. In the case of DeepMatch, the smaller set of existing test image patches also affects the quality of the final set of tie points. Overall, the performance of the cGAN-based framework is more constant between the different test image scenes, whereas the Siamese-based framework is able to provide tie point with an accuracy of less than two pixel for three out of six image pairs. On the other hand, the tie point of both methods were successfully utilized for the absolute geo-localization improvement of the optical images, where the initial alignment error of the test images (see Table 5.2) could be reduced by up to 83% (see Table 5.10).

**Potential for Future Developments:** Both methods enable to achieve the goal of an accurate and precise tie point generation and hence an improvement of the absolute geo-localization accuracy of optical images. Nevertheless, both methods have the potential for a further increase in performance through the use of additional training data and developments or structural changes in the training process. In the case of the Siamese-based methods, a training on more data acquired from different sensors and with different pixel spacing would be of great interest. In the case of the cGAN-based method, the introduction of the actual problem of image matching into the training process provides great potential. So far, the training of the cGANs is geared to the problem of generating images, which look realistic enough to "fool" the discriminator. The results reveal that patches, which look more like real SAR images not necessarily lead to better matching results. Therefore, it is more important to preserve features such as edges or corners, which are beneficial for a matching technique, in the artificial patches. This could be realized through the combination of both approaches by including the generator network into the Siamese architecture. By replacing the discriminator with the a Siamese matching network the training of the generator could be tailored towards the problem of generating artificial patches, which lead to better matching results than using the original optical patches.

# 6

# Conclusion and Future Work

This work was motivated by a common problem of optical satellite imagery, namely their lower absolute geo-localization accuracy compared to SAR images. Utilizing imprecisely geo-localized images directly or in combination with additional data can cause reduced information retrieval especially in joint data evaluation. The problem can be tackled by adjusting the optical sensor model parameters through the use of tie points computed between optical images and images with a highly better absolute geo-localization accuracy such as TerraSAR-X images. Therefore, a framework for the generation of accurate and reliable tie points between high-resolution optical and SAR satellite imagery was developed, evaluated and discussed in this thesis.

Previous research studies investigated several strategies in order to handle the problem of optical and SAR image registration. The utilized matching concepts of these approaches can mainly be divided into intensity-based approaches, which rely on pixel intensity values in order to measure the similarity between images, and feature-based approaches, which rely on the detection, extraction and matching of salient image features. Recent studies further investigated the combination of both matching types in order to combine their strengths and to overcome their individual weaknesses, and thereby achieving promising results. However, all of these traditional methods are not able to learn the detection and extraction of features and thus rely on carefully tailored processing steps in order to handle the different imaging properties of optical and SAR imagery. As a consequence, most methods are developed ad hoc for the matching of a certain image feature, and are hence limited to the registration of specific image scenes.

In contrast to the traditional methods we base our work on deep learning techniques, which provide new possibilities for this research field. Our main contribution is the development of a framework for the absolute geo-localization accuracy enhancement of optical satellite images, which is based on the usage of two novel and general optical and SAR image matching methods. These two methods built on existing knowledge about the utilized images (in order to provide an optimal initial situation) and on neural networks (in order to realize an automatic image matching). The presented framework is thereby divided into three parts: the selection of suitable matching areas, the generation of reliable and accurate tie points, and a sensor model adjustment of the optical sensors.

The first step of the framework selects suitable matching areas. The main outcomes of this step are the following:

- Based on existing knowledge about the geometric properties of optical and SAR imagery, a semi-automatic concept has been created and applied in order to select suitable matching areas such as street and street crossings. The selection is based on the usage of a manual refined CORINE land cover layer. Through the pre-selection geometric differences between the images are strongly reduced and the existence of reliable and salient features in the areas to be matched is enhanced. As a consequence, the reliability of the later generated tie points with regard to their geo-localization increases. The achieved dataset (later split into a training, validation and test set) includes around $160,000$ patches organized in pairs, cropped from 46 optical and SAR images spread across Europe. A drawback of the approach is the manual refinement step, which is time consuming, and hence hampers the extension to new image pairs.

- To overcome the drawback of a manual refinement and to simplify the future usage and further developments, we developed a fully automatic area selection approach, extracting areas along existing road networks. The road network is thereby identified with the help of available OSM data and a novel deep learning-based concept developed especially for the task of street detection in SAR images. This enables a faster adaption to new image pairs, and hence further improves the quality and abilities of the matching framework. On the other hand, the true quality of automatic selected areas has to be evaluated in the future.

The second step of the framework includes a tie point generation process realized through two deep learning-based matching approaches. The main outcomes of this step can be summarized in the following points:

- Two deep learning-based architectures have been developed that are capable of generating a set of reliable and accurate tie points from a set of pre-selected optical and SAR image patches. Because of the variety in our training image pairs (in landscape and acquisition time) our methods are applicable to a wide range of images acquired over different locations or at different times of the year. The evaluation of both approaches on an independent set of test image pairs revealed their capacity of generating reliable sets of tie points with a stable quality across different image scenes, and hence their potential to set the basis for a general image registration framework. Both methods outperformed state-of-the-art approaches.

- The cGAN-based matching frameworks helped to overcome radiometric differences between optical and SAR images by translating the former to the later. Through the use of the artificially generated images patches, the application of traditional SIFT- and BRISK-based image matching become feasible. More precisely, the matching accuracy of SIFT could be increased around 57% to an average accuracy of 2.4 pixels (average Euclidean distance to the ground truth locations) and with a precision of 1.05 pixels (standard deviation). In the case of a BRISK-based matching, the accuracy could be increased around 38% to an average accuracy of 2.22 pixels and with a precision of 1.10 pixels. An open problem of this approach is still the time consuming cGAN training, along with its difficult performance analysis with regard to its tie point generation ability. Furthermore, this approach still relies on the success of the handcrafted matching approaches SIFT and BRISK. A benefit is the fast applicability of the proposed method to new image scenes once the image generator network is trained.

- The developed Siamese neural network-based tie point generation approach, on the other hand, does not rely on a single handcrafted processing step. In contrast to other deep learning-based matching approaches, our network is able to match multi-modal images with different radiometric properties. Through the implementation of a target-oriented training procedure, the network learns to measure the similarity between optical and SAR images, and therefore to automatically generate tie points. After the training, new tie points between arbitrary optical and SAR image pairs can be computed within seconds. Furthermore, the network provides an integrated confidence score in order to assess the predicted tie points. This enables a quick and efficient

removal of outliers. Overall, a matching accuracy of 1.91 pixels with a precision of 1.14 pixels was achieved. A drawback of this approach is the large amount of required data in order to successfully train the network. As a consequence, the extension to new data from different sensors requires a more efficient and faster dataset generation concept.

In order to complete the image registration framework and to enable the geo-localization accuracy improvement of optical imagery we utilize, in a third and final step, the generated tie points and well-proven methods to adjust the corresponding sensor model parameters of the optical images. For both tie point generation approaches only 5-10% of the tie points were removed during this process, highlighting the uniform quality of the generated points between the different test images. Note that the distribution of the final set of tie points within the images depends thereby highly on the nature of the image scene (rural or semi-urban area), or more precisely, on the distribution of the suitable features within the image and the quality of the image pre-selection process. Through our image registration process, the overall alignment error of the test images could be reduced from about 23 m to 5 m. A quantitative analysis of the newly orthorectified optical images clearly showed the improvement in absolute geo-localization accuracy of the optical images from our test set, and hence the potential of our image registration framework. Overall, the usage of deep learning techniques for the problem of matching optical and SAR images enabled an automatic registration, and thus the absolute geo-localization accuracy enhancement of optical imagery acquired over various cities across Europe. In contrast to other approaches, our method is based on reliable (in terms of equal geometric properties of the optical and SAR image patches) features, e.g. streets and street crossings, which frequently appear in many satellite images. Additionally, our neural networks are extendable to images from other optical and SAR sensors and with different spatial resolutions and pixel spacing, respectively. Nevertheless, more tests have to be performed in the future in order to assess and verify the applicability to a variety of different optical and SAR sensors. As this step requires a time-consuming and costly manual registration of optical and SAR image pairs for the creation of a new training dataset, it went beyond the scope of this work.

**Future Prospects:** The developed framework tackles the problem of optical and SAR image matching for the first time with the help of deep learning techniques. For a further improvement of the quality of the image registration results, the following is proposed:

- As the results of the proposed framework highly depend on the quality of the training dataset, the realization of the proposed automatic matching area selection process should be pursued in the future. This step requires the combination of the extracted road information from the SAR images, the extracted streets crossings from OSM data and information about the desired land classes from the CORINE layer. The benefit of such an extension is that both tie point generation methods could be quickly and easily adapted to new optical and SAR training pairs acquired from a variety of different sensors. Furthermore, having such an automatic training dataset generation process would enable a better validation of both tie point generation methods and, hence the overall framework, for in a more general setting.

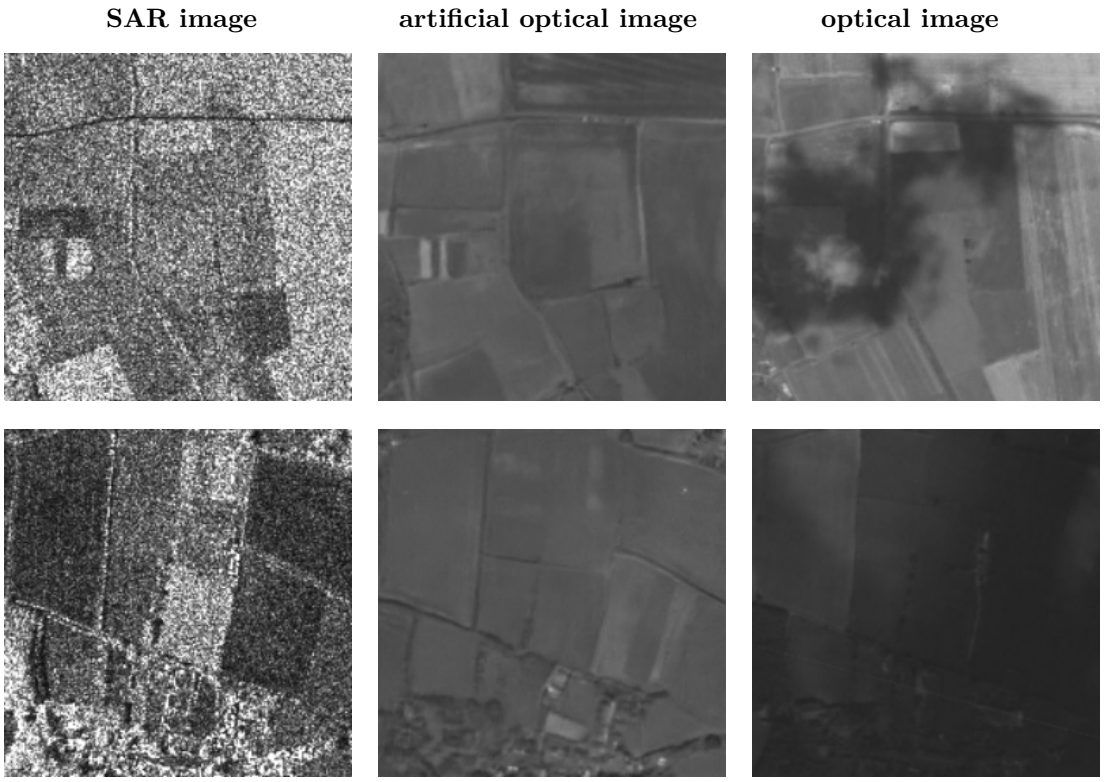



**Figure 6.1:** Side by side comparison between SAR, artificial optical and original optical sample patches with a ground sampling distance of 3.75 m.

- In case of the cGAN matching framework, the investigation of different generator architectures would be of interest. In particular, the influence of the network depth, number of parameters and skip connections on the image generation process would be an in interesting field of research. Furthermore, including the problem of image matching in the training process could further improve the quality of the tie point generation. For this set, a new training strategy and objective have to been developed.

- So far, the artificial optical images generated from SAR images could not be successfully applied for the task of accurately and reliably generating tie points. Nevertheless, this direction could be interesting for other applications, such as a better interpretation or visual understanding of SAR images or to enable a first indication of the image content for areas covered by clouds or their shadows in optical images as exemplified in Figure 6.1. In order to reach this goal, the latest cGAN architectures and training concepts should be investigated in order to find the best setup for an SAR to optical image translation process.

- In the case of the Siamese-based matching framework, the influence of data augmentation techniques (e.g. rotation, flipping), alternative network architectures (e.g. fully shared weights, less layers/weights), similarity measures (e.g. Minkowski distance, Euclidean distance, fully connected layers) and loss functions (e.g. hinge loss, triplet loss) on the accuracy and precision of the tie points is of great interest and should be fruther investigated in the future. Additionally, the resulting score maps could be

further enhanced by using curve fitting or interpolation techniques to enable sub-pixel accuracies.

- Aside from the improvement of the each single methods, the combination of a generator network with a deep learning-based matching approach represents an promising future extension. Thereby, more suitable artificial images patches could be generated, which could further improve the quality of the image matching process. In detail, by replacing the discriminator with a Siamese matching network, the image generation process could be geared towards the problem of image matching and directly be integrated in the matching frameworks. This would lead to the generation of artificial images, which will probably not look like real SAR images anymore, but facilitate the image matching process. A challenge of this idea, is to develop a suitable training procedure and to combine both networks into one architecture, while keeping the number of training parameters within a reasonable bounds (the network has to fit on the available GPUs).

# List of Symbols

| | |
|---|---|
| $*$ | Convolution operator |
| $*_d$ | Dilated convolution operator |
| $\boldsymbol{\varepsilon}$ | Boresight angles |
| $\lambda$ | Learning rate |
| $\phi_i$ | Unary potential |
| $\phi_{i,j}$ | Pairwise potential |
| $\sigma$ | Activation function |
| $\boldsymbol{\theta}$ | Set of trainable parameters |
| $\widehat{\boldsymbol{\theta}}$ | Predicted parameters |
| $\boldsymbol{\theta}^*$ | Optimal parameters |
| $\boldsymbol{\theta}^{(D)}$ | Discriminator parameters |
| $\boldsymbol{\theta}^{(G)}$ | Generator parameters |
| $a^{(t)}$ | Activation values of layer $t$ |
| $b$ | Biases of a neural network |
| $c$ | Number of channels |
| $D$ | Discriminator network |
| $d$ | Dilation factor |
| $D_{\mathrm{JS}}$ | Jensen-Shannon divergence |
| $D_{\mathrm{KL}}$ | Kullback-Leibler divergence |
| $\mathcal{D}_{\mathrm{test}}$ | Test dataset |
| $\mathcal{D}_{\mathrm{train}}$ | Training dataset |
| $\mathcal{D}_{\mathrm{val}}$ | Validation dataset |
| $\mathcal{E}$ | Overall error |
| $E$ | Expected value |

| | |
|---|---|
| $\boldsymbol{f}$ | Feature vector |
| $G$ | Generator network |
| $\boldsymbol{g}^{(D)}$ | Gradients of the discriminator |
| $\boldsymbol{g}^{(G)}$ | Gradients of the generator |
| $\boldsymbol{h}$ | Feature matrix |
| $H(\boldsymbol{I})$ | Marginal entropy of image $I$ |
| $H(\boldsymbol{I}_1, \boldsymbol{I}_2)$ | Joint entropy between image $\boldsymbol{I}_1$ and $\boldsymbol{I}_2$ |
| $h^{(t)}$ | Hidden values of layer $t$ |
| $h(x, y)$ | Joint histogram between two images $\boldsymbol{I}_1$ and $\boldsymbol{I}_2$ |
| $\boldsymbol{I}$ | Image |
| $\boldsymbol{k}$ | Filter or kernel |
| $L$ | Number of layers |
| $\mathcal{L}$ | Loss function |
| $l_i$ | The $i$-th layer of a neural network |
| $N$ | Number of input-output pairs $(x, y)$ |
| $N_b$ | Batch size |
| $n_{\text{disc}}$ | Number of discriminator iterations |
| $n_{\text{train}}$ | Number of training iterations |
| $p_{\text{data}}$ | Real data distribution |
| $p_g$ | Generator distribution |
| $p_z$ | Noise distribution |
| $\boldsymbol{R}$ | Reference image |
| $\mathcal{R}$ | Regularization term |
| $\boldsymbol{R}^{\text{Earth}}_{\text{body}}$ | Rotation from the body to the Earth coordinate system |
| $\boldsymbol{R}^{\text{body}}_{\text{sensor}}$ | Rotation from the sensor to the body coordinate system |

| | |
|---|---|
| $s$ | Search range |
| $S$ | Search space |
| $\boldsymbol{s}$ | Score map |
| $\boldsymbol{s}$ | Calibrated score map |
| $s_{\mathrm{DEM}}$ | Pixel scaling factor defined by the DEM |
| $\boldsymbol{T}$ | Template image |
| $t_{\mathrm{clip}}$ | Clipping Parameter |
| $W$ | Weights of a neural network |
| $\mathcal{W}$ | Wasserstein distance |
| $X$ | Set of input samples $x$ |
| $x^{(n)}$ | The $n$-th input element of $X$ |
| $Y$ | Set of output samples $y$ |
| $\boldsymbol{Y}_{\mathrm{bin}}$ | Ground truth distribution |
| $\widehat{y}^{(n)}$ | Predicted $n$-th output element |
| $\boldsymbol{Y}_{\mathrm{tol}}$ | Smooth target distribution |
| $z^{(n)}$ | The $n$-th input element of $Y$ |

# List of Abbreviations

| | |
|---|---|
| **(A)NN** | (Artificial) Neural Network |
| **ADAM** | Adaptive Moment Estimation |
| **ALOS** | Advanced Land Observing Satellite |
| **BRISK** | Binary Robust Invariant Scalable Key Point |
| **BN** | Batch Normalization |
| **CCRE** | Cross-cumulative Residual Entropy |
| **(c)GAN** | (Conditional) Generative Adversarial Network |
| **(c)LSGAN** | (Conditional) Least Square Adversarial Network |
| **CNN** | Convolutional Neural Network |
| **CORINE** | Coordination of Information on the Environment |
| **CRA** | Cluster Reward Algorithm |
| **CT** | Computed Tomograph |
| **(c)WAN** | (Conditional) Wasserstein Adversarial Network |
| **DEM** | Digital Elevation Model |
| **DoG** | Differences of Gaussian |
| **DOP** | Digital Orthophoto |
| **FAST** | Features from Accelerated Segment Test |
| **GCP** | Ground Control Point |
| **GT** | Ground Truth |
| **JS** | Jensen-Shannon |
| **KL** | Kullback-Leibler |
| **LSS** | Local Self-similarity |
| **MI** | Mutual Information |
| **MLP** | Multi-Layer Perceptron |
| **MRI** | Magnetic Resonance Imaging |

| | |
|---|---|
| **MSE** | Mean Squared Error |
| **NCC** | Normalized Cross-Correlation |
| **NDVI** | Normalized Difference Vegetation Index |
| **OSM** | OpenStreetMap |
| **PPB** | Probabilistic Patch-Based |
| **PRISM** | Panchromatic Remote-sensing Instrument for Stereo Mapping |
| **RANSAC** | Random Sampling Consensus |
| **ReLU** | Rectified Linear Unit |
| **RMSProp** | Root Mean Square Propagation |
| **SAR** | Synthetic Aperture Radar |
| **SID** | Squared Intensity Differences |
| **SIFT** | Scale-Invariant Feature Transform |
| **SURF** | Speeded Up Robust Features |

# List of Figures

# List of Tables

# Bibliography


[1] J. Pappachen and D. Belur. Medical Image Fusion: A Survey of the State of the Art. *Information Fusion*, 19:4–19, 2014.

[2] B. Khaleghi, A. Khamis, F. Karray, and S. Razavi. Multisensor Data Fusion: A Review of the State-of-the-Art. *Information Fusion*, 14(1):28–44, 2013.

[3] C. Pohl and J. van Genderen. Remote Sensing Image Fusion: An Update in the Context of Digital Earth. *International Journal of Digital Earth*, 7(2):158–172, 2014.

[4] M. Schmitt and X. X. Zhu. Data Fusion and Remote Sensing: An Ever-growing Relationship. *IEEE Geoscience and Remote Sensing Magazine*, 4(4):6–23, 2016.

[5] D. Brunner, G. Lemoine, and L. Bruzzone. Earthquake Damage Assessment of Buildings Using VHR Optical and SAR Imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 48(5):2403–2420, May 2010.

[6] G. Lisini, P. Gamba, F. Dell'Acqua, and F. Holecz. First Results on Road Network Extraction and Fusion on Optical and SAR Images Using a Multi-scale Adaptive Approach. *International Journal of Image and Data Fusion*, 2(4):363–375, 2011.

[7] D. Amarsaikhan, H. Blotevogel, J.L. van Genderen, M. Ganzorig, R. Gantuya, and B. Nergui. Fusing High-resolution SAR and Optical Imagery for Improved Urban Land Cover Study and Classification. *International Journal of Image and Data Fusion*, 1(1):83–97, 2010.

[8] B. Mishra and J. Susaki. SAR and Optical Data Fusion for Land Use and Cover Change Detection. In *IEEE Geoscience and Remote Sensing Symposium (IGARSS), Quebec City, QC, Canada*, pages 4691–4694, July 2014.

[9] C. Tison, F. Tupin, and H. Maitre. A Fusion Scheme for Joint Retrieval of Urban Height Map and Classification From High-Resolution Interferometric SAR Images. *IEEE Transactions on Geoscience and Remote Sensing*, 45(2):496–505, 2007.

[10] T. Perciano, F. Tupin, R. Hirata Jr., and R. Cesar Jr. A Two-level Markov Random Field for Road Network Extraction and its Application with Optical, SAR, and Multitemporal Data. *International Journal of Remote Sensing*, 37(16):3584–3610, 2016.

[11] H. Sportouche, F. Tupin, and L. Denise. Extraction and Three-Dimensional Reconstruction of Isolated Buildings in Urban Scenes From High-Resolution Optical and SAR Spaceborne Images. *IEEE Transactions on Geoscience and Remote Sensing*, 49(10):3932–3946, 2011.

[12] F. Tupin and M. Roux. Detection of Building Outlines Based on the Fusion of SAR and Optical Features. *ISPRS Journal of Photogrammetry and Remote Sensing*, 58(1):71–82, 2003.

[13] M. Schmitt and X. Zhu. On the Challenges in Stereogrammetric Fusion of SAR and Optical Imagery for Urban Areas. In *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, volume XLI-B7, pages 719–722, July 2016.

[14] M. Eineder, C. Minet, P. Steigenberger, X. Cong, and T. Fritz. Imaging Geodesy-Toward Centimeter-Level Ranging Accuracy with TerraSAR-X. *IEEE Transactions on Geoscience and Remote Sensing*, 49(2):661–671, 2011.

[15] J. Gonçalves and I. Dowman. Precise Orientation of Spot Panchromatic Images with Tie Points to a SAR Image. In *Photogrammetric Computer Vision - ISPRS Commission III Symposium, Graz, Austria*, pages 1–6, Graz, Austria, September 2002.

[16] R. Perko, H. Raggam, K. Gutjahr, and M. Schardt. Using Worldwide Available TerraSAR-X Data to Calibrate the Geo-location Accuracy of Optical Sensors. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Vancouver, BC, Canada*, pages 2551–2554, July 2011.

[17] P. Reinartz, R. Müller, P. Schwind, S. Suri, and R. Bamler. Orthorectification of VHR Optical Satellite Data Exploiting the Geometric Accuracy of TerraSAR-X Data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 66(1):124–132, 2011.

[18] N. Merkle, R. Müller, and P. Reinartz. Registration of Optical and SAR Satellite Images based on Geometric Feature Templates. In *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, International Conference on Sensors & Models in Remote Sensing & Photogrammetry, Kish Island, Iran*, volume XL-1/W5, pages 23–25, November 2015.

[19] Q. Zhao, C. Hütt, V. Lenz-Wiedemann, Y. Miao, F. Yuan, ZF. hang, and G. Bareth. Georeferencing Multi-source Geospatial Data Using Multi-temporal TerraSAR-X Imagery: a Case Study in Qixing Farm, Northeast China. *Photogrammetrie - Fernerkundung - Geoinformation*, 2015(2):173–185, 2015.

[20] J. Inglada and A. Giros. On the Possibility of Automatic Multisensor Image Registration. *IEEE Transactions on Geoscience and Remote Sensing*, 42(10):2104–2120, 2004.

[21] S. Suri and P. Reinartz. On the Possibility of Intensity Based Registration for Metric Resolution SAR and Optical Imagery. In *2th AGILE International Conference on Geographic Information Science, Hannover, Germany*, pages 1–19, June 2003.

[22] H. Chen, M. Arora, and P. Varshney. Mutual Information-based Image Registration for Remote Sensing Data. *International Journal of Remote Sensing*, 24(18):3701–3706, 2003.

[23] S. Suri and P. Reinartz. Mutual-Information-Based Registration of TerraSAR-X and Ikonos Imagery in Urban Areas. *IEEE Transactions on Geoscience and Remote Sensing*, 48(2):939–949, 2010.

[24] M. Siddique, M. Sarfraz, D. Bornemann, and O. Hellwich. Automatic Registration of SAR and Optical Images Based on Mutual Information Assisted Monte Carlo. In *IEEE International Geoscience and Remote Sensing Symposium, Munich, Germany*, pages 1813–1816, July 2012.

[25] M. Hasan, M. R. Pickering, and X. Jia. Robust Automatic Registration of Multimodal Satellite Images Using CCRE with Partial Volume Interpolation. *IEEE Transactions on Geoscience and Remote Sensing*, 50(10):4050–4061, 2012.

[26] L. Huang, Z. Li, and R. Zhang. SAR and Optical Images Registration Using Shape Context. In *IEEE International Geoscience and Remote Sensing Symposium, Honolulu, HI, USA*, pages 1007–1010, July 2010.

[27] H. Cheng. Optical Image and SAR Image Matching Based on Scattered Edge Feature. In *International Conference on Multimedia Information Networking and Security, Nanjing, Jiangsu, China*, pages 1–4, November 2010.

[28] J. Zhao, S. Gao, H. Sui, Y. Li, and L. Li. Automatic Registration Of SAR And Optical Image Based On Line And Graph Spectral Theory. In *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, pages 377–382, April 2014.

[29] H. Sui, C. Xu, J. Liu, and F. Hua. Automatic Optical-to-SAR Image Registration by Iterative Line Extraction and Voronoi Integrated Spectral Point Matching. *IEEE Transactions on Geoscience and Remote Sensing*, 53(11):6058–6072, 2015.

[30] C. Pan, Z. Zhang, H. Yan, G. Wu, and S. Ma. Multisource Data Registration Based on NURBS Description of Contours. *International Journal of Remote Sensing*, 29(2):569–591, 2008.

[31] C. Shah, Y. Sheng, and L. Smith. Automated Image Registration Based on Pseudoinvariant Metrics of Dynamic Land-Surface Features. *IEEE Transactions on Geoscience and Remote Sensing*, 46(11):3908–3916, 2008.

[32] P. Dare and I. Dowman. An Improved Model for Automatic Feature-Based Registration of SAR and SPOT Images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 56(1):13–28, 2001.

[33] B. Fan, C. Huo, C. Pan, and Q. Kong. 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, 2013.

[34] Y. Ye, L. Shen, M. Hao, J. Wang, and Z. Xu. Robust Optical-to-SAR Image Matching Based on Shape Properties. *IEEE Geoscience and Remote Sensing Letters*, 14(4):564–568, 2017.

[35] C. Xu, H. Sui, H. Li, and J. Liu. An Automatic Optical and SAR Image Registration Method with Iterative Level Set Segmentation and SIFT. *International Journal of Remote Sensing*, 36(15):3997–4017, 2015.

[36] The Electromagnetic Spectrum. `https://www.miniphysics.com/electromagnetic-spectrum_25.html`. Accessed: 2018-07-12.

[37] J. Albertz. *Einführung in die Fernerkundung: Grundlagen der Interpretation von Luft- und Satellitenbildern*. WBG (Wissenschaftliche Buchgesellschaft), 2013.

[38] C. Heipke. *Photogrammetrie und Fernerkundung: Handbuch der Geodäsie, herausgegeben von Willi Freeden und Reiner Rummel*. Springer Reference Naturwissenschaften. Springer Berlin Heidelberg, 2017.

[39] I. Cumming. and F. Wong. *Digital Processing of Synthetic Aperture Radar Data: Algorithms and Implementation*. Number Bd. 1 in Artech House Remote Sensing Library. Artech House, 2005.

[40] C. Girard and M. Girard. *Processing of Remote Sensing Data*. Taylor & Francis, 2003.

[41] S. Auer. *3D Synthetic Aperture Radar Simulation for Interpreting Complex Urban Reflection Scenarios*. Dissertation, Technische Universität München, München, 2011.

[42] S. Auer and S. Gernhardt. Linear Signatures in Urban SAR Images - Partly Misinterpreted? *IEEE Geoscience and Remote Sensing Letters*, 11(10):1762–1766, 2014.

[43] R. Werninghaus and S. Buckreuss. The TerraSAR-X Mission and System Design. *IEEE Transactions on Geoscience and Remote Sensing*, 48(2):606–614, 2010.

[44] A. Samuel. Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, 3(3):210–229, 1959.

[45] T. Mitchell. *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1997.

[46] C. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[47] E. Alpaydin. *Introduction to Machine Learning*. The MIT Press, 2014.

[48] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`, Accessed: 2018-07-12.

[49] R. Sutton and A. Barto. *Introduction to Reinforcement Learning*. MIT Press, Cambridge, MA, USA, 1998.

[50] A. Ng and M. Jordan. On Discriminative vs. Generative Classifiers: A Comparison of Logistic Regression and Naive Bayes. In *Advances in Neural Information Processing Systems 14*, pages 841–848. MIT Press, 2002.

[51] W. McCulloch and W. Pitts. A Logical Calculus of the Ideas Immanent in Nervous Activity. *The Bulletin of Mathematical Biophysics*, 5(4):115–133, 1943.

[52] F. Rosenblatt. Principles of Neurodynamics. Perceptrons and the Theory of Brain Mechanisms. Technical report, Cornell Aeronautical Laboratory, Inc., Buffalo, New York, 1961.

[53] CS231n: Convolutional Neural Networks for Visual Recognition. `http://cs231n.github.io/neural-networks-1/`. Accessed: 2018-07-12.

[54] K. He, X. Zhang, S. Ren, and J. Sun. Delving Deep Into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *IEEE International Conference on Computer Vision (ICCV), Santiago, Chile*, pages 1026–1034, 2015.

[55] D. Rumelhart, G. Hinton, and R. Williams. Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1. In *Learning Internal Representations by Error Propagation*, pages 318–362. MIT Press, Cambridge, MA, USA, 1986.

[56] A. Choromanska, M. Henaff, M. Mathieu, G. Arous, and Y. LeCun. The Loss Surfaces of Multilayer Networks. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics (AISTATS), San Diego, California, USA*, pages 1–13, May 2015.

[57] S. Ioffe and C. Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of the 32nd International Conference on Machine Learning (ICML), Lille, France*, pages 448–456, 2015.

[58] J. Bergstra and Y. Bengio. Random Search for Hyper-parameter Optimization. *The Journal of Machine Learning Research*, 13(1):281–305, 2012.

[59] J. Snoek, H. Larochelle, and R. Adams. Practical Bayesian Optimization of Machine Learning Algorithms. In *Conference on Neural Information Processing Systems (NIPS),Lake Tahoe, Nevada, US*, pages 2951–2959, December 2012.

[60] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum Learning. In *International Conference on Machine Learning (ICML), New York, NY, USA*, pages 41–48, New York, NY, USA, June 2009.

[61] J. Duchi, E. Hazan, and Y. Singer. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *The Journal of Machine Learning Research*, 12:2121–2159, 2011.

[62] T. Tieleman and G. Hinton. Lecture 6.5-RMSProp: Divide the Gradient by a Running Average of Its Recent Magnitude. *COURSERA: Neural Networks for Machine Learning*, 2012.

[63] D. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR), San Diego, California, US*, pages 1–13, San Diego, California, US, May 2015.

[64] K. Hornik. Approximation Capabilities of Multilayer Feedforward Networks. *Neural Networks*, 4(2):251–257, 1991.

[65] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[66] Y. Zhou and R. Chellappa. Computation of Optical Flow Using a Neural Network. In *IEEE International Conference on Neural Networks, San Diego, CA, USA*, pages 71–78, July 1988.

[67] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems NIPS, Montreal, Canada*, pages 1–9, December 2014.

[68] P. Luc, C. Couprie, S. Chintala, and J. Verbeek. Semantic Segmentation Using Adversarial Networks. In *Conference on Neural Information Processing Systems (NIPS) Workshop on Adversarial Training, Barcelona, Spain*, pages 1–12, Dec 2016.

[69] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA*, pages 105–114, July 2017.

[70] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative Adversarial Text-to-Image Synthesis. In *Proceedings of The 33rd International Conference on Machine Learning (ICML), New York, US*, pages 1–10, June 2016.

[71] A. Kadurin, A. Aliper, A. Kadurin, P. Mamoshina, Q. Vanhaelen, K. Khrabrov, and A. Zhavoronkov. The Cornucopia of Meaningful Leads: Applying Deep Adversarial Autoencoders for New Molecule Development in Oncology. *Oncotarget*, 8(7):10883–10890, 2017.

[72] D. Nie, R. Trullo, C. Petitjean, S. Ruan, and D. Shen. Medical Image Synthesis with Context-Aware Generative Adversarial Networks. In *Medical Image Computing and Computer-Assisted Intervention, Quebec City, QC, Canada*, pages 417–425, Quebec City, QC, Canada, September 2017.

[73] J. Guo, B. Lei, C. Ding, and Y. Zhang. Synthetic Aperture Radar Image Synthesis by Using Generative Adversarial Nets. *IEEE Geoscience and Remote Sensing Letters*, 14(7):1111–1115, July 2017.

[74] L. Ratliff, S. Burden, and S. Sastry. Characterization and Computation of Local Nash Equilibria in Continuous Games. In *51st Annual Allerton Conference on Communication, Control, and Computing (Allerton), Monticello, IL, USA*, pages 917–924, October 2013.

[75] B. Zitová and J. Flusser. Image Registration Methods: A Survey. *Image and Vision Computing*, 21(11):977–1000, 2003.

[76] J. Starck, F.. Murtagh, and A. Bijaoui. *Image Processing and Data Analysis: The Multiscale Approach.* Cambridge University Press, 1998.

[77] J. Walters-Williams and Y. Li. *Estimation of Mutual Information: A Survey*, pages 389–396. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

[78] J. Pluim, J. Maintz, and M. Viergever. Mutual-information-based Registration of Medical Images: A Survey. *IEEE Transactions on Medical Imaging*, 22(8):986–1004, 2003.

[79] W. Förstner and E. Gülch. A Fast Operator for Detection and Precise Location of Distinct Points, Corners and Centers of Circular Features, Interlaken, Switzerland. In *ISPRS Intercommision Workshop*, pages 149–155, June 1987.

[80] C. Harris and M. Stephens. A Combined Corner and Edge Detector. In *In Proc. of Fourth Alvey Vision Conference*, pages 147–151, 1988.

[81] M. Hassaballah, Aly A. Abdelmgeid, and H. Alshazly. *Image Features Detection, Description and Matching*, pages 11–45. Springer International Publishing, Cham, 2016.

[82] J. Canny. A Computational Approach to Edge Detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 8(6):679–698, 1986.

[83] D. Marr and E. Hildreth. Theory of Edge Detection. *Proceedings of the Royal Society of London B: Biological Sciences*, 207(1167):187–217, 1980.

[84] N. Pal and S. Pal. A Review on Image Segmentation Techniques. *Pattern Recognition*, 26(9):1277–1294, 1993.

[85] D. Lowe. Object Recognition from Local Scale-Invariant Features. In *IEEE International Conference on Computer Vision (ICCV), Kerkyra, Greece*, pages 1150–1158, September 1999.

[86] S. Leutenegger, M. Chli, and R. Siegwart. BRISK: Binary Robust Invariant Scalable Keypoints. In *IEEE International Conference on Computer Vision (ICCV), Barcelona, Spain*, pages 2548–2555, Barcelona, Spain, November 2011. IEEE Computer Society.

[87] M. Brown and D. Lowe. Invariant Features from Interest Point Groups. In *British Machine Vision Conference (BMWC)*, pages 253–262, September 2002.

[88] E. Rosten, R. Porter, and T. Drummond. Faster and Better: A Machine Learning Approach to Corner Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(1):105–119, 2010.

[89] A. Goshtasby. Image Registration by Local Approximation Methods. *Image and Vision Computing*, 6(4):255–261, 1988.

[90] M. Ehlers and D. Fogel. High-precision Geometric Correction of Airborne RemoteSensing Revisited: The Multiquadric Interpolation. In *Proc. SPIE*, pages 814–824, January 1994.

[91] L. Brown. A Survey of Image Registration Techniques. *ACM Computing Surveys*, 24(4):325–376, 1992.

[92] P. Doucette, H. Theiss, E. Mikhail, and D. Motsko. Spatial Analysis of Image Registration Methodologies for Fusion Applications. In *SPIE Defense, Security, and Sensing, Orlando, Florida, USA*, pages 1–12, Orlando, Florida, USA, May 2011.

[93] J. LeMoigne, N. Netanyahu, and R. Eastman. *Image Registration for Remote Sensing*. Cambridge University Press, 2011.

[94] A. Goshtasby. *2-D and 3-D Image Registration: For Medical, Remote Sensing, and Industrial Applications*. Wiley-Interscience, 2005.

[95] R. Müller, T. Krauß, M. Schneider, and P. Reinartz. Automated Georeferencing of Optical Satellite Data with Integrated Sensor Model Improvement. *Photogrammetric Engineering & Remote Sensing*, 78(1):61–74, 2012.

[96] J. Inglada. Similarity Measures for Multisensor Remote Sensing Images. In *IEEE International Geoscience and Remote Sensing Symposium, Toronto, Ontario, Canada*, pages 104–106, June 2002.

[97] L. Fonseca and B. Manjunath. Registration Techniques for Multisensor Remotely Sensed Imagery. *Journal of Photogrammetry Engineering and Remote Sensing*, 62(9):1049–1056, 1996.

[98] X. Liu, Q. Lei, Z.and Yu, X. Zhang, Y. Shang, and W. Hou. Multi-Modal Image Matching Based on Local Frequency Information. *EURASIP Journal on Advances in Signal Processing*, 2013(1):1–11, 2013.

[99] H. Cheng, S. Zheng, Q. Yu, J. Tian, and J. Liu. Matching of SAR Images and Optical Images Based on Edge Feature Extracted via SVM. In *7th International Conference on Signal Processing, Beijing, China*, pages 930–933, August 2004.

[100] R. Schowengerdt T. Hong. A Robust Technique for Precise Registration of Radar and Optical Satellite Images. *Photogrammetric Engineering & Remote Sensing*, 71(5):585–593, 2005.

[101] G. Lehureau, F. Tupin, C. Tison, G. Oller, and D. Petit. Registration of Metric Resolution SAR and Optical Images in Urban Areas. In *7th European Conference on Synthetic Aperture Radar, Friedrichshafen, Germany*, pages 1–4, June 2008.

[102] Y. Jiang. Optical/SAR Image Registration Based on Cross-correlation with Multi-scale and Multi-direction Gabor Characteristic Matrices. In *IET International Radar Conference, Xi'an, China*, pages 1–4, April 2013.

[103] I. Dowman and P. Dare. Automated Procedures for Multisensor Registration and Orthorectification of Satellite Images. In *International Archives of Photogrammetry and Remote Sensing*, volume 32, pages 37–44, 2000.

[104] H. Li, B. Manjunath, and S. Mitra. A Contour-Based Approach to Multisensor Image Registration. *IEEE Transactions on Image Processing*, 4(3):320–334, 1995.

[105] R. Touzi, A. Lopes, and P. Bousquet. A Statistical and Geometrical Edge Detector for SAR Images. *IEEE Transactions on Geoscience and Remote Sensing*, 26(6):764–773, 1988.

[106] N. Sang, T. Zhang, W. Li, and G. Wang. Fast and Effective Algorithm for Radar-to-optical Scene Matching Based on the Knowledge of Object Region. In *Proc. SPIE*, volume 2738, pages 2738–2738, 1996.

[107] C. Ai, T. Feng, J. Wang, and S. Zhang. A Novel Image Registration Algorithm for SAR and Optical Images Based on Virtual Points. In *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, pages 1–4, July 2013.

[108] E. Shechtman and M. Irani. Matching Local Self-Similarities Across Images and Videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Minneapolis, Minnesota, USA*, June 2007.

[109] A. Kelman, M. Sofka, and C. Stewart. Keypoint Descriptors for Matching Across Multiple Image Modalities and Non-linear Intensity Variations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Minneapolis, Minnesota, USA*, pages 1–7, June 2007.

[110] Y. Ye and J. Shan. A Local Descriptor Based Registration Method for Multispectral Remote Sensing Images with Non-linear Intensity Differences. *ISPRS Journal of Photogrammetry and Remote Sensing*, 90:83–95, 2014.

[111] T. Long, W. Jiaoa, G. Hea, Z. Zhanga, B. Chenga, and W Wanga. A Generic Framework for Image Rectification Using Multiple Types of Feature. *ISPRS Journal of Photogrammetry and Remote Sensing*, 102:161–171, 2015.

[112] W. Jia, J. Zhang, and J. Yang. Automatic Registration of SAR and Optics Image Based on Multi-features on Suburban Areas. In *Joint Urban Remote Sensing Event (JURSE), Shanghai, China*, pages 1–7, June 2009.

[113] Z. Wang, J. Zhang, Y. Zhang, and B. Zou. Automatic registration of sar and optical image based on multi-features and multi-constraints. In *IEEE International Geoscience and Remote Sensing Symposium, Honolulu, Hawaii, USA*, pages 1019–1022, July 2010.

[114] Y. Ye, J. Shan, L. Bruzzone, and L. Shen. Robust Registration of Multimodal Remote Sensing Images Based on Structural Similarity. *IEEE Transactions on Geoscience and Remote Sensing*, 55(5):2941–2958, 2017.

[115] M. Chen, A. Habib, H. He, Q. Zhu, and W. Zhang. Robust Feature Matching Method for SAR and Optical Images by Using Gaussian-Gamma-Shaped Bi-Windows-Based Descriptor and Geometric Constraint. *Remote Sensing*, 9(9):1–25, 2017.

[116] W. Shi, F. Su, R. Wang, and Y. Lu. Optical and SAR Image Registration Based on Improved Nonsubsampled Wavelet Transform for Sea Islands. *Acta Oceanologica Sinica*, 33(5):86–95, 2014.

[117] R. Hänsch, O. Hellwich, and X. Tu. Machine-learning Based Detection of Corresponding Interest Points in Optical and SAR Images. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China*, pages 1492–1495, July 2016.

[118] Y. Han and Y. Byun. Automatic and Accurate Registration of VHR Optical and SAR Images Using a Quadtree Structure. *International Journal of Remote Sensing*, 36(9):2277–2295, 2015.

[119] B. Xiong, W. Li, L. Zhao, J. Lu, X. Zhang, and G. Kuang. Registration for SAR and Optical Images Based on Straight Line Features and Mutual Information. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China*, pages 2582–2585, July 2016.

[120] M. Salehpour and A. Behrad. Nonrigid Synthetic Aperture Radar and Optical Image Coregistration by Combining Local Rigid Transformations Using a Kohonen Network. *Journal of the Optical Society of America A*, 34(10):1865–1876, 2017.

[121] M. Gong, S. Zhao, L. Jiao, D. Tian, and S. Wang. 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, 2014.

[122] S. Suri, P. Schwind, P. Reinartz, and J. Uhl. Combining Mutual Information and Scale Invariant Feature Transform for Fast and Robust Multisensor SAR Image Registration. In *75th Annual ASPRS Conference, Baltimore, MD, USA*, pages 1–12, March 2009.

[123] G. Palubinskas and P. Reinartz. Template Based Matching of Optical and SAR Imagery. In *Joint Urban Remote Sensing Event (JURSE), Lausanne Switzerland*, pages 1–4, March 2015.

[124] N. Merkle, R. Müller, P. Schwind, G. Palubinskas, and P. Reinartz. A New Approach for Optical and SAR Satellite Image Registration. In *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, Munich Germany*, volume II-3/W4, pages 119–126, March 2015.

[125] A. Krizhevsky, I. Sutskever, and G. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Conference on Neural Information Processing Systems (NIPS), Lake Tahoe, USA*, pages 1097–1105. Curran Associates, Inc., December 2012.

[126] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, US*, pages 770–778, June 2016.

[127] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems 28*, pages 91–99. Curran Associates, Inc., 2015.

[128] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu. Deep Learning-Based Classification of Hyperspectral Data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(6):2094–2107, 2014.

[129] G. Matthyus, S. Wang, S. Fidler, and R. Urtasun. HD Maps: Fine-grained Road Segmentation by Parsing Ground and Aerial Images. In *IEEE International Conference on Computer Vision (ICCV), Las Vegas, NV, USA*, pages 3611–3619, June 2016.

[130] J. Geng, J. Fan, H. Wang, X. Ma, B. Li, and F. Chen. High-Resolution SAR Image Classification via Deep Convolutional Autoencoders. *Geoscience and Remote Sensing Letters, IEEE*, 12(11):2351–2355, 2015.

[131] W. Luo, A. G. Schwing, and R. Urtasun. Efficient Deep Learning for Stereo Matching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, USA*, pages 1–9, June 2016.

[132] J. Zbontar and Y. LeCun. Stereo Matching by Training a Convolutional Neural Network to Compare Image Patches. *Journal of Machine Learning Research*, 17:1–32, 2016.

[133] A. Shaked and L. Wolf. Improved Stereo Matching with Constant Highway Networks and Reflective Confidence Learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA*, pages 6901–6910, July 2017.

[134] A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, and A. Bry. End-to-End Learning of Geometry and Context for Deep Stereo Regression. In *IEEE International Conference on Computer Vision (ICCV), Las Condes, Chile*, pages 1–9, December 2017.

[135] H. Park and K. M. Lee. Look Wider to Match Image Patches With Convolutional Neural Networks. *IEEE Signal Processing Letters*, 24(12):1788–1792, Dec 2017.

[136] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid. DeepFlow: Large Displacement OpticalFlow with Deep Matching. In *IEEE International Conference on Computer Vision (ICCV), Sydney, Australia*, pages 1385–1392, 2013.

[137] A. Dosovitskiy, P. Fischer, E. Ilg, P. Häusser, C. Hazırbaş, V. Golkov, P. v.d. Smagt, D. Cremers, and T. Brox. FlowNet: Learning Optical Flow with Convolutional Networks. In *IEEE International Conference on Computer Vision (ICCV), Las Condes, Chile*, pages 1–9, December 2017.

[138] M. Bai, W. Luo, K. Kundu, and R. Urtasun. Exploiting Semantic Information and Deep Matching for Optical Flow. In *European Conference on Computer Vision (ECCV), Amsterdam, Netherlands*, pages 154–170, October 2016.

[139] H. Altwaijry, J. Trulls, E.and Hays, P. Fua, and S. Belongie. Learning to Match Aerial Images with Deep Attentive Architectures. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, USA*, pages 1–9, June 2016.

[140] T. Lin, Y. Cui, S. Belongie, and J. Hays. Learning Deep Representations for Ground-to-Aerial Geolocalization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA*, pages 5007–5015, 2015.

[141] I. Rocco, R. Arandjelovic, and J. Sivic. Convolutional Neural Network Architecture for Geometric Matching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA*, pages 1–9, July 2017.

[142] B. de Vos, F. Berendsen, M. Viergever, M. Staring, and I. Isgum. End-to-End Unsupervised Deformable Image Registration with a Convolutional Neural Network. In *International Workshop on Deep Learning in Medical Image Analysis, Quebec City, QC, Canada*, pages 204–212, September 2017.

[143] P. Fischer, A. Dosovitskiy, and T. Brox. Descriptor Matching with Convolutional Neural Networks: a Comparison to SIFT. Technical Report 1405.5769, arXiv, May 2014.

[144] F. Ye, Y. Su, H. Xiao, X. Zhao, and W. Min. Remote Sensing Image Registration Using Convolutional Neural Network Features. *IEEE Geoscience and Remote Sensing Letters*, 15(2):232–236, 2018.

[145] K. Yi, E. Trulls, V. Lepetit, and P. Fua. LIFT: Learned Invariant Feature Transform. In *European Conference on Computer Vision (ECCV), Amsterdam, Nederlands*, October 2016.

[146] H. Bay, T. Tuytelaars, and L. Van Gool. SURF: Speeded Up Robust Features. In *European Conference on Computer Vision (ECCV), Graz, Austria*, pages 404–417, May 2006.

[147] A. Alahi, R. Ortiz, and P. Vandergheynst. FREAK: Fast Retina Keypoint. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, Rhode Island, US*, pages 510–517, June 2012.

[148] H. Altwaijry, A. Veit, and S. Belongie. Learning to Detect and Match Keypoints with Deep Architectures. In *British Machine Vision Conference (BMVC), York, UK*, pages 1–12, 2016.

[149] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah. Signature Verification using a "Siamese" Time Delay Neural Network. In *Conference on Neural Information Processing Systems (NIPS), Denver, Colorado, USA*, pages 737–744, 1994.

[150] X. Han, T. Leung, Y. Jia, R. Sukthankar, and A. Berg. MatchNet: Unifying Feature and Metric Learning for Patch-Based Matching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA*, pages 1–9, June 2015.

[151] S. Zagoruyko and N. Komodakis. Learning to Compare Image Patches via Convolutional Neural Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA*, pages 1–9, 2015.

[152] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, and F. Moreno-Noguer. Discriminative Learning of Deep Convolutional Feature Point Descriptors. In *IEEE International Conference on Computer Vision (ICCV), Santiago, Chile*, pages 1–9, 2015.

[153] I. Melekhov, J. Kannala, and E. Rahtu. Siamese Network Features for Image Matching. In *23rd International Conference on Pattern Recognition (ICPR), Cancun, Mexico*, pages 378–383, December 2016.

[154] I. Melekhov, J. Kannala, and E. Rahtu. Image Patch Matching Using Convolutional Descriptors with Euclidean Distance. In *ACCV 2016 International Workshops, Taipei, Taiwan*, pages 1–16, November 2017.

[155] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, Rhode Island, US*, pages 1–9, June 2012.

[156] C. Villamil-Lopez, L. Petersen, R. Speck, and D. Frommholz. Registration of Very High Resolution SAR and Optical Images. In *Proceedings of EUSAR 2016: 11th European Conference on Synthetic Aperture Radar, Hamburg, Germany*, pages 1–6, June 2016.

[157] L. Mou, M. Schmitt, Y. Wang, and X. Zhu. A CNN for the Identification of Corresponding Patches in SAR and Optical Imagery of Urban Scenes. In *Joint Urban Remote Sensing Event (JURSE), Dubai, United Arab Emirates*, pages 1–4, March 2017.

[158] M. Bossard, J. Feranec, and J. Otahel. CORINE Land Cover Technical Guide - Addendum 2000. *European Environmental Agency, Copenhagen*, 2000.

[159] B. Jeon, J. Jang, and K. Hong. Road Detection in Spaceborne SAR Images Using a Genetic Algorithm. *IEEE Transactions on Geoscience and Remote Sensing*, 40(1):22–29, 2002.

[160] M. Koch, M. Moya, J. Chow, J. Goold, and R. Malinas. Road Segmentation Using Multipass Single-pol Synthetic Aperture Radar Imagery. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Boston, Massachusetts, US*, pages 151–160, June 2015.

[161] J. Geng, H. Wang, J. Fan, and X. Ma. Deep Supervised and Contractive Neural Network for SAR Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(4):2442–2459, 2017.

[162] C. Henry, S. Azimi, and N. Merkle. Road Segmentation in SAR Satellite Images with Deep Fully-Convolutional Neural Networks. *IEEE Geoscience and Remote Sensing Letters (Early Access)*, pages 1–5, 2018.

[163] J. Long, E. Shelhamer, and T. Darrell. Fully Convolutional Networks for Semantic Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, Massachusetts, US*, pages 3431–3440, June 2015.

[164] M. Zeiler, D. Krishnan, G. Taylor, and R. Fergus. Deconvolutional Networks. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, US*, pages 2528–2535, June 2010.

[165] N. Homayounfar, S. Fidler, and R. Urtasun. Sports Field Localization via Deep Structured Models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, Hawaii, US*, pages 4012–4020, July 2017.

[166] D. Eigen and R. Fergus. Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-scale Convolutional Architecture. In *IEEE International Conference on Computer Vision (ICCV), Las Condes, Chile*, pages 2650–2658, December 2015.

[167] P. Krähenbühl and V. Koltun. Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials. In *Conference on Neural Information Processing Systems (NIPS), Granada, Spain*, pages 109–117, Granada, Spain, December 2011.

[168] L. Chen, G. Papandreou, I.s Kokkinos, K. Murphy, and A. Yuille. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2018.

[169] N. Merkle, P. Fischer, and R. Müller S. Auer. On the Possibility of Conditional Adversarial Networks for Multi-Sensor Image Matching. In *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Fort Worth, Texas, US*, pages 1–4, July 2017.

[170] N. Merkle, S. Auer, R. Müller, and P. Reinartz. 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, 2018.

[171] P. Isola, J. Zhu, T. Zhou, and A. Efros. Image-to-Image Translation with Conditional Adversarial Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, Hawaii, US*, pages 1–9, July 2017.

[172] M. Mirza and S. Osindero. Conditional Generative Adversarial Nets. In *CoRR*, volume abs/1411.1784, 2014.

[173] E. Denton, S. Chintala, A. Szlam, and R. Fergus. Deep Generative Image Models Using a Laplacian Pyramid of Adversarial Networks. In *Advances in Neural Information Processing Systems 28*, pages 1486–1494. Curran Associates, Inc., 2015.

[174] D. Yoo, N. Kim, S. Park, A. Paek, and I. Kweon. Pixel-Level Domain Transfer. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, Hawaii, US*, pages 1–9, July 2017.

[175] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. Efros. Context Encoders: Feature Learning by Inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, US*, pages 1–9, June 2016.

[176] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI), Munich, Germany,*, pages 234–241, October 2015.

[177] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In *European Conference on Computer Vision (ECCV), Amsterdam, Netherlands*, pages 694–711, October 2016.

[178] X. Wang and A. Gupta. Generative Image Modeling Using Style and Structure Adversarial Networks. In *European Conference on Computer Vision (ECCV), Amsterdam, Netherlands*, pages 318–335, October 2016.

[179] C. Li and M. Wand. Precomputed Real-Time Texture Synthesis with Markovian Generative Adversarial Networks. In *European Conference on Computer Vision (ECCV), Amsterdam, Netherlands*, pages 702–716, October 2016.

[180] S. Ruder. An Overview of Gradient Descent Optimization Algorithms. In *CoRR*, volume abs/1609.04747, 2016.

[181] X. Mao, Q. Li, H. Xie, R. Lau, Z. Wang, and S. Smolley. Least Squares Generative Adversarial Networks. In *IEEE International Conference on Computer Vision (ICCV), Venice, Italy*, pages 1–9, Venice, Italy, Oct 2017.

[182] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein Generative Adversarial Networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML), Sydney, Australia*, pages 214–223, Sydney, Australia, August 2017.

[183] C. Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2016.

[184] M. Fischler and R. Bolles. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Commun. ACM*, 24(6):381–395, 1981.

[185] N. Merkle, W. Luo, S. Auer, R. Müller, and R. Urtasun. Exploiting Deep Matching and SAR Data for the Geo-Localization Accuracy Improvement of Optical Satellite Images. *Remote Sensing*, 9(6), 2017.

[186] F. Yu and V. Koltun. Multi-Scale Context Aggregation by Dilated Convolutions. In *IEEE International Conference on Learning Representations (ICCV), San Juan, Puerto Rico*, pages 1–13, 2016.

[187] Basics of Remote Sensing: Along Track Scanning. `http://grindgis.com/what-is-remote-sensing/know-basics-of-remote-sensing`. Accessed: 2018-07-12.

[188] J. Jeong and T. Kim. Comparison of Positioning Accuracy of a Rigorous Sensor Model and two Rational Function Models for Weak Stereo Geometry. *ISPRS Journal of Photogrammetry and Remote Sensing*, 108:172–182, 2015.

[189] Google Earth Pro Version 7.3.2 (December 14, 2012). Europe. SIO, NOAA, U.S. Navy, NGA, GEBCO. Landsat/Copernicus. IBCAO. Accessed: 2018-08-20.

[190] M. Schneider, R. Müller, T. Krauss, P. Reinartz, B. Hörsch, and S. Schmuck. Urban Atlas - DLR Processing Chain for Orthorectification of PRISM and AVNIR-2 Images and TerraSAR-X as possible GCP Source. In *Internet Proceedings: 3rd ALOS PI Symposium, Kona, Hawaii, USA*, pages 1–6, Nov 2009.

[191] C. Deledalle, L. Denis, and F. Tupin. Iterative Weighted Maximum Likelihood Denoising with Probabilistic Patch-Based Weights. *IEEE Transactions on Image Processing*, 18(12):2661–2672, 2009.

[192] A. Buades and B. Coll. A Non-Local Algorithm for Image Denoising. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA*, pages 60–65, 2005.

[193] J. Muscat. *Functional Analysis: An Introduction to Metric Spaces, Hilbert Spaces, and Banach Algebras*. Springer International Publishing, 2014.

[194] Y. Wang, Q. Yu, and W. Yu. An Improved Normalized Cross Correlation Algorithm for SAR Image Registration, Munich, Germany. In *IEEE International Geoscience and Remote Sensing Symposium*, pages 2086–2089, Munich, Germany, July 2012.

[195] P. Schwind and P. d'Angelo. Evaluating the Applicability of BRISK for the Geometric Registration of Remote Sensing Images. *Remote Sensing Letters*, 6(9):677–686, 2015.

[196] W. Burger and M. Burge. *Principles of Digital Image Processing: Core Algorithms*. Springer Publishing Company, Incorporated, 2009.

# Acknowledgements

*"It doesn't matter where you're going, it's who you have beside you."*

– Author Unknown

Over the last few years, many people have helped me to grow, both academically and professionally. Without their support, this work would not have been possible.

First of all, I am deeply grateful to my supervisor Rupert Müller. Without his foresight to look into the field of Deep Learning and his enthusiasm for even the smallest succeeded experiment, I wouldn't have continued my researches nor finished this thesis.

Second, I am very grateful to Prof. Dr. Reinartz, for giving me the opportunity to work at DLR, carrying out this thesis at the University of Osnabrück and supporting me during the entire time. I would like to thank Dr. Auer, for proof-reading all my papers and providing great advices and helpful discussions. I would also like to thank Prof. Dr. Hinz for reviewing and evaluating this thesis.

In addition, I am glad that I had the opportunity to visit the Machine Learning Group at the Computer Science Department of the University of Toronto twice and to work with Prof. Dr. Urtasun and Wenjie Luo. I had a truly rewarding stay and had the luck to meet an amazing group of people. In particular, Kaustav, Lluis, Eleni, Jamie, Namdar, Andreaa, Emil and Hanna, who supported my during this rough, but important time of my live and made me feel at home.

I was also lucky to work every day around a wonderful group of people at DLR and I am thankful to all my colleagues. Special thanks to Peter and Matthias for helping me from the beginning by providing data, code and helpfully advices, Daniele and Tobias for proof-reading papers and this thesis, Corentin and Majid for all their efforts enabling the writing of our paper and, Jiaojiao, Peter, Olli, Franz, Wei, Xiangyu, and Janja for helpful discussions, listening to my ideas and complaints, supporting me and becoming friends.

I am more than grateful for meeting Chara and Victor, who supported me during difficult times, welcomed my into their family and always reminded me that live has so much more to offer than just work. Ευχαριστώ για όλα!

Last but not least, I would like to express my deepest gratitude to my family. My parents Karin and Fritz, who supported, encouraged and covered my back my entire live. My siblings and partners in crime since day one, Leonie and Nils, who always stand by my side. My latest family members Romina and Fritz, who completed my brothers live and, Karen and Hans, who suddenly stepped into our lives and welcomed us with open arms into their own family. Thank you for taking me for who I am!

*Nina Merkle*

*Munich, Germany, September 2018*