
# Building Information Extraction and Refinement from VHR Satellite Imagery using Deep Learning Techniques

Dissertation

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

Vorgelegt von

*Ksenia Bittner*

Prüfer der Dissertation:
Prof. Dr. Peter Reinartz, Universität Osnabrück
Prof. Dr. techn. Friedrich Fraundorfer, Technische Universität Graz, Österreich

Tag der mündlichen Prüfung: 29. November, 2019

# Abstract

Building information extraction and reconstruction from satellite images is an essential task for many applications related to 3D city modeling, planning, disaster management, navigation, and decision-making. Building information can be obtained and interpreted from several data, like terrestrial measurements, airplane surveys, and space-borne imagery. However, the latter acquisition method outperforms the others in terms of cost and worldwide coverage: Space-borne platforms can provide imagery of remote places, which are inaccessible to other missions, at any time. Because the manual interpretation of high-resolution satellite image is tedious and time consuming, its automatic analysis continues to be an intense field of research. At times however, it is difficult to understand complex scenes with dense placement of buildings, where parts of buildings may be occluded by vegetation or other surrounding constructions, making their extraction or reconstruction even more difficult. Incorporation of several data sources representing different modalities may facilitate the problem. The goal of this dissertation is to integrate multiple high-resolution remote sensing data sources for automatic satellite imagery interpretation with emphasis on building information extraction and refinement, which challenges are addressed in the following:

*Building footprint extraction* from *Very High-Resolution (VHR)* satellite images is an important but highly challenging task, due to the large diversity of building appearances and relatively low spatial resolution of satellite data compared to airborne data. Many algorithms are built on spectral-based or appearance-based criteria from single or fused data sources, to perform the building footprint extraction. The input features for these algorithms are usually manually extracted, which limits their accuracy. Based on the advantages of recently developed *Fully Convolutional Networks (FCNs)*, i.e., the automatic extraction of relevant features and dense classification of images, an *end-to-end* framework is proposed which effectively combines the spectral and height information from *red, green, and blue (RGB)*, *pan-chromatic (PAN)*, and *normalized Digital Surface Model (nDSM)* image data and automatically generates a full resolution binary building mask. The proposed architecture consists of three parallel networks merged at a late stage, which helps in propagating fine detailed information from earlier layers to higher levels, in order to produce an output with high-quality building outlines. The performance of the model is examined on new unseen data to demonstrate its generalization

capacity.

The availability of *detailed Digital Surface Models (DSMs)* generated by dense matching and representing the elevation surface of the Earth can improve the analysis and interpretation of complex urban scenarios. The generation of DSMs from VHR optical stereo satellite imagery leads to high-resolution DSMs which often suffer from mismatches, missing values, or blunders, resulting in coarse building shape representation. To overcome these problems, a methodology based on *conditional Generative Adversarial Network (cGAN)* is developed for generating a good-quality *Level of Detail (LoD)* 2 like DSM with enhanced 3D object shapes directly from the low-quality photogrammetric half-meter resolution satellite DSM input.

Various deep learning applications benefit from *multi-task learning* with multiple regression and classification objectives by taking advantage of the similarities between individual tasks. Therefore, an observation of such influences for important remote sensing applications such as realistic elevation model generation and roof type classification from stereo half-meter resolution satellite DSMs, is demonstrated in this work. Recently published deep learning architectures for both tasks are investigated and a new end-to-end cGAN-based network is developed, which combines different models that provide the best results for their individual tasks.

To benefit from information provided by *multiple data sources*, a different cGAN-based work-flow is proposed where the generative part consists of two encoders and a common decoder which blends the intensity and height information within one network for the DSM refinement task. The inputs to the introduced network are single-channel photogrammetric DSMs with continuous values and pan-chromatic half-meter resolution satellite images. Information fusion from different modalities helps in propagating fine details, completes inaccurate or missing 3D information about building forms, and improves the building boundaries, making them more rectilinear.

Lastly, additional comparison between the proposed methodologies for DSM enhancements is made to discuss and verify the most beneficial work-flow and applicability of the resulting DSMs for different remote sensing approaches.

# Zusammenfassung

Die Extraktion von Gebäudeinformationen und die Rekonstruktion von Gebäuden aus Bilddaten ist ein zentraler Arbeitsschritt zahlreicher Anwendungen im Bereich der 3D Stadtmodellierung und -planung, Katastrophenmanagement, Navigation und Entscheidungsfindung. Diese Gebäudeinformationen können aus Daten unterschiedlicher Art gewonnen werden, darunter Landvermessungsdaten, Luftbilder oder hochaufgelöste (VHR) satellitengestützte Erdbeobachtung. Letztere übertrifft jedoch die anderen Methoden der Datenerhebung in Bezug auf Kosten und weltweite Flächenabdeckung: Satellitenplattformen können zu jeder Zeit Daten auch an solchen Orten erheben, welche für andere Methoden unzugänglich sind. Allerdings ist die manuelle Interpretation von Satellitenbildern mühsam und langwierig. Deshalb ist und bleibt die automatisierte Verarbeitung von hochauflösenden Satellitenbildern ein intensives Forschungsthema. Für Algorithmen ist es jedoch schwierig, komplexe Szenen mit dichter Bebauung zu interpretieren, da Teile der Gebäude aufgrund von Vegetation oder eines flachen Aufnahmewinkels verdeckt sein können, was die Extraktion oder Rekonstruktion sogar noch erschwert. Dieses Problem kann durch die Integration mehrerer Datenquellen unterschiedlicher Modalitäten gelöst werden.

Das Ziel der vorliegenden Arbeit ist somit mehrere hochaufgelöste Fernerkundungsdatenquellen für eine voll automatisierte Satellitenbildanalyse zu integrieren. Der Fokus liegt dabei auf der Extraktion von Gebäudeinformationen und deren Verbesserung hinsichtlich der Bildqualität. Dabei werden zwei Problemstellungen erforscht, die im Folgenden adressiert werden: Die Extraktion von Gebäudegrundflächen aus sehr hochauflösenden Satellitenbildern ist ein wichtiger Prozessschritt, der aber aufgrund der äußerst hohen Diversität der Gebäudeformen, sowie der—im Vergleich zu Luftaufnahmen—relativ niedrigen räumlichen Bildauflösung sehr anspruchsvoll ist. Die Bildmerkmale ("features"), die für diese Algorithmen nötig sind, werden für gewöhnlich manuell erfasst, was ihre Genauigkeit limitiert. Basierend auf den Vorteilen, welche der kürzlich entwickelte Ansatz der Fully Convolutional Networks (FCN) bietet—nämlich der automatischen Extraktion relevanter Merkmale und der dichten Bildklassifikation—wird hier ein "end-to-end framework" vorgeschlagen, welches Spektral- und Höheninformationen aus rot-grün-blau (RGB) und panchromatischen (PAN) Bildern, sowie normalisierten digitalen Oberflächenmodellen (nDSM) effizient kombiniert und daraus automatisiert eine vol-

lauflösende binäre Gebäudemaske errechnet. Die vorgeschlagene Architektur besteht aus drei parallelen Netzwerken, welche zu einem späteren Prozesszeitpunkt fusioniert werden. Dies hat den Vorteil, dass feingranulare Informationen von früheren Verarbeitungsschritten zu höheren Ebenen des Netzwerks propagiert werden, sodass der Output in hochqualitativen Gebäudeumrandungen besteht. Die Leistungsfähigkeit dieses Modells wird anhand von neu erhobenen, nie zuvor prozessierten Daten untersucht, um die Generalisierbarkeit bzw. Übertragbarkeit des Modells auf weitere Anwendungen (z.B. andere Städte) zudemonstrieren.

Die steigende Verfügbarkeit detaillierter digitaler Oberflächenmodelle (DSM), welche mittels Dense Matching und der Representation von Höhenprofilen auf der Erdoberfläche gewonnen werden, kann die Analyse und Interpretation komplexer urbaner Szenarien verbessern. DSMs, welche aus optischen stereo VHR-Satellitenbildern gewonnen werden, sind zwar hochauflösend, beinhalten aber Höhen-Diskrepanzen und fehlende oder sogar falsche Fragmente, was bei der Gebäuderekonstruktion in einer groben bzw. falscgen Gebäudeform resultiert. Um dieses Problem zu lösen wurde im Rahmen dieser Arbeit ein auf einem Conditional Generative Adversarial Network (cGAN) basierender Ansatz entwickelt, welcher aus einem photogrammetrischen halb-Meter aufgelösten DSM schlechter Qualität direkt ein hochqualitatives DSM entsprechend eines Level of Details (LoD) 2 mit erweiterten 3D Objektformen produziert. Diverse Deep-Learning-Anwendungen können von dem hier präsentierten Multi-Task Learning mit multipler Regression und Klassifikation profitieren, indem die Lösungsansätze der einzelnen Prozesschritte auf ähnliche Teilaufgaben übertragen werden. Um die Vorzüge dieses Ansatzes zu bewerten wird die Wirkungsweise bei wichtigen Fernerkundungsanwendungen, wie z.B. der realitischen Schätzung von Höhenprofilmodellen, und der Dachtypenklassifikation aus stereo halb-Meter-aufgelösten Satelliten-DSMs demonstriert. Es werden kürzlich veröffentlichte Deep-Learning-Architekturen für beide Anwendungen geprüft, und ein neues “end-to-end” cGAN-basiertes Netzwerk entwickelt, welches verschiedene Modelle in einer Weise kombiniert, in der sie zunächst separat genutzt werden, um möglichst spät im Rechenprozess die bestmöglichen Ergebnisse der individuellen Aufgaben zu kombinieren. Um die Vorzüge der Verfügbarkeit von Informationen aus multiplen Datenquellen zu nutzen, wird zudem ein weiterer cGAN-basierter Prozess vorgeschlagen, dessen generativer Anteil aus zwei Encodern und einem gemeinsamen Decoder besteht, wobei die Intensitäts- und Höheninformationen aus den Bilddaten innerhalb eines Netzwerks kombiniert werden, zum Zwecke der Verfeinerung des DSMs. Als Inputs für dieses vorgeschlagene Netzwerk werden photogrammetrische Einzelkanal-DSMs mit kontinuierlichen Werten, sowie halb-Meter aufgelöste panchromatische Einzelkanal-Satellitenbilder benutzt. Die Fusion der Informationen aus verschiedenen Modalitäten hilft dabei, feine Details, vollständig falsche oder fehlende 3D information über Gebäudeformen zu propagieren, und verbessert die Gebäudeumrisse, indem sie sie rechtwinkliger macht. Zuletzt werden die vorgeschlagenen Lösungsmethoden zur DSM-Verbesserung miteinander verglichen, um den nützlichsten Ansatz zu identifizieren und zu verifizieren und um die Anwendbarkeit der resultierenden DSMs für verschiedene Fernerkundungsaufgaben zu beurteilen.

# Acknowledgments

I would like to express my deepest gratitude to all people, whose support and encouragement made this work possible:

First of all, I would like to emphasize my appreciation to Prof. Dr. Peter Reinartz for giving me the opportunity to join his Department of Photogrammetry and Image Analysis at German Aerospace Center (DLR) and be able to carry out this dissertation at the University of Osnabrück. I would like to thank him for introducing me to the interesting remote sensing problems and supporting me with many discussions, valuable suggestions and constructive comments which helped me to reach the final goal. I feel very grateful that he was always taking time out of his busy days and being there when I needed his help and advice. Moreover, I would like to thank Prof. Reinartz for the possibility to attend different international workshops and conferences, and for the freedom to participate in various research projects.

I would like to express my sincerest thanks to Prof. Dr. Friedrich Fraundorfer from Technical University of Graz for showing interest in my work and giving me the opportunity to visit his computer vision research group for two months where I had the luck to meet an amazing group of people. Under guidance of Prof. Fraundorfer I was able to approach my topic differently and develop an extended version of my methodology which helped me further to improve results. His brilliant insights into computer vision problems, knowledge and experience, have been continuously inspiring me. Moreover, I would like to thank him for agreeing to be a co–referee of this dissertation.

I would also like to acknowledge the German Academic Exchange Service (DAAD) for financial support of my PhD study and the German Aerospace Center (DLR) for providing facilities and satellite data for this research.

Further, I would like to thank Dr. Marco Körner and his team for giving me support and encouragement, sharing their thoughts and ideas with me. At many stages of this research I benefited from Dr. Körner's advice and enlightening discussions, particularly when exploring new ideas and methodologies. His positive outlook and confidence in my research gave me a strong motivation. Moreover, his careful editing and proofreading contributed enormously to my publications and this thesis. I also would like to thank Dr. Shiyong Cui for sharing with me his knowledge of machine learning techniques, teaching me programming and being patient responding many of my questions and queries at the

beginning of my PhD.

I cannot find words to express my gratitude to all colleagues at DLR with whom I was lucky to work together every day. Special thanks to Dr. Pablo d'Angelo for the valuable discussions, sharing with me his profound scientific knowledge and experience, providing me a technical support and generating all of the datasets required for this research. I would like to thank Peter Schwind and Dr. Emiliano Carmona for their technical and programming support during my research work. I am very grateful to Thomas Krauß, Dr. Jiaojiao Tian, Maximilian Langheinrich, Dr. Xiangyu Zhuo, Adam Fathalrahman, Zeinab Gharib Bafghi and Dr. Tahmineh Partovi, along with many others for valuable discussions and sharing with me their time, knowledge and experience, listening to my ideas or problems and supporting me continuously.

Completing this work would have been all the more difficult without my best friends who have been always there for me no matter what, and I appreciate it!

Last but not least, I would like to express my deepest gratitude to my family for always believing in me, encouraging and unconditionally supporting me in all my pursuits entire live: my beloved grandparents, Olga and Alexander, who loved me to the Moon and back; my cousin Elena and her husband Christian, who help and support me since the first day of coming to Germany; my uncle Sergey, who accepts me as his own child and guides me with his valuable advises, my dear mother, Irina, who coveres my back my entire live, loves me enough with her compassion, understanding and unconditional acceptance, my beloved husband, Nikolaj, who enriches my every day with love, sharing with me not only happy moments but strongly supporting me in case of difficulties, always having the solution for any problem. I can't thank you all enough for being there for me and constantly loving me.

Ksenia Bittner
Oberpfaffenhofen, December 2019

# Contents

Chapter 1

# Introduction

In the last 10 years, satellite technologies with high-resolution sensors have become an important tool for large-scale Earth observation. They provide essential information by orbiting around the Earth and observing areas of interest for a large diversity of applications. Moreover, satellite images cover much greater areas compared, for example, to aerial mapping. The constant development of space-borne instruments makes it possible to collect digital data at spatial resolutions up to 30 cm which allows detailed terrestrial scene understanding. This is especially important for controlling the impact of urbanization, because a large part of the population is still moving from less developed regions to urban areas due to modernization and industrialization. Consequently, this entails changes in urban spaces and appears as a formation of new building constructions or their destruction. Remote sensing is a valuable tool to monitor these changes for large areas and in a time-efficient manner, because the speed of urban changes has provoked a huge demand of reliable building information extraction, which was not possible two decades ago due to the low-resolution of satellite images. However, the analysis of current commercial high-resolution satellite imagery with very fine detail and applying automatic algorithms instead of slow and expensive manual image interpretation allows the disclosure of individual, industrial, and residential buildings from images. This information is favorable for a vast amount of remote sensing applications, like the update of *Geographic Information System (GIS)* databases, civil engineering, city management, emergency problems handling, 3D city modeling, etc. Although numerous attempts have been made to develop methodologies towards automatic buildings extraction from satellite imagery, this topic is still an open issue for scientists due to low-resolution in comparison to airborne data and the large diversity of building appearances resulting in scene complexity.

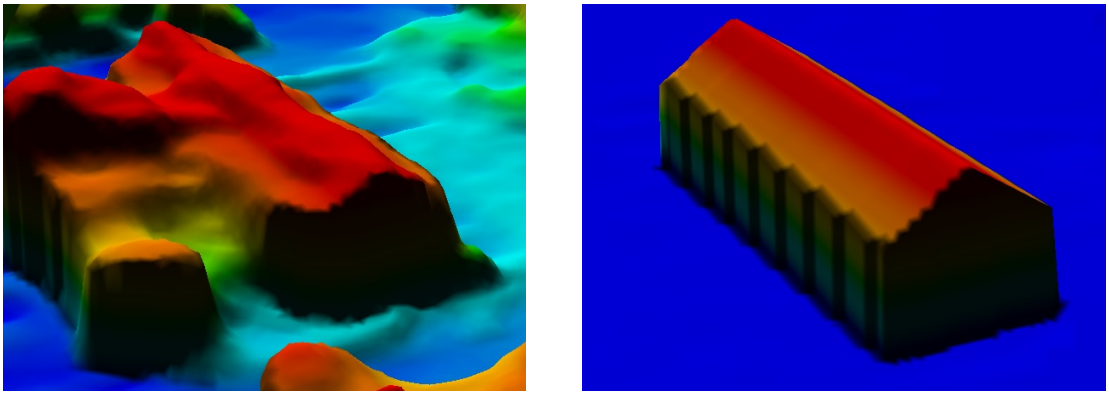
The development of innovative algorithms for building information extraction is also motivated by the type of data which advanced satellite technologies produce. For example, the modern WorldView-2 satellite provides multi-spectral imagery (2 m pixel spacing) allowing objects recognition through their color differences, and pan-chromatic imagery (0.5 m pixel spacing) allowing detailed extraction through much finer informa-

tion embedded in the images. In some cases, however, it is not enough to only use spectral information since roads and building roofs can have similar texture appearance. Moreover, in observing only two-dimensional images, we lose the third dimension—height. The availability of *Digital Surface Models (DSMs)* and their combination with spectral information can be used for solving these problems and as a result, improve image interpretation. This is especially important for building extraction and reconstruction applications, because geometrical information is crucial for 3D object discrimination.

In previous studies, buildings are delineated utilizing *pan-chromatic (PAN)*, multi-spectral images, height information from DSMs, or even the combination of spectral and height data. However, the majority of approaches showed the potential for extracting buildings with similar color, size, or simple shapes, i.e., rectangular or square. In this dissertation, buildings with arbitrary shapes and sizes are automatically delineated, taking advantage of spatial, spectral, and height information from multi-view stereo satellite data sources.

A DSM represents the elevation of the Earth surface including topography and objects above the bare Earth, be it natural objects, e.g., trees, or man made activities, e.g., buildings. Because DSMs provide the geometry and structure of an urban environment with buildings being the most prominent objects in it, they can increase the understanding and explanation of complex urban scenarios. DSMs can be produced from a variety of source data. Their production, however, is most common from multi-view stereo satellite imagery, e.g., WorldView or Pleiades satellites, because these satellites provide global coverage and exhibit sub-meter resolution. Moreover, DSMs from satellite platforms are less expensive and are not subject to bad weather conditions like *Light Detection and Ranging (LiDAR)* surveys from aerial platforms. However, despite the available techniques capable of generating large-scale elevation models from high-resolution satellite images, DSMs still feature many mismatches and noise in forms of blunders or spikes due to occlusions by dense and complex building structures, perspective differences, or stereo matching errors during their generation [1, 2]. This introduces difficulties for building extraction and reconstruction. Hence, the development of methodologies able to improve photogrammetric DSMs (see example in Figure 1.1a) automatically to a higher quality level (see example in Figure 1.1b), e.g., close to *Level of Detail (LoD)* 2 regarding the *City Geography Markup Language (CityGML)* standard, with more realistic and complete building geometries is in demand, since earlier approaches which mainly utilized filtering and interpolation strategies, have not been very successful and affect the steepness and details of raised objects.

With recent advances in the field of artificial neural networks, it is possible to learn image features automatically instead of extracting them by classical methods, e.g., hand crafted feature extraction. Innovative architectures, such as *Convolutional Neural Networks (CNNs)*, have demonstrated the ability to classify high-dimensional data sources accurately and robustly and have become the state-of-the-art for image recognition tasks. Considering these new advantages in the field of image processing, this dissertation aims to develop novel deep learning based algorithms for extracting building footprints from *Very High-Resolution (VHR)* satellite photography as well as improving building ge-



**(a)** Photogrammetric DSM (input) **(b)** Desired building shape (output)

**Figure 1.1:** Illustration of the building shape refinement we aim to achieve in this dissertation. (a) demonstrates the original low-quality photogrammetric DSM we want to refine, (b) shows the desired enhancement regarding building shape in DSM which is generated from CityGML data.

ometries in photogrammetrically generated DSMs. Moreover, we aim to investigate if the inclusion of auxiliary knowledge in form of spectral, height, segmentation or categorization, e.g., roof type, information into the algorithmic procedures can improve the estimation results of both tasks.

## 1.1 Scope of the Dissertation

This thesis is primarily focused on the development of new methodologies for detailed and robust terrestrial scene interpretation from VHR satellite imagery regarding building objects. We accomplished this goal by tackling two specific objectives:

- **Objective 1: Building Footprint Extraction**

  The terrestrial scene understanding includes the identification of various natural or man-made targets in an image. Buildings are one of the most difficult, but also one of the most significant objects among them to extract, since they are the most pronounced elements of urban organization. Satellite data provide comprehensive information, applicable for building footprint extraction. Modern satellite platforms capture images from multiple sensors. To improve the accuracy of extracted building footprints, information from multiple sensors, e.g., pan-chromatic and multi-spectral, can be combined to compensate the limitations of each independent sensor. However, not only spectral knowledge can be merged. The height information derived from multi-view stereo satellite imagery is also very useful for improving building outlines extraction, because it provides detailed information about the building geometry. Extracting spectral and height related features for building hypothesis generation is a complex task if done manually or with traditional machine learning approaches. A new potential solution is automatic

extraction of relevant features using deep learning techniques which demonstrated their superiority over traditional ones. In order to produce improved building footprints, we aim to develop an end-to-end neural network for remote sensing imagery understanding which benefits from information combined from different modalities.

- **Objective 2: Space-borne Photogrammetric Digital Surface Model Refinement**

  Height information embedded in DSMs can greatly contribute to the scene interpretation, since it provides knowledge about object silhouettes—a crucial clue for automatic building extraction and reconstruction tasks. Although the DSM can be generated from different data sources, in this thesis we focus on space-borne platforms because these data offer large area coverage around the world and are especially challenging. However, DSMs from multi-view satellite images have some limitations. For example, due to the presence of vegetation building constructions are often hidden under the tree crown and appear incomplete in the DSM. The fronto-parallel plane assumption in the *Semi-Global Matching (SGM)* algorithm affects inclined roofs, like hip and gable, by reconstructing them as piece-wise horizontal planes which are far from realistic building appearances. Moreover, the dense matching algorithm can cause errors in the presence of homogeneous or low-textured areas. All these problems can influence the performance of approaches which utilize photogrammetric DSMs. Therefore the quality enhancement of photogrammetric DSMs with emphasis on building shapes is in demand. Previously developed approaches did not show significant improvement, resulting only in over-smoothed DSMs or being constrained to a few building forms. In order to automatically improve DSMs to a high level of accuracy, deep learning can be applied, as it has already shown promising results towards depth image reconstruction task. We also aim to positively influence building boundaries as well as roof plane reconstruction by introducing auxiliary knowledge via pixel-wise roof type classification masks or intensity information from PAN images additionally incorporated into the learning process. This additional information can make the network more confident when performing the detailed DSMs refinement.

## 1.2 Guidelines for Reading

This is a *cumulative* dissertation which is organized as follows. Chapter 2 gives a brief introduction to the fundamental knowledge related to this thesis. Chapter 3 summarizes the state-of-the-art as well as the contributions of this dissertation related to the aforementioned objectives. Chapter 4 defines problems related to building footprint extraction task and describes the developed methodology to tackle them. Moreover, the experimental results for two different cities applying the proposed deep learning approach together with qualitative and quantitative evaluations are shown and discussed in this chapter. The chapter represents a peer-reviewed journal paper

- K. Bittner, F. Adam, S. Cui, M. Körner, and P. Reinartz, "Building Footprint Extraction from VHR Remote Sensing Images Combined with Normalized DSMs using Fused Fully Convolutional Networks," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 8, pp. 2615–2629, 2018 [3].

Chapter 5 introduces open questions to our second objective related to DSM refinements. It states three possible approaches to solve the problem. Furthermore, experimental results of three proposed methodologies together with qualitative and quantitative analysis are presented in this chapter. An additional comparison between three proposed methodologies is also followed by detailed discussions. Chapter 5 is based on the combined findings of the following two peer-reviewed journal papers:

- K. Bittner, P. d'Angelo, M. Körner, and P. Reinartz, "DSM-to-LoD2: Space-borne Stereo Digital Surface Model Refinement," *Remote Sensing*, vol. 10, no. 12, p. 1926, 2018 [4]
- K. Bittner, M. Körner, F. Fraundorfer, and P. Reinartz, "Multi-Task cGAN for Simultaneous Space-borne DSM Refinement and Roof-Type Classification," *Remote Sensing*, vol. 11, no. 11, p. 1262, 2019 [5]

and one double-blind peer-reviewed workshop paper:

- K. Bittner, M. Körner, and P. Reinartz, "Late or Earlier Information Fusion from Depth and Spectral Data? Large-Scale Digital Surface Model Refinement by Hybrid-cGAN," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019 [6].

Chapter 6 concludes the dissertation and gives an outlook on future work.

# Chapter 2

# Background

*"Man must rise above the Earth—to the top of the atmosphere and beyond—for only thus will he fully understand the world in which he lives."*
— Socrates, Philosopher

## 2.1 High-Resolution Satellite Imagery

We live on a planet which changes very rapidly, and there is no confidence about the changes that are coming in the future. It is beneficial to gain a new vision, and prove all facts, before reaching any critical decision. To monitor changes globally, we need a bird's-eye view of our planet from above. Satellite imagery can help us to achieve this goal. In this technology-driven era, our planet is continually being observed and imaged by satellites providing long-term global observations of the land surface, oceans, biosphere, and atmosphere in very high spatial, spectral, and temporal resolution and with large-area coverage. As a result, the aforementioned advantages make a satellite a valuable instrument which can collect more data more quickly than instruments on the ground or by using airplanes and helicopters.

### 2.1.1 Active Imagery

Remote sensing instruments collect light energy within specific regions of the electromagnetic spectrum, and comprises the range of all types of electromagnetic radiation covering visible light, radio waves, microwaves, infrared light, ultraviolet light, X-rays and gamma-rays as depicted in Figure 2.1. Two types of remote sensing equipments can be distinguished: *active* and *passive*. Active sensors, e.g., *Synthetic Aperture Radar (SAR)* or *Light Detection and Ranging (LiDAR)*, provide their own energy source for target illumination (see Figure 2.2a). Regular pulses of energy are emitted at a known electromagnetic wave towards the surface of the Earth to be investigated. The radiation reflected, refracted, or scattered by the Earth's surface or its atmosphere back to

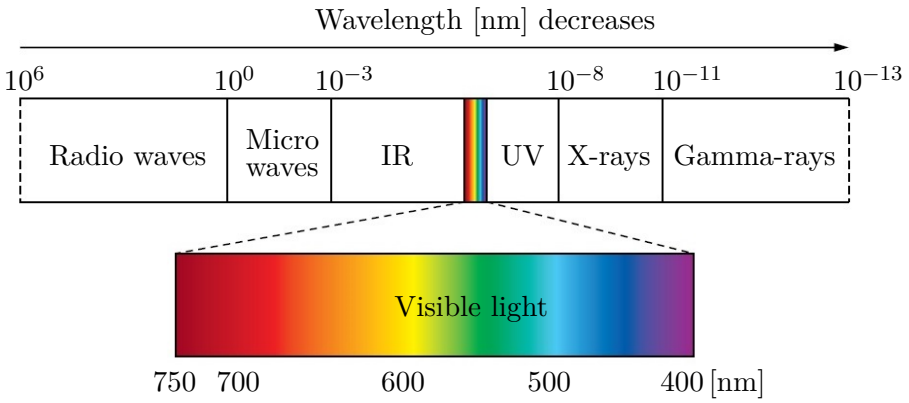



**Figure 2.1:** The electromagnetic spectrum illustration from the longest wavelength (at the left) to the shortest wavelength (at the right).

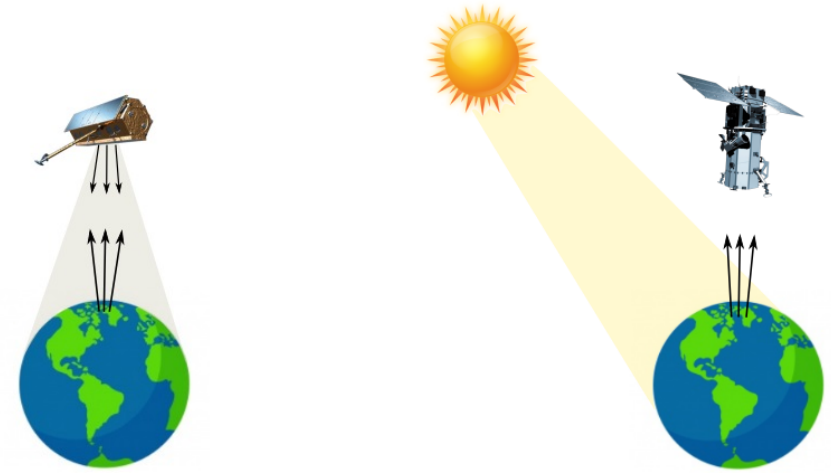
the satellite is detected and measured by the sensor. Radar imaging systems, such as *European Remote Sensing Satellite (ERS)*, TanDEM-X, TerraSAR-X, and *Japan Earth Resources Satellite (JERS)*, are examples of active sensors. The advantage of using microwaves is that they are able to obtain measurements at any moment, without any limitations from the time of day or the season. They can even penetrate clouds. However, to adequately illuminate targets on the Earth surface, active sensor instruments require a large amount of energy. This type of data also cannot differentiate the visible color of objects which is crucial for many remote sensing applications.

### 2.1.2 Passive Imagery

On the other hand, passive instruments detect natural energy in the wavelength range they are capable of that either emitted or reflected back to space from the observed area (see Figure 2.2b). The sunlight reflected by objects on the ground is the most usual external radiation source sensed by passive instruments. By analyzing the strength of reflection, a land surface can be understood, e.g., urban areas or farm fields, forests, rivers and lakes, or distribution of plants. Passive instruments can only be utilized to detect reflection when energy of natural origin is available. This can only occur while the Sun illuminates the Earth. Weather conditions also influence data acquisition. For instance, optical sensors cannot observe areas under clouds. The examples of passive instruments that measure only sunlight radiation and reflected or emitted radiation by the Earth are WorldView, Landsat, Sentinel-2 and Pleiades satellites.

#### 2.1.2.1 Pan-chromatic Band

Earth observation sensors can capture data in single or multiple bands producing *panchromatic (PAN)* and multi-spectral images, respectively. The PAN image is the result of the measure of light energy in the full visible, and often partially the *near-infrared*



(a) (b)

**Figure 2.2:** Schematic illustration of (a) active and (b) passive remote sensing instruments in action.

*(NIR)* spectrum. This measurement would typically cover wavelengths between 0.47 µm and 0.83 µm, like for WorldView-2 satellite. The broad bandwidth allows to maintain a high signal-to-noise ratio, resulting in the PAN data with a high spatial resolution. As the PAN band combines the information from this spectral range, it returns a single intensity value per pixel that is often displayed in a gray-scale image (see Figure 2.3a).

#### 2.1.2.2 Multi-Spectral Band

With more spectral bands, the information gathered by the sensor increases. Most commercial Earth observation satellites generate multi-spectral images covering several, and commonly narrow bands, over the visible and infrared portions of the electromagnetic spectrum. This means that each multi-spectral image layer is gathered at a specific wavelength band. Therefore, the range of wavelengths contributing to the radiation energy collected by the sensor is reduced and multi-spectral instruments typically have to acquire energy on a larger spatial extent to enhance the signal-to-noise ratio. This leads to a lower spatial resolution opposite to PAN images. Typically, multi-spectral sensors provide a discrete and limited number of bands, generally under 15, although on some sensors, such as *hyperspectral*, it can contain more than 100 spectral bands.

A common example of multi-spectral images is the production of "natural color" photographs which combine three bands of the visible spectrum: blue, green and red (see Figure 2.3b). However, current remote sensing multi-spectral cameras are not only restricted to the visible spectrum. They can detect different infrared spectral ranges, e.g., NIR, shortwave, *infrared (IR)*, thermal IR or even the ultraviolet ranges. For example, the image which couples the green, red, and near infrared bands is a valuable data source for remote sensing applications related to the land, forest and agriculture monitoring because it highlights the presence and health of the vegetation: healthy vegetation absorbs

**(a)** PAN **(b)** RGB **(c)** Pan-sharpened

**Figure 2.3:** An example of (a) PAN image with *Ground Sampling Distance (GSD)* of 0.5 m, (b) *red, green, and blue (RGB)* image with GSD of 2 m, and (c) pan-sharpened image with GSD of 0.5 m depicting a central cathedral in the city of Munich, Germany. Images are acquired with WorldView-2 satellite. The pan-sharpened image is generated by combining the high-resolution pan-chromatic image with low-resolution RGB image using a pan-sharpening technique developed by Krauß *et al.* [7].

blue- and red-light energy to fuel photosynthesis and creates chlorophyll which reflects strongly in the NIR spectrum, and therefore appears in darker red in the image. Numerous other combinations of wavelength bands are possible, depending on the information to be extracted for the Earth's surface investigation (soil, water, geological formations, etc.).

#### 2.1.2.3 Pan-Sharpening

Current commercial satellites, like Pleiades or WorldView, commonly include both lower-resolution multi-spectral bands and a single higher-resolution pan-chromatic band. The reason to configure the satellites in this way is to keep their weight, cost, bandwidth and data rate lower. The desire to have space-borne images with both high spatial and spectral resolution motivates researchers to develop pan-sharpening techniques that merge high-resolution pan-chromatic data with medium-resolution multi-spectral data to create a multi-spectral image with higher-resolution features. An example of pan-sharpened image is illustrated in Figure 2.3c. Obtaining high spatial and at the same time spectral resolution data, can be crucial in many remote sensing applications.

### 2.1.3 Digital Surface Models

Traditional satellite image provides valuable 2D information from which the physical state of the Earth surface can be analyzed. However, the world is three-dimensional and two-dimensional imagery only represents a part of the reality. Elevation models can measure and analyze the real world in 3D for a more complete picture. The digital format of a height model is given by a *Digital Surface Model (DSM)*, consisting of elevation data and representing the Earth surface as seen from an orthogonal aerial viewpoint. The

DSM is a 2.5D height representation because it considers the maximum height over ground rather than looking at all objects that might be below.

Remote sensing technologies provide several approaches to measure the surface topography. Most common data sources are ground survey, stereo airborne, or space-borne photogrammetry, satellite *Interferometric Synthetic Aperture Radar (InSAR)*, and aerial LiDAR.

#### 2.1.3.1 Ground survey

Ground survey is used to collect detailed topographic and land cover characteristic data for smaller target areas such as city districts, dumping area, reservoir site, etc. It involves *Global Positioning System (GPS)* observations, total stations and laser ranging to gather *X,Y,Z* coordinates of terrestrial points in a gridded way for DSM generation. Although this technique gives detailed elevation models, it has limitations consisting of high cost, narrow area coverage, and is highly time consuming because it requires extensive labor involvement [8, 9].

#### 2.1.3.2 Light Detection and Ranging Technology

The LiDAR technology utilizes the NIR band of the electromagnetic spectrum and measures the time it takes for the emitted laser pulse to travel from the transmitter to the object on the Earth's surface and back to the receiver. Because the speed of light is known, the distance can be calculated. The LiDAR technique has an up to centimeter level accuracy as well as a regular scan pattern with high density. This allows the production of a high-quality DSMs. The pulse laser light transmitted by the LiDAR is also able to partially penetrate vegetation and also acquires ground points through the vegetation cover. However, LiDAR data acquisition is expensive and does not provide global coverage compared to satellites. Furthermore, LiDAR data only provide the building or roof top sparse depth information. It influences the precise identification and measure, as the exact corners of building construction usually do not contain collected points [10].

#### 2.1.3.3 Interferometric Synthetic Aperture Radar

The DSM generation can be obtained through radar interferometry, using the difference in phase between at least two complex SAR images. The interferometry image pair is either taken sequentially via repeat-pass mode, or simultaneously via single-pass mode by airborne or space-borne sensors from slightly different view angles focused on the same area of interest [11, 12]. SAR is a cost-effective technology which can perform wide coverage measurements with ground resolution up to 1 m under all weather conditions and at day and night times [13]. However, densely settled areas, steep topography, shadow regions, or vegetation presence are hard to handle in InSAR processing, as they lead to limited visibility of objects due to occlusion issues and as a result, produce erroneous or missing height estimates [14]. Moreover, because of the side-looking sensor

principle, InSAR technology is not as useful for some remote sensing applications, such as building recognition and reconstruction.

#### 2.1.3.4 Multi-View Stereo Photogrammetry

Height information extraction is possible from either airborne or space-borne multi-view stereo photogrammetry by using correlations between every pixel of overlapping images over an area. Multi-view stereo photogrammetry contributes to the scene completeness by minimizing the occlusion between landscapes and by allowing a multitude of matching points. Airborne platforms provide a good performance with respect to spatial and spectral resolution. The resulting images are not affected as much by atmospheric perturbations. Airborne platforms are very flexible in operation and all acquired data is available to the customer as soon as the survey mission is over [15]. However, turbulences, clouds, and air traffic control, as well as the high costs, limit the efficient use of airborne platforms. Furthermore, they cannot be used during volcanic eruptions or hurricanes because it puts the crew in danger. Also, airborne survey missions are usually planned as one-time operations compared to the space-borne platforms which continuously observe the Earth. A good overview of existing airborne stereo systems is given by Zhang [16].

In recent years, the techniques of stereo imagery acquisition from space-borne platforms have rapidly evolved, allowing researchers to obtain DSMs of improved quality and resolution. The data is very useful for many applications, as it provides regular surveys of the same area, e.g., for security surveys or when monitoring natural disasters [16]. When compared to airborne platforms, they are cost-effective and provide a much wider area of coverage. Multi-view stereo data can be collected by across-track stereoscopy from two or more adjacent orbits in different days. However, the quality of DSMs generated from across-track stereo imagery is influenced by temporal changes of the surface reflection due to the different acquisition dates. The recently employed same-date along-track stereo data acquisition from the same orbit using multiple cameras or by adjusting the viewing angles overcomes this problem. It reduces the radiometric image variations allowing higher quality 3D reconstructions than previously possible. An example of such a DSM is illustrated in Figure 2.4.

Satellite imagery products vary in their spectral and spatial resolution, geographic and temporal coverage, cloud cover, security regulations, and price [18–20]—characteristics and operation mechanisms which can limit specific Earth observation applications. Technological progress and increasing accessibility of high-resolution satellite imagery expand traditional application domains in remote sensing. Examples of modern, submeter-resolution commercial space-borne platforms are the GeoEye-1, WorldView 1–4 and Pléiades 1A/1B satellites. They represent the most important very high-resolution satellites with stereo acquisition capabilities.

- **GeoEye-1**

GeoEye-1 is an American very high-resolution satellite that was launched on September 6, 2008. It collects 0.41 m resolution PAN and 1.64 m resolution multi-spectral

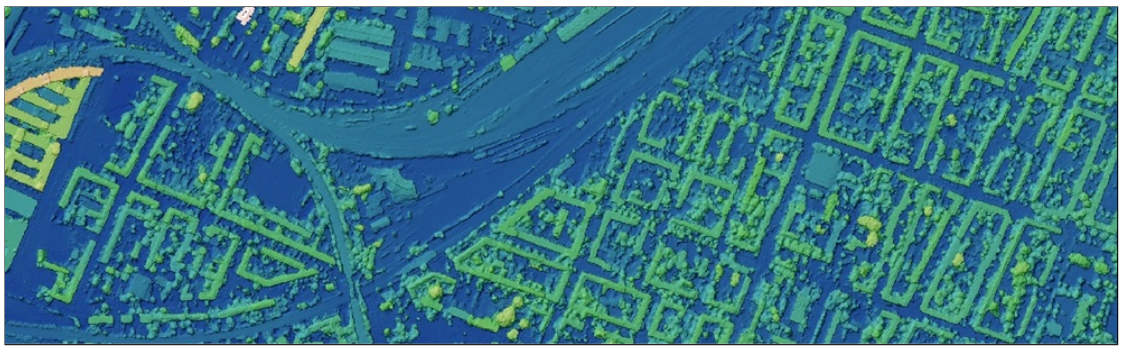


**Figure 2.4:** An example of optical DSM with a resolution of 0.5 m produced from six pan-chromatic Worldview-1 images using *Semi-Global Matching (SGM)* [17] method. The depicted area is located in Berlin, Germany and represents a 1.5 $km^2$ coverage. The image is color-shaded for better visualization.

images providing 15.3 km swath at nadir. The multi-spectral images consist of four color bands including blue, green, red, and near-infrared channels. The satellite can be rotated forward, backward or sidewards with high-precision and has a revisit capability of one to three days, depending on latitude. This sensor is meant to be utilized for large projects, as it can collect over 350 000 $km^2$ of pan-sharpened multi-spectral satellite imagery per day. The description of the GeoEye-1 satellite is taken from the Satellite Imaging Corporation web page [21].

- **WorldView 1-4**

WorldView-1 is a commercial Earth observation satellite that has provided high-resolution pan-chromatic multi-view stereo imagery collection since its launch on September 18, 2007. The satellite has revisit time of 1.7 days on average and is capable of covering up to 750 000 $km^2$ per day of half-meter resolution PAN imagery.

WorldView-2 is another commercial Earth observation satellite operated by DigitalGlobe that provides a 0.5 m pan-chromatic images as well as 2 m multi-spectral images with four standard colors (red, green, blue, and near-infrared 1) and four additional bands (coastal, yellow, red edge, and near-infrared 2). The satellite was launched on October 8, 2009. WorldView-2 has a revisit time of 1.1 days and is able to collect nearly 1 million $km^2$ every day of multi-view stereo imagery in a single pass by adjusting the camera viewing angle.

WorldView-3 is the first multi-payload, super-spectral, high-resolution commercial satellite sensor. It was launched on August 13, 2014. Along with 0.31 m pan-chromatic and 1.24 m multi-spectral bands, the WorldView-3 sensor collects eight-band 3.7 m resolution *short wave infrared (SWIR)* and *Clouds, Aerosols, Water Vapor, Ice and Snow (CAVIS)* data at 30 m resolution. The CAVIS offers standardized data independent of where or when the images have been acquired by correcting the inconsistencies caused by certain conditions.

WorldView-4 satellite was successfully launched on November 11, 2016. Similar to WorldView-3, it is capable of delivering the pan-chromatic image with 0.31 m resolution

and collecting four-band multi-spectral image at 1.24 m resolution. WorldView-4 satellite is able to revisit any point on the Earth every 4.5 days, depending upon the required look angle. The very high-resolution of provided images offers customers precise views for 2D or 3D mapping, change detection and image analysis tasks.

The description of the WorldView 1-4 satellites is taken from the web pages of the Satellite Imaging Corporationg [21] and European Space Imaging [22].

- **Pléiades 1A/1B**

Pléiades is an optical observation system built by Airbus Defence and Space consisting of two identical satellites, Pléiades 1A and Pléiades 1B, that deliver 50 cm color images. Pléiades 1A was successfully launched on December 16, 2011 followed by Pléiades 1B launched on December 2, 2012. Operating 180° apart on a phased orbit, the Pléiades system allows a daily revisit to any location on the Earth, which makes it ideal for mapping large scale areas. This information about the Pléiades satellites is taken from the Satellite Imaging Corporation web page [21].

### 2.1.4 Satellite Stereo-Based Digital Surface Model Generation

The high agility of modern *Very High-Resolution (VHR)* satellites, like WorldView-1 and 2, makes the collection of multiple images of the same area under different viewing angles in a single pass possible. By merging DSMs, generated from several image pairs applying SGM algorithms [17, 23, 24] which avoid matching windows, a better-quality DSM with minimum number of outliers and sharper object boundaries can be achieved. Elevation information extraction in the form of DSMs from along-track stereo satellite imagery is a highly valuable component for many remote sensing applications.

The DSM generation pipeline incorporates two main steps: 1) stereo imagery orientation computation considering given *Rational Polynomial Coefficient (RPC)* and 2) dense image matching. An accurate stereo matching is a very important procedure, because the density and accuracy of derived matching points directly influence the quality of DSMs.

- **Multi-View Image Orientation**

Elevation model extraction from multi-view very high-resolution satellite imagery requires accurate calculation of RPC camera parameters for each image. The initial RPCs known from orbit and attitude information cannot meet fitting requirements between multiple images in most situations. Thus, it is necessary to refine RPCs as the satellite position and orientation errors will result in systematic failures during dense image matching for generating DSMs.

Errors in sensor position and orientation can be corrected by a *bundle adjustment* process which ensures that the observations in multiple images of a single ground point are in a unified geodetic framework [25]. Several works [24, 26, 27] have been applying RPC bundle adjustment refinement procedure already to establish a good relative orientation

between the images. The correction of RPCs first requires accurate tie points between the stereo images. Initial tie points for stereo pair can be located at one pixel accuracy by sliding the pattern area over all the search region. These tie points are then refined to sub-pixel accuracy using local *Least Squares Matching (LSM)* [28]. Second, to correct absolute orientation, optimal *Ground Control Points (GCPs)* at sub-pixel accuracy are needed which are 3D locations of points known a priori from GPS observations or already existing DSMs and high-resolution orthorectified images. However, the acquisition of required GCPs could be tedious or even impossible, if a rapid response is needed in cases like large scale processing or crises situations. Because this type of information is not always available, contemporary research uses automatic image-based techniques to extract and correct GCPs [29].

- **Dense Image Matching**

To generate a detailed DSM from multi-view stereo images, a dense stereo matching method called SGM [17, 23] is used. The SGM defines a pixel-wise matching of mutual information and approximates global 2D smoothness constrains by following 1D constrains in multiple directions through the image [23]. The main fundamentals of SGM are matching and disparity map generation.

After establishing a good relative orientation, matching is performed for all potentially matching pixels in the stereo pair using epipolar geometry. SGM does not involve window matching technique [17, 23]. As a result, a reliable reconstruction of object edges can be achieved. To avoid strong local assumption on the local surface shape, the matching step is viewed as an energy minimization problem solved by optimizing the aggregation costs from 16 directions and finding disparity image $D$ with low energy

$$\mathcal{E}(D) = \Sigma_p \left\{ C\left(p, D_p\right) + \sum_{q \in N_p} P_1 T\left[|D_p - D_q| = 1\right] + \sum_{q \in N_p} P_2 T\left[|D_p - D_q| > 1\right] \right\}. \quad (2.1)$$

The function $C$ is a data term which defines the pixel-wise matching cost between the image pixels $p$ and corresponding pixels in disparity map $D_p$. The second and third terms of the energy function are regularization terms which favor similar disparities for neighboring pixels $N_p$ but also allows large jumps in areas with high contrast, respectively. As a result, if disparity changes are small (disparity change = 1 pixel), a constant penalty $P_1$ is added to the energy function. In contrast, if a higher disparity change occurs (disparity change > 1 pixel), a larger constant penalty $P_2$ is added. The function $T$ is a condition which is equal to 1 if the argument is true, and to 0 otherwise. The detailed description of cost aggregation is given by Hirschmuller [23].

Image pairs are matched from the first to the second, and from the second to the first image by applying the SGM on multiple images. It allows to keep only consistent disparities and avoids to fuse the most mismatched regions. Small isolated regions are rejected as irregularities. The obtained separate disparity maps from each image pair are first reprojected to the desired projection and merged afterwards to generate a single DSM by using a median filter [24].

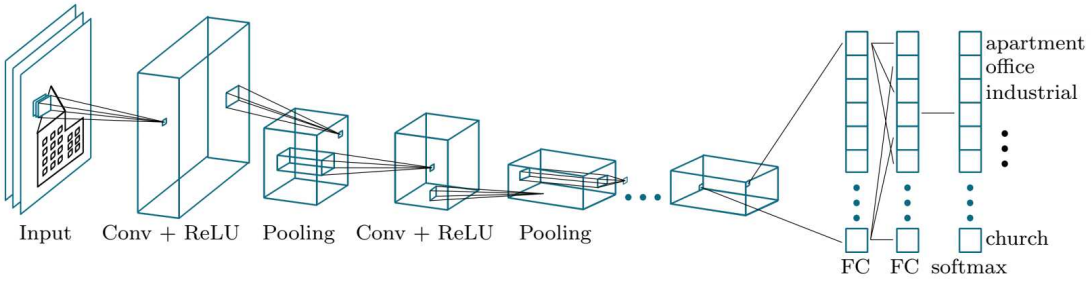



**Figure 2.5:** The schematic representation of CNN hierarchically structured of multiple convolutional, *Rectified Linear Unit (ReLU)* and pooling layers. The top layers are fully connected layers, which after applying the softmax normalization on each of the neurons represent a class of probability distributions.

## 2.2 Satellite Image Processing using Deep Learning Techniques

Currently, the amount of data obtained by remote sensors are large and complicated. As a result, only inspecting them manually is impractical or even impossible. Numerous image analysis approaches have been developed to interpret and extract information at the greatest extent possible from remote sensing images. The strategies applied for image analysis are strongly dependent on the goals of each individual project.

Until approximately five years ago, machine learning algorithms mainly used hand crafted features from the images, like spectral and textural, calculated from a few training samples. With the recent advances of deep learning methods for remote sensing imagery processing, it is now possible to solve many Earth observation problems automatically and at a high-level of accuracy. Deep learning algorithms demonstrate excellent results at recognizing patterns captured in neighboring pixels within the images. As a result, they learn to represent the image context as a hierarchy of patterns where each pattern is a combination of simpler and more abstract features. The most extensively used deep learning concept is *Convolutional Neural Networks (CNNs)* [30] that are particularly powerful for computer vision tasks, such as image recognition [31, 32], segmentation [33, 34] and classification [30, 35–37].

### 2.2.1 Convolutional Neural Networks

The search for good internal information about image objects has been the main intention since the beginning of computer vision. Many studies are based on a mathematical operation, called convolution. The general expression of discrete two-dimensional convolution is

$$(k * f)_{x,y} = \sum_{s=-W_1}^{W_1} \sum_{t=-W_2}^{W_2} k_{s,t} f_{x-s,y-t} \tag{2.2}$$

where $f$ represents the input image and $k$ defines the filter given by $s \in [-W_1, \ldots, W_1]$ and $t \in [-W_2, \ldots, W_2]$.

CNNs take the general advantages of neural networks and manage to work with 2D or 3D data. As a result, their training parameters represent 2D or 3D matrices. CNNs are typically composed of multiple trainable blocks. Each of the blocks receives a collection of feature maps and transforms it to a new collection of feature maps through differentiable functions. A common CNN block is built of three layers: a convolutional layer, a layer with non-linear activation function and a pooling layer.

A convolutional layer consists of $S$ filters of size $(2 \cdot W + 1) \times (2 \cdot W + 1)$ in spatial direction and of size $C$ in spectral direction with trainable weights $k_{m,n,c}$, where $m, n \in [-W, \ldots, W]$ determine the spatially localized region and $c \in [1, \ldots, C]$ determine the spectral one. The spectral size defines a number of features per pixel position, i.e., a number of channels, such as three for a RGB image. The feature map $y^l_{i,j,s}$ of layer $l$ at spatial location $i, j$ and at spectral channel $s$ is computed as

$$y^l_{i,j,s} = \sigma\Big(\sum_{m=-W}^{W}\sum_{n=-W}^{W}\sum_{c=1}^{C}(k^{l-1,s}_{m,n,c} \cdot y^{l-1}_{i-m,j-n,s} + b^{l-1}_{i,j,s})\Big) \tag{2.3}$$

where $b^{l-1}_{i,j,s}$ is a bias matrix. It is not necessary that the number of spectral channels $s$ stays the same along the whole network architecture. In most cases, it gradually increases from the bottom to the top of the networks.

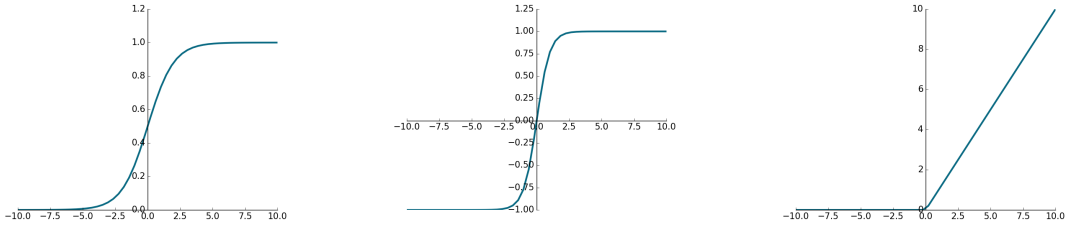
Due to the translation equivariance property [38], if one feature is useful to compute at some image location $(x_1, y_1)$, then it should also be useful to compute at a different image location $(x_2, y_2)$. As a result, the same weight parameters defining identical features can be applied around the whole image. In CNNs, this strategy is extensively exploited allowing the sharing of weights and biases across the entire layer. With local connectivity and shared weights, CNNs tend to reduce a number of learning parameters compared to the standard fully connected neural networks and provide a better generalization when dealing with computer vision problems [39].

In Equation (2.3), the $\sigma(x)$ defines a non-linear activation function which helps the network to learn a complex data. In other words, non-linearities ensure that the model is able to perform more complex function approximation than a linear one. Common non-linear functions used for feed-forward neural networks are the *sigmoid* (see Figure 2.6a)

$$\sigma_{\text{sigmoid}}(x) = \frac{1}{1 + \mathrm{e}^{-x}} \tag{2.4}$$

or the *hyperbolic tangent* (see Figure 2.6b)

$$\sigma_{\tanh}(x) = \frac{\mathrm{e}^{x} - \mathrm{e}^{-x}}{\mathrm{e}^{x} + \mathrm{e}^{-x}} \ . \tag{2.5}$$



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

**Figure 2.6:** Examples of (a) sigmoid, (b) tanh, and (c) ReLU non-linear activation functions commonly used for feed-forward neural networks.

For CNNs it is common to use the *rectified linear unit* non-linearity

$$\sigma_{\text{ReLU}}(x) = \max(0, x) \tag{2.6}$$

which sets all negative values in the convolution matrix to zero introducing the sparsity to the network and keeps the positive values unchanged (see Figure 2.6c). This is beneficial for training the network with backpropagation [40] because it prevents the gradient vanishing problem and makes the computational implementation more efficient.

A pooling layer progressively reduces the information in the spatial range $[-P, \ldots, P]$. The most common pooling operation is *max pooling*

$$p^l_{i,j,s} = \max_{m \in [i-P, i+P], n \in [j-P, j+P]} \left\{ y^l_{m,n,s} \right\} \tag{2.7}$$

which takes the maximum element in a local region, thereby providing a small degree of invariance. Pooling operation reduces the number of parameters and computation in the network, and as a result is able to control overfitting.

CNNs were initially developed for image classification tasks. Thus, the top layers of CNNs are usually fully connected layers, which merge the information of the whole image to predict the correct class associated with it. The final layer is a 1D array of neurons representing a class of probability distributions after applying softmax normalization on each of the neurons. The size of the last 1D group of neurons is equal to the number of possible classes established for the task. A multi-stage design of CNN architectures allows to automatically and adaptively learn spatial hierarchies of features, from low-level components at the bottom, such as edges and corners, to high-level semantic information at the top layers. In other words, convolutional and pooling layers transform the input image to a higher spectral but low spatial resolution abstract representation which are easier to separate.

### 2.2.2 Training the Network with Backpropagation

The signal of the entering data is distributed in forward direction through neural network parameters towards the point where a decision will be made. Training a network means

finding such parameters which minimize dissimilarities between predicted values by the network and given ground-truth labels on a training dataset. The misclassifications are penalized by a task-dependent loss function. Tasks based on classification use the cross-entropy loss function

$$\mathcal{L}_{\text{CE}}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{p}) = -\sum_i y_i \log p(x_i) \tag{2.8}$$

paired with sigmoid (*cf.* Equation (2.4)) or softmax

$$\sigma_{\text{softmax}}(\boldsymbol{x}) = \frac{e^{x_k}}{\sum_j e^{x_j}} \tag{2.9}$$

normalization, while regression tasks use the Euclidean error function

$$\mathcal{L}_{\text{Eucl}}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{p}) = \sum_i \|y_i - p(x_i)\|_2^2 . \tag{2.10}$$

The Euclidean error function can be also used in case of binary classification. However, the common choice of the cross-entropy loss function over the Euclidean loss function for training the CNNs for the binary classification task can be explained by its ability to avoid the slowdown problem during the learning process and to provide a more numerically stable gradient when coupled with softmax normalization [41]. Here, $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ is the set of input examples in the training dataset and $\boldsymbol{y} = \{y_1, \ldots, y_n\}$ is the corresponding set of target values for those input examples. The $p(x_i)$ represents the predicted output of the neural network for provided input $x_i$.

The main objective of neural networks is to create a model from the training data which correctly generalizes on a very wide range of previously unseen data. It is possible, due to the network's ability to learn relevant features in the data. A method which is only able to work on the training data is not effective.

To prevent a tendency of neural network to overfit the training data, a process of prior knowledge insertion, called regularization, is usually employed. It can appear in many forms, e.g., data augmentation, but a common way is to add an additional term to Equation (2.8) called "weight decay",

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{p}) = \sum_i \mathcal{L}_0(y_i, p(x_i)) + \lambda \|w\|_2^2 , \tag{2.11}$$

which prevents the weights growing too large. Here, the term $\mathcal{L}_0$ defines the cross-entropy loss function (*cf.* Equation (2.8)) and $\lambda$ is a constant controlling how strong large weights should be penalized.

By minimizing the loss function (*cf.* Equation (2.11)), w.r.t the weights $\boldsymbol{w}$ and biases $\boldsymbol{b}$, the network can be trained. Generally, neural networks are trained using the back-propagation algorithm [40], which computes the derivatives of the loss function with respect to parameters $\frac{\partial \mathcal{L}}{\partial w_i}$ and $\frac{\partial \mathcal{L}}{\partial b_i}$, and propagates them back through the network so

that they can alter the parameters

$$w_i \leftarrow w_i - \alpha \frac{\partial \mathcal{L}}{\partial w_i} \tag{2.12}$$

$$b_i \leftarrow b_i - \alpha \frac{\partial \mathcal{L}}{\partial b_i} \tag{2.13}$$

with learning rate $\alpha$ one step at a time. This standard loss minimization technique is called *gradient descent*. Since a neural network is a composition of multiple simple functions, the computation of derivatives for the gradient descend algorithm is done by applying the chain rule

$$f(g(x))' = f'(g(x)) \cdot g'(x). \tag{2.14}$$

Computing the gradient descend for data with thousands (or more) elements for every iteration is impractical. Therefore, an approximation called the *Stochastic Gradient Descent (SGD)* with a mini-batch is a typical choice, as the gradient computation is done on a small number of data points rather than on the entire dataset. The training can be further optimized by SGD with momentum [42] which is often considered an essential component to train deep networks. In specific cases, the extensions like Adagrad [43], Nesterov's accelerated gradient descent [44], Adadelta [45] and Adam [46], tend to be very similar in quality, if not better than vanilla momentum.

Deep neural networks have a large amount of parameters. As a result, they are prone to overfitting on the training data. In addition to the classical weight decay regularization (*cf.* Equation (2.11)), a *dropout* [47] regularization technique shows its power and strength to improve generalization for deep neural networks. The term "dropout" refers to ignoring a random amount of units (along with their connections) from the neural network during training [47]. Consequently, only a random subgroups of neurons are trained in a single iteration of SGD. This prevents neurons from co-adapting too much to the problem.

### 2.2.3 Diversities of Convolutional Neural Networks

#### 2.2.3.1 Recurrent Neural Networks

*Recursive Neural Networks (RNNs)* are powerful and robust types of neural networks which have gained more popularity in the last years. Unlike feed-forward networks, the RNNs are networks with loops in them, also called recurrent connections. The major advantage of RNNs is that with these connections the network is able to refer to last states, allowing information to persist over time, and can therefore making use of sequential information. This is similar to how the human brain processes the information. Humans do not start their thinking from scratch every second. For example, when we read a poem, we understand each sentence based on our understanding of the previous words.

A diagram in Figure 2.7 demonstrates that a block of the RNN has connections from

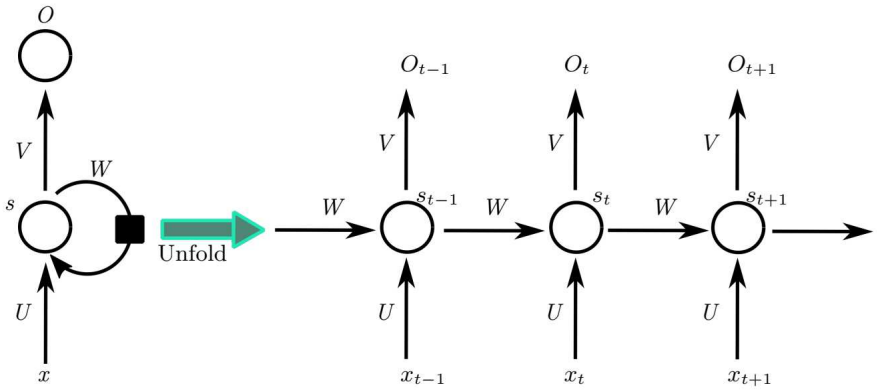



**Figure 2.7:** A diagram of an unfolded RNN represented as a chain of repeated units where $\boldsymbol{U}$ defines weight vector for hidden layer, $\boldsymbol{V}$ represents weight vector for output layer, $\boldsymbol{W}$ represents the same weight vector for output layer but for different time steps, $\boldsymbol{x}$ is an input vector and $\boldsymbol{O}$ is an output vector.

its input to its output as well as a connection from its output again to its input. This extra loop connection allows information to be passed from one step of the network to the next.

The illustrated RNN block can be unrolled (or unfolded) into a full network (Figure 2.7) of $k$ instances. The number of $k$ instances depends on the length of the whole sequence. For example, if an input sentence consists of three words, the network is unfolded into a three-layer neural network, where one layer is responsible for one word. If we specify the hidden state at time step $t$ as $s_t$, the process of carrying memory forward can be defined as

$$s_t = \sigma\left(U \cdot x_t + W \cdot s_{t-1}\right) \tag{2.15}$$

where $x_t$ is the input or the second word of a sentence at a time step $t$ modified by a weight matrix $W$, added to the hidden state of the previous time step $s_{t-1}$ multiplied by its own hidden-state-to-hidden-state matrix $U$. The function $\sigma$ is one of non-linearity functions described in Section 2.2.1.

The training of RNN is done on unfolded networks, like in Figure 2.7, with a backpropagation algorithm the same as for the feed-forward networks, except that each epoch has to run through each unfolded layer. The extension of backpropagation algorithm for RNN receives the name *Backpropagation Through Time (BPTT)*.

RNNs have demonstrated superior results in many tasks. Their ability to deal with sequential data is significant for applications like image or video captioning [48], handwriting recognition [49], speech synthesis [50] and speech recognition [51], where a model produces sequential outputs. In the case of time series forecasting [52], video analysis [53], and musical information retrieval [54], a model must learn from inputs that are sequenced. Some interactive tasks, like translating natural language [55, 56], engaging in dialogue [57], and controlling a robot [58], often require both capabilities.

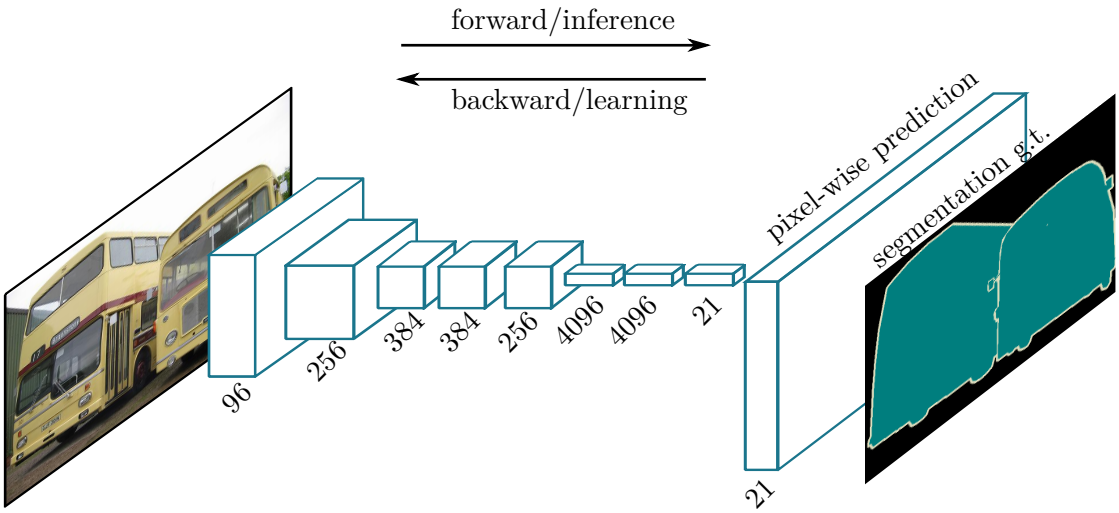



**Figure 2.8:** Fully convolutional network representation. Compared to standard CNNs, *Fully Convolutional Networks (FCNs)* are able to learn to make dense predictions for each pixel within a given input image. The image is adapted from Long *et al.* [59].

#### 2.2.3.2 Fully Convolutional Networks

The semantic segmentation task is different from the classification task because of its requirement to assign a class for each pixel from the input image (see Figure 2.8), instead of only one class for the whole input image. This problem is extremely difficult because the method should be able to classify and to locate the objects at the same time. Moreover, the produced output should have the same size as the original input image.

Recently introduced FCNs [59] have become the state-of-the-art methodologies for the task of image segmentation. They are an adaptation of traditional CNNs, where the top fully connected layers are changed with convolutional layers. As a result, unlike the basic CNNs, the spatial information in the top layers is not lost when applying FCNs, but can be tracked back [60]. This also reduces the number of parameters and computation time as well as makes the network independent from the input size.

The output layer of FCNs is the same height and width as the input image, but the number of channels is equal to the number of classes of the task and represents the per-class probability maps $cl_i(x, y)$. In the case of binary classification, the shape of the final output layer will be $height \times width \times 2$.

Semantic segmentation requires a mechanism to project the lower-resolution features learned at different stages of the encoder onto the higher-resolution space to obtain a dense classification. For the encoder, a pre-trained classification network like VGG [36] or ResNet [61] is commonly used. The recovering of spatial information is done in the so-called decoder part of the network using transposed convolution layers which perform a learned interpolation from a set of nearby points.

Applying up-sampling to bring the classification maps to the original size does not guarantee the production of accurate segmentation masks with very detailed object

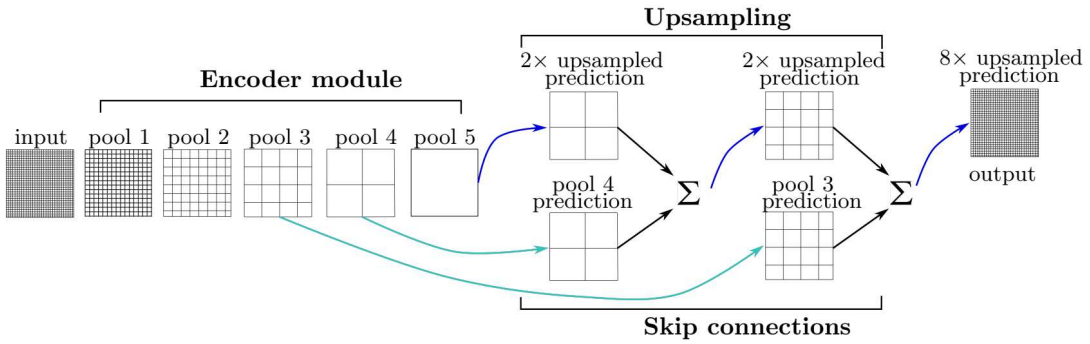



**Figure 2.9:** The schematic representation of FCN-8s architecture introduced by Long *et al.* [59] which integrates the mechanism called "skip" connections (——) into up-sampled (——) probability maps for more detailed recovery. The illustration is adapted from Long *et al.* [59].

boundaries beacause too much spatial information has been lost by all down-samplings in the network. Long *et al.* [59] propose to combine the high-frequency information from encoder feature, skipping some layers of non-linear processing, with the output from transposed convolution layers at the same resolution using an *element-wise addition* (see Figure 2.9). These skip connections from earlier layers in the network provide more precise spatial information which helps to recover more detailed shapes for segmentation boundaries.

#### 2.2.3.3 Region-Based Convolutional Neural Networks

Unlike the traditional image classification tasks where only one label is assigned by CNNs to an image context, object detection algorithms attempt to draw a bounding box around the object of interest to locate it within the image. In most real-case scenarios, there could be many bounding boxes representing different objects of interest within the image. The problem of object detection and localization requires precise estimation of the objects location as well as their classification. Object detection is a challenging task because multiple objects with varying sizes can be present in one image.

As opposed to traditional exhaustive sliding window approaches, the region proposal algorithms output a set of object proposals in the form of object-like image patches at multiple scales. Girshick *et al.* [31] employ such selective search approach [62] to generate approximately 200 category independent region proposals on a test image. The wrapped region proposals are then fed to a CNN trained for classification, which in turn generates a fixed size high-level semantic feature. At the end, each region is classified by a set of category specific linear *Support Vector Machines (SVMs)*. The schematic representation of described architecture, named *Region-based CNN (R-CNN)*, is illustrated in Figure 2.10.

Further efforts were performed to improve the R-CNN approach in a way to make it more computationally efficient with a high rate of detection accuracy. For example, Ren

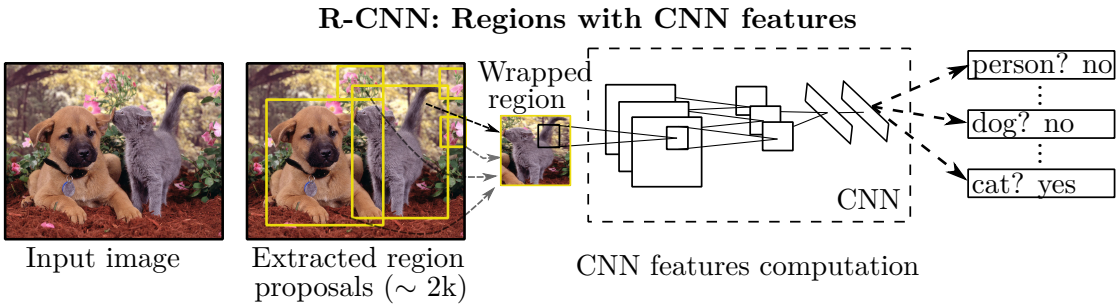



**Figure 2.10:** Schematic representation of Region-based CNN developed for object detection problems. After generating a set of region proposals, the R-CNN passes the wrapped proposals through a CNN to classify them. The illustration is adapted from Girshick *et al.* [63].

*et al.* [64] introduce Faster R-CNN where the selective search algorithm [62] is replaced by FCN called *Region Proposal Network (RPN)* to compute region proposals at no extra cost. Another advantage of proposed work compared to the earlier approach [31] is a decreased amount of proposals (as low as 20).

### 2.2.4 Generative Adversarial Networks

*Generative Adversarial Networks (GANs)* were introduced by Goodfellow *et al.* [65] in 2014 as a new class of neural networks. Their potential is tremendous, because they can learn to approximate any distribution of data. It means that such generative models are able to create a new content.

GANs are adversarial frameworks composed of two competing neural network models: the *generator* and the *discriminator*. The task of a generator is to produce samples of fake data which will fool the discriminator. The task of a discriminator is to evaluate the inputs for authenticity. The generator never sees the genuine data. It must learn to generate realistic data by considering feedback from the discriminator.

The idea behind the adversarial training is the competition between the generator and the discriminator which drives both networks to improve their methods: both networks try to beat each other and by doing so, they sharpen each other's skills. As a result, the discriminator is able to detect fake samples, while the generator finds a way to produce data indistinguishable from the real samples.

GANs are formulated as a minimax game

$$\min_G \max_D \mathcal{L}_{\text{GAN}}(G, D) = \mathbb{E}_{\boldsymbol{y}\sim p_{\text{real}}(\boldsymbol{y})}[\log D(\boldsymbol{y})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))], \quad (2.16)$$

where a generator $G(\boldsymbol{z}) = \boldsymbol{y}$ maps a latent vector $\boldsymbol{z} \sim p_z(\cdot)$ from normal or uniform distribution $p_z(\cdot)$ to a fake sample $\boldsymbol{y}$ which resembles a real one $\boldsymbol{y}^* \sim p_{\text{real}}$. A discriminator $D(\boldsymbol{y}) \in [0, 1]$ outputs a single scalar estimating the probability that $\boldsymbol{y}$ comes from the real data rather than from generated one. In Equation (2.16), the $\mathbb{E}[\cdot]$ defines the expectation value. The diagram of GAN's working process is visualized in Figure 2.11.

Currently, GANs are an intense research topic and many different types of GAN are

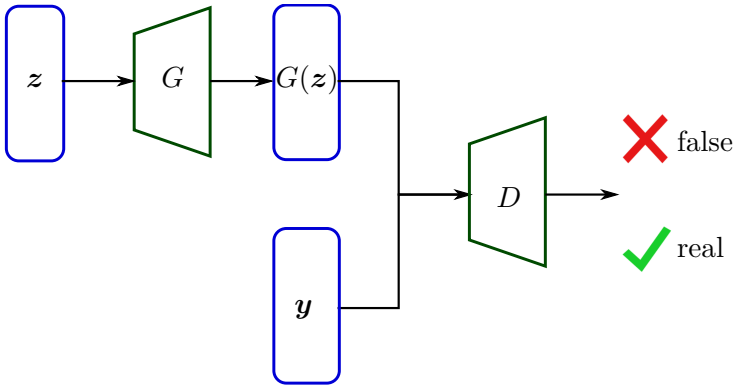



**Figure 2.11:** Schematic representation of GAN model developed for generating new data $G(\boldsymbol{z})$ from the random noise $\boldsymbol{z}$ with the same context as the one learned from ground-truth data $\boldsymbol{y}$.

implemented. Some of these implementations which are being used the most actively are briefly described below.

- **Vanilla GAN**

Vanilla GAN [65] is the fundamental type of GAN. Its generator and discriminator are built of simple multiple layers of perceptrons. The input to the generator is the data with random distribution. The generator learns to convert this random data into the meaningful information that meets the requirements of the real data distribution. The algorithm tries to optimize the mathematical Equation (2.16) using SGD.

- **Conditional GAN**

*Conditional Generative Adversarial Networks (cGANs)* are an extension of vanilla GANs which were first introduced by Mirza *et al.* [66] in 2014. It is a deep learning framework where extra information conditions both the generator and the discriminator. The added contextual information helps the generator to produce the corresponding data as well as helps the discriminator to distinguish the real data from the generated data. The range of cGANs applications is broad. It includes reconstructing objects from edge maps, generating photos from label maps, colorizing images [67, 68], predicting future frames in a natural video sequence [69], etc.

- **Laplacian Pyramid GAN**

The Laplacian pyramid [70] is a class of linear transform which breaks down an image into diverse components through a set of transform functions. *Laplacian Pyramid GANs (LAPGANs)* [71] integrate the different levels of the Laplacian pyramid into the cGAN framework to generate high-quality samples of natural images. The image is first down-sampled at each layer of the pyramid and then, after acquiring a noise generated by a cGAN at each layer, it is up-scaled again back to its own size.

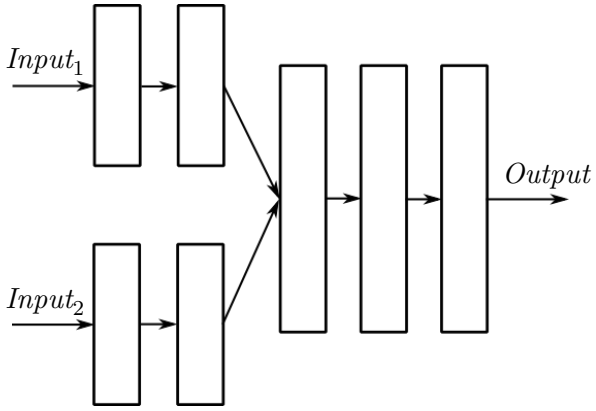



**Figure 2.12:** Schematic representation of the multi-modal learning concept.

- **Super-Resolution GAN**

*Super-Resolution GAN (SRGAN)* [72] employs a deep neural network together with an adversarial network to produce images with higher-resolution. During the learning process, a low-resolution image is obtained by down-sampling a high-resolution image. The task of the generator is to up-sample low-resolution images to super-resolution ones. The task of the discriminator to differentiate between real and generated high-resolution images, and backpropagates the GAN loss to train the network. SRGANs are particularly useful for upscaling native low-resolution images to enhance their details.

### 2.2.5 Multi-Modal Networks

In the past five years, the integration of information from multiple data sources ($Input_1$ and $Input_2$ in schematic representation of neural network in Figure 2.12) for one specific task retrieval has become very popular. Several studies [73, 74] show that data fusion can make information more confident, intelligent, consistent, and accurate compared to the results from individual data sources. The logic behind it is similar to the human ability to process information. For instance, humans combine signals from the five body senses (sight, sound, smell, taste, and touch) with the knowledge about the environment to understand and perceive the world around them. Based on this information, the individual interacts with the environment and makes decisions about present and future actions [75]. This natural ability to merge multiple data sources has expanded into many fields of science.

Multi-model fusion can be fundamentally categorized into early and late fusion [76]. While the joint information is directly extracted from the input data by the early fusion, the late fusion accumulates the individual decisions made by the neural networks for each modality [77].

The most common types of combined modalities are from the image space, like RGB-D and motion information from the videos. Numerous works have investigated that the inclusion of depth information into the system positively influences the final results, as

it contains additional information about object shape. Socher *et al.* [78] explored the integration of appearance and texture information from spectral images together with depth images to address the classification problem. A single-layer CNN processes the data separately for low-level feature extraction. The obtained low-level features are given to a set of RNNs to generate a joint representation for the final softmax layer. Eitel *et al.* [79] proposed CNN architecture where the RGB and depth images are simultaneously processed in two individual streams joint at the end with a late fusion network. Wang *et al.* [80] approached the problem differently and merged the RGB-D data at the beginning, forming four-channel input, to perform a scene labeling.

Compared to still images, video frames provide additional information in form of dynamics. The earlier methodologies used to feed the raw video frames into CNNs. Jhuang *et al.* [81] proposed to use a predefined set of spatio-temporal filters in the initial layer of an introduced model for action recognition task. Ji *et al.* [82] approached a human action recognition problem by introducing for the first time an end-to-end training of CNNs with 3D convolutions over series of consecutive video frames. However, it was found that learning spatio-temporal filters was not suitable to efficiently capture the motion patterns. Therefore, later approaches, such as the one introduced by Simonyan *et al.* [83] and Kuehne *et al.* [84], proposed a two-stream model which separates videos into spatial and temporal components.

There are also approaches that combine images with other forms of data to learn different tasks. For example, to perform an accurate meteorological analysis, data acquired from various devices, such as rain gauges, radars, and space-borne remote sensing devices, are often incorporated together due to their complementarity for better accuracy, coverage and resolution [85–87]. The idea behind the human-machine interaction problem is the combination of multiple interaction modes based on audio, vision, touch, smell, movement, and interpretation of human language commands [88, 89]. Moreover, data fusion applications include medical and industrial robotics for objects recognition and processes monitoring [90–92], military systems designed for battlefield surveillance, automated threat recognition systems and remote sensing [93], autonomous driving [94], etc.

### 2.2.6 Multi-Task Learning

Machine learning algorithms usually solve a single problem at one time, i.e., we expect a single output at the end, like image classification, segmentation or even object generation. However, several surveys [95–98] demonstrate that neural networks are able to jointly solve multiple tasks ($Output_1$ and $Output_2$ in schematic representation of neural network in Figure 2.13) at the same time and benefit from similarities between them. As a result, this helps to contribute with an improvement to the generalization performance of every task. The logic behind the improvement of related tasks is borrowed from transfer learning [99] which aims to transfer knowledge, such as feature or attributes, learned from an earlier task to another task. However, the multi-task learning problem uses multiple tasks to mutually benefit each other while transfer learning uses one or more tasks to share the knowledge with a target task.

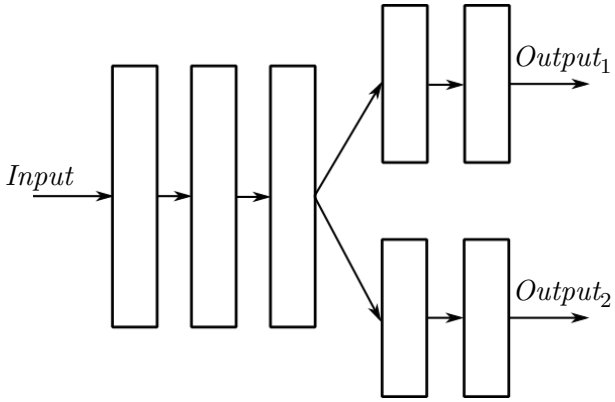



**Figure 2.13:** Schematic representation of the multi-task learning concept.

Multi-task learning is embedded in many applications, ranging from modeling consumer buying decision [100], genomic[101], natural language processing [102] and speech recognition [103], to a huge variety of computer vision tasks [104–107]. The majority of them focus on pose estimation and action recognition [108], classification and segmentation [109] or recognition, localization and classification [110], etc. Misra *et al.* [111] proposed the *CrossStich* network with a new sharing unit which combines the activations from several networks for learning the appropriate combination of shared and task-specific information. Teichmann *et al.* [112] introduced the *MultiNet* architecture for joint classification, detection and semantic segmentation in the autonomous driving filed. Uhrig *et al.* [113] developed a FCN that predicts semantic labels, depth and an instance-based encoding using each pixel's direction towards its corresponding object center. Kendall *et al.* [114] approached the similar task differently and developed a model which learns per-pixel depth regression, semantic and instance segmentation from a monocular input image by incorporating multi-task weighting algorithm.

Some researchers made an attempt to perform multi-task learning for geometry and regression tasks. Eigen *et al.* [115] proposed a multi-scale CNN architecture which predicts depth, surface normals, and semantic labels at the same time from a single RGB image. Kokkinos [116] developed an *UberNet* which uses diverse training datasets and simultaneously handles a number of different regression and classification tasks. Contrary to previous works, Liebel *et al.* [117] proposed a concept of employing seemingly unrelated auxiliary tasks, such as time or weather prediction, together with a single-image depth estimation and semantic segmentation into a multi-task learning paradigm. Those additional tasks play the role of a regularization measure and enhance the robustness, performance, and speed of the training. Recently, Lin *et al.* [118] presented an approach very similar to ours where they addressed the joint depth estimation together with semantic segmentation from a single RGB input image. Their developed CNN architecture follows the standard multi-task idea: substituting the encoder part of the architecture by a common feature-extraction network for both tasks, while separating the global semantic segmentation and depth estimation task at the network bottleneck.

The network responsible for depth estimation was constructed by three components: a global depth network, a gradient network, and a refining network which are all based on *AlexNet* structure. The network responsible for semantic segmentation extracted the features using VGGNet architecture and was followed by the atrous up-sampling network to output a class score maps. The developed network was intended for real-time autonomous driving applications.

## 2.3 Summary

Modern technologies open broad opportunities for analyzing and monitoring the changes happening on the Earth. The best way to obtain fast and large-scale information about the Earth is to explore worldwide satellite imagery data. Worldwide satellite imagery data gives extensive knowledge about buildings and terrain for hundreds of cities which is crucial for many remote sensing applications. Information such as the outline of buildings and their 3D representations can be obtained by taking advantage of the currently available very high spatial and spectral resolutions, together with the height information retrieved from multi-view stereo satellite images. However, those tasks are challenging due to the variety of building shapes as well as their complexity. The strength of recently developed methodologies which are able to extract task dependent features automatically promises to achieve good results with a high degree of accuracy. Therefore in this dissertation, two problems related to automatic building information extraction and refinement are investigated utilizing deep learning-based approaches.

# Chapter 3

# State-of-the-Art

Information extraction from remote sensing images, which includes rich geometric, topological, and semantic knowledge of buildings, has been an active research field in photogrammetry and computer vision for many years. Building information extraction plays a significant role in a wide range of practical applications, such as updating *Geographic Information System (GIS)* databases, city planning, land use analysis, estimation of population density, disaster management, etc. [119–121]. Because the manual processing of geometric and semantic information of buildings is time-consuming and costly when applied to both regional or global scales, the development of a methodology that requires less human interaction is under demand. In this chapter, we review the existing methodologies dedicated to both semantic information extraction of a building as well as its geometry refinement that motivates our new developments in the remote sensing research field. The chapter is an extended version of the literature review introduced in our four publications [3–6].

## 3.1 Building Footprint Extraction from Very High-Resolution Satellite Imagery

Automatic building detection and footprint extraction from high-resolution satellite imagery (see Figure 3.1) is an intense field of research in the area of remote sensing. However, due to the sophisticated nature of urban surroundings, large diversity of building structures, and relatively low spatial resolution, the collection of detailed building outlines from remotely sensed imagery is very challenging. Therefore, automatic methods are needed to perform time and cost effective extraction of building footprints from *Very High-Resolution (VHR)* satellite imagery within urban areas containing numerous, and at times very complex constructions.

**Figure 3.1:** An example of a building footprint map overlaid to a *red, green, and blue (RGB)* image with a resolution of 0.5 m acquired from the Worldview-2 satellite. The depicted area is located in the city of Munich, Germany and represents a 1.1 $km^2$ coverage.

### 3.1.1 Traditional Methodologies

In general, the existing methods can be grouped into three classes according to the data used for building footprint extraction. The first class uses single aerial or high-resolution satellite imagery, the second class utilizes structural information in the form of *Digital Surface Models (DSMs)* and the third class combines the spectral and height information because they provide complementary knowledge and increase the opportunity for enhancing final results.

#### 3.1.1.1 Spectral Information Based Approaches

Monocular imagery provides the spatial and spectral information useful for terrestrial scene interpretation. The early studies on building extraction emphasize the importance of features obtained from this type of information. Usually low-level features—such as edges, corners, line segments and their combinations—are detected and grouped together based on various rules to form building hypotheses [122–125]. Moreover, it is observed that building rooftops within relatively homogeneous areas have regular shapes represented by rectangles or combinations of them. As a result, methodologies extensively employ the shape information.

- **Geometric Relationships**

Mohan *et al.* [126] proposed to detect buildings in aerial images based on perceptual organization taking into account the structural relationships. The authors applied a number of rules to detected linear features while grouping them into parallels: either linear features form an U-structure in case of their collation with aligned endpoints or a complete rectangle hypothesis if two such U-structures are detected. The best consistent rectangles were found by minimizing the cost of the constraint satisfaction network.

Krishnamachari *et al.* [127] performed a *Markov Random Field (MRF)*-based building delineation in aerial images. The method involved straight line segments extraction from the generated edge map of an image and applied the MRF on line interactions to provide final building contours. Liu *et al.* [128] introduced a method which segmented a *pan-chromatic (PAN)* image into different regions, and used all detected contours as candidates. To distinguish "true" buildings among all possible candidates, the probability model was used. The assumption that building borders are perpendicular or parallel was incorporated into the algorithm. Saeedi *et al.* [129] used the *Blue Line* edge detection approach to identify straight segment lines and group them into building hypothesis. To remove unwanted extracted objects, like pools and green areas, the authors applied the multi-level verification process on a constructed hypothesis. The drawback of this methodology is that it only produced squared building contours and in some cases did not follow the exact contour of the building shape. Wang *et al.* [130] presented a framework which first applied edge-preserving bilateral filter to enhance edge contrast. The line segment detector *EDLines* was then used for real-time building line segments extraction which were finally grouped into candidate rectangular buildings by graph search-based approach. In general we can say that the detection of relevant building-correspondence lines is a complicated problem, as many other irrelevant and randomly distributed lines are usually fragmented in the images.

- **Graph Structure Models**

Some methodologies establish the building detection in terms of graph models. Kim *et al.* [131] modeled low-level linear features extracted from aerial images as vertices of a line-relation graph. By searching the graph, building hypotheses were found. Building hypotheses were finally verified using the intensity and shadow information. Segl *et al.* [132] combined supervised shape classification process with unsupervised image segmentation algorithm in an iterative procedure to allow an object-oriented inspection like building shapes in high-resolution satellite images. Molinier *et al.* [133] proposed the detection of man-made constructions by training self-organizing maps on satellite images. Sirmacek *et al.* [134] formalized the problem of urban areas and buildings detection as a graph theory where extracted *Scale Invariant Feature Transform (SIFT)* key points were assigned to graph vertices and special relationships between these vertices, e.g., intensity values, form the edges of the graph. The approach first detected urban areas using a graph matching method and then identified separate buildings applying graph-cut algorithms. The method was able to extract building hypothesis but failed to generate accurate building boundaries. Ok [135] proposed to extract building candidates from single VHR multi-spectral images based on a graph theory framework. The methodology mainly consisted of a two-level partitioning algorithm which employed the relationship between buildings and their shadows to obtain accurate building regions. As the approach relays on *near-infrared (NIR)* band and solar angles, only partial availability from relative prior knowledge can lead to inefficiency. Although the algorithms based on geometrical primitives achieve good results, they experience difficulties, especially with more complex, non-rectangular building shapes.

- **Active Contour Models**

Active contour models, also called snake models, are other effective approaches for building detection. The active contour model was first defined by Kass *et al.* [136] as an energy-minimizing spline which works towards object contours approximation under external constraint forces. Compared to traditional edge-based methodologies, active contour models are more flexible for detecting unstructured or complex shapes of objects. Active contour models are used as the fundamental step for building segmentation task because they approximate the building regions. It reduces unnecessary information processing, since a whole scene may contain noises and non-building features, and improves the performance of building segmentation in terms of computational time and accuracy. With time, the concept of active contours has been expanded to several improved models, which includes Balloons snake [137], *Generalized Gradient Vector Flow (GGVF)* snake [138], region-based active contours [139], *Curvature Vector Flow (CVF)* snake [140], etc. Numerous active contour-based approaches have been developed for extracting buildings from aerial and satellite images. Ahmady *et al.* [141] proposed to extract buildings from aerial urban images using a Chan-Vese [139] active contour model that is based on the Mumford-Shah image segmentation algorithm [142]. The major disadvantage of this approach appears due to a series of regular circles in initial snake algorithms that lead to a large computational time and a detection of other objects if they feature similar spectral information as buildings. The method proposed by Fazan *et al.* [143] was based on dynamic programming which was used to reduce the number of required processing steps when looking for a set of optimal variables for the energy function. The method is semi-automatic because the human operator provides initial approximated curves describing the building roof contours for the snake model. The selection of the best set of edge points for forming building roof contours was done by an optimization process based upon initial snake curve results. The method did not show superior results because the snake algorithm was not able to converge correctly to weak edges. It also experienced difficulties with convergence for concave contours. Overall, the approaches based on active contour models can achieve high performance in extracting buildings but suffer from low-quality in case of deep concavities which complex buildings exhibit at times.

- **Shadow Information Assistance**

Some methodologies reported challenges in building detection from monocular images referring to color resemblance between buildings, roads and squares or the variation of building appearance due to the difference in illumination, reflectance, and characteristics of the optical sensors [144]. To overcome these challenges, different aspects of knowledge, like shadow and illumination, were investigated to improve the building extraction problem. For example, the shadow presence can serve as hints for building location [145] or prediction of its shape and height properties [124]. The idea behind shadow information usability is that its spectral characteristics—like color, shape and brightness—is common for different buildings. Huertas *et al.* [123] employed the shadow evidence to estimate the

building corners and wall sides, and established building hypothesis under the assumption that their structures are rectangular or built of rectangular components. Liow *et al.* [146] also used the shadow information to complete the grouping process of the boundary segments as well as to improve their accuracy. McGlone *et al.* [147] incorporated the shadow information together with image orientation to check a building hypothesis based on the vanishing point as well as the projective geometry calculations. In general, the methods based on shadow characteristics are failing when the solar illumination of the ground surface is affected by the cloud cover.

- **Multi-Spectral Information Contribution**

The studies which only use single-band images are limited in their performance. With the recent availability of VHR multi-spectral aerial and satellite imagery, attention was drawn to the investigation of multi-band information as another useful data source for the building detection task. For example, the *Normalized Difference Vegetation Index (NDVI)* data extracted from *red* and NIR channels of a multi-spectral image indicate vegetation and as a result, can help to eliminate trees. Early approaches for image classification typically employed task-specific features like color histograms or local binary patterns and passed them to machine learning algorithms to generate a labeled image [144, 148, 149]. Ngo *et al.* [150] decomposed an image into small homogeneous regions, which were then grouped into clusters. The assumption that buildings are typically accompanied with shadows was used to merge these building segments with their neighboring regions in the same cluster to produce final building proposals.

For further overview about building detection in aerial and satellite imagery the reader is refered to Mayer [151] and Ünsalan *et al.* [152].

#### 3.1.1.2 Elevation Information Involvement

Features can be extracted not only from spectral images. Because buildings are elevated objects, the height information, if existing, is a valuable feature for detecting and extracting buildings. Therefore, integration of elevation information into building detection methodologies can significantly increase the accuracy and precision as well as robustness of extracted building footprints. Sirmacek *et al.* [153] delineated building outlines from DSM data using building skeletons, which were separated into diverse pieces for a box-fitting algorithm. The active rectangular shape growing was then performed, until the difference between the previously extracted building edges and the rectangle was reduced. Brédif *et al.* [154] proposed building footprint extraction from DSMs through a two-step algorithm based on a global optimization solver. The first energy aimed to extract rectangular building footprints directly from elevation models using a *Marked Point Process (MPP)* of rectangles preventing their overlapping but forcing the alignment with DSM discontinuities. The second energy constructed rectangular building tiles into the complete building polygon. The method is even able to detect complex shapes of buildings but is limited to low-height buildings as well as the ones located in inner courtyards.

#### 3.1.1.3 Data Combination

The combination of spectral imagery and DSMs, derived from either VHR remote sensing images acquired from two or more viewing angles, *Light Detection and Ranging (LiDAR)* or *Synthetic Aperture Radar (SAR)* interferometry, is the most prominent application for data fusion, as both modalities have their advantages and limitations, and can complement each other for achieving the final goal. Rottensteiner *et al.* [155] fused features extracted from a LiDAR DSM and a single nadir RGB image using the Dempster-Shafer methodology [156] for building delineation. First, a polymorphic feature extraction algorithm [157] was applied to the first derivatives of the DSM to compute the surface roughness. Initial building regions were also derived by classifying the image pixels as buildings or non-buildings. The final building detection was performed applying the Dempster-Shafer theory with defined cues in forms of two surface roughness parameters and height differences between the DSM and the *Digital Terrain Model (DTM)*, the last laser pulse, NDVI and initial building regions. Although the approach showed promising results, it failed to detect small size buildings due to the low number of LiDAR points. Sohn *et al.* [158] first determined the building objects by analyzing the height property of laser points and NDVI computed from IKONOS imagery. A full description of building outlines was then accomplished by merging convex polygons obtained from the hierarchical division of proposed building regions by rectilinear lines using the *Binary Space Partitioning (BSP)* tree. Zabuawala *et al.* [159] extracted the initial building footprint, based on an iterative morphological filtering approach. This initial segmentation result was afterwards enhanced with color aerial imagery by first creating a combined gradient surface and then applying the watershed algorithm to find ridge lines on the surface. Guercke *et al.* [125] first detected building edges and separated them from other above-ground information using DSM data and NDVI, then iteratively fitted a rectangle to the building contour until all building parts became rectangles. Turlapaty *et al.* [160] first obtained an initial test dataset by thresholding those samples from DSMs that did not correspond to buildings. The block-based features were then extracted from the potential building segments. Finally, these features were used for *Support Vector Machine (SVM)* classification to discriminate buildings from non-building objects in the initial test dataset. In the recent work of Partovi *et al.* [161], the process of building outline delineation started by detecting rough building contours using DSMs and refining them using high-resolution pan-chromatic satellite images. The finer building contours were then parameterized to simpler line segments which were later used to determine buildings main orientations. Final polygons representing the building's outlines were formed through adjusting the intersected and connected line segments.

Although the methodologies based on hand-crafted features have shown promising results in the past, their main drawback is that they are not robust for the large variety of shapes and appearances of buildings within remote sensing images of different scales.

### 3.1.2 Deep Learning-Based Methodologies

With a tremendous jump in development of artificial neural networks, it became possible to learn image features automatically instead of retrieving them by classical methods. Moreover, making network architectures deeper permits more abstract learning and discriminative semantic features, and obtains better classification performance compared to traditional approaches. Another big advantage of using the deep learning concept is its ability to transfer knowledge across tasks similar to human ability. It means that low-level and high-level features acquired while learning one task can be utilized to solve target problems from related domains. This is especially practical if target problems have only limited amounts of data that makes the learning process very difficult. As a result, one can directly use deep *Convolutional Neural Network (CNN)* models pre-trained on natural image data sets, e.g., *VGGNet* [162], *GoogLeNet* [163] and *ResNet* [61], to generalize to images or as a local feature extractor for combining the extracted features with feature coding techniques to obtain final classification results. Another possibility is using the pre-trained model to adapt to a new task. Numerous research works [164, 165] have demonstrated that fine-tuning the models pre-trained on images from the computer vision domain on a set of satellite images helps to obtain superior results than only directly using the pre-trained CNNs.

Recent applications have demonstrated impressive results in learning large-extent spatial contextual features for labeling high-resolution remote sensing data [166–168]. They aim to produce either a semantic segmentation of images with classes like building, road, vegetation and water [166, 167] or a binary classification of the image with a single class [168–170]. For example, Farabet *et al.* [171] assigned patch-wise predictions from a CNN with three convolutional layers and a fully connected layer to predefined superpixels which are combined into meaningful regions after applying a *Conditional Random Field (CRF)*. The approach processed each scale from the generated image pyramids separately with the CNN while sharing filter weights across scales. Mnih [170] utilized a specific patch-based architecture, where instead of the inference of a single value to classify a whole image, a dense classification patch was retrieved as a final outcome. In order to enhance the performance of the proposed algorithm, the results were processed by CRFs, because this approach improves the predictions by encouraging smoothness between similar adjacent pixels. Inspired by the patch-based approach of Mnih [170], Saito *et al.* [172] proposed to train the CNN with multi-labeled patches from visible and infrared bands. Moreover, they introduced an output function named *Channel-wise Inhibited Softmax (CIS)* into the learning process and demonstrated an improvement on the object extraction results. Vakalopoulou *et al.* [164] used AlexNet architecture with pre-trained weights to train it on VHR multi-spectral satellite imagery for building extraction task. Final building objects were extracted by applying the MRF on the classification results. However, due to cropping the images to small regions, the patch-based approaches introduce discontinuities on the border of the classified patches. In addition to this, the patches or predefined segments can cover only fragments of buildings and as a result, are not able to capture the entire information of an individual building.

In the field of semantic segmentation, it has become more popular to follow the idea of *Fully Convolutional Networks (FCNs)* proposed by Long *et al.* [59]. The FCNs are the type of CNNs which consist of convolutional and pooling layers plus activation functions. Thus, there are no *Fully Connected (FC)* layers in this type of network. As a result, they can efficiently compute spatially explicit label maps and are independent from the input size. In the context of building footprint extraction, Yuan [173] proposed a type of FCN architecture where the outputs of each stage of the network were up-sampled, stacked together, and fed into a convolutional layer with a filter of size $1 \times 1 \times n$ (where $n$ is the number of stacked feature maps). A prediction map was generated, where, in contrast to our methodology presented in Chapter 4, the values of pixels corresponded to their distance to the building boundaries and not to class 0 or class 1. Going even further, Zuo *et al.* [174] proposed a *Hierarchically Fused FCN (HF-FCN)* which approached a similar strategy as Yuan [173] by hierarchically fusing the information from the multi-scale receptive fields of the network built on the basis of VGG-16 architecture. Maggiori *et al.* [41] converted the fully connected network proposed by Mnih [170] to FCN and generated a building mask out of RGB satellite imagery by first, training the network on possibly inaccurate *OpenStreetMap (OSM)* data, and then refining the model on a small amount of hand-labeled data. The major differences from the methodology we are proposing in Chapter 4 are that the network architecture was much shallower and did not produce the output map of the same size as the input image. On the other hand, like in our methodology, the approach combined coarse and fine information from different layers in order to produce more detailed results. Moreover, Maggiori *et al.* [175] investigated the network built on the basis of FCN proposed by Long *et al.* [59] combined with a *Multilayer Perceptron (MLP)* on top of it. However, MLP is a FC network applied to every pixel individually and it significantly enlarges the number of parameters in the neural network. Recently, Yang *et al.* [176] performed a comparison study of four state-of-the-art CNN models, namely the branch-out CNN, the FCN, the *Conditional Random Field as a Recurrent Neural Network (CRFasRNN)*, and the *SegNet* [177], for buildings extraction problem across the entire United States continent using aerial images provided by the *National Agriculture Imagery Program (NAIP)*. The result showed that the performance of the SegNet architecture left behind other models on both the F1-score and *Intersection over Union (IoU)*.

However, the FCNs and other convolutional encoder-decoder models, like the SegNet or *DenseNet* [178], only apply some layers in the generation process of final output discarding the fine details. Therefore, Ronneberger *et al.* [179] presented an *UNet* approach for recovering high-frequency information. The UNet architecture employs "up-convolution" operators rather than pooling operations to concatenate the correspondingly cropped feature maps with higher resolution features from the encoder part of the network to better learn the representations for the next convolutions. Nevertheless, the classical UNet implementations has two main limitations: 1) the parameters on both sides of the bottleneck layers are updated before the intermediate layers. This makes intermediate layers less powerful in terms of semantic representations [180]; 2) the sparsity applied in the intermediate features restricts the generalization performance. To

overcome these limitations and make the UNet architecture applicable to remote sensing problems, Wu *et al.* [180] proposed an U-shape network with multi-constraints that are computed for specific layers between the prediction after applying the binary cross entropy and the corresponding ground truth. In this way, parameters are updated using multi-constraints in each iteration that eliminates the bias in a single constraint. This leads to better building extraction results in terms of evaluation metrics compared to the original UNet.

There were also many attempts to investigate the potential of *Generative Adversarial Networks (GANs)* [65] in the remote sensing domain. Isola *et al.* [181] tried to generate a mapping function to convert a satellite photo into a map and vice versa. Merkle *et al.* [182] presented promising methodologies for synthetic SAR images generation based on optical images. Marmanis *et al.* [183] proposed to use GANs for artificial SAR images generation in order to increase the training dataset.

In our earlier work [184], we demonstrated that artificial neural networks originally generated for spectral imagery processing were also applicable for height information interpretation. We presented a four-layer FC neural network for building footprint extraction from *normalized Digital Surface Models (nDSMs)*. This approach was able to extract the complete building footprints to a high degree of accuracy. But the computation of such network was heavily influenced by the FC layers and the level of details, which directly depended on the patch size. In continuation with this work, we proposed methodology in [185] to use a deep learning framework FCN8s developed by Long *et al.* [59] for segmenting the buildings from nDSM data. In contrast to our methodology presented in Chapter 4, in work [185] we copied the nDSM three times and initialized the network with the model pre-trained on RGB images. However, there is no influence on the final result, as the elevation information has different statistics in comparison to spectral information and thus requires different feature representation.

An important milestone for semantic segmentation of remote sensing images with deep learning are multi-stream architectures that learn separate convolution layers for different data modalities. A study of Lagrange *et al.* [186] showed that combining the spectral image with a DSM is essential for retrieving some specific classes. They fine-tuned separate CNNs related to the different image bands and afterwards combined them using a SVM. A further development of deeper networks and the late fusion of the spectral and height information was investigated in the work of Marmanis *et al.* [60]. This work is most closely related to the one presented in Chapter 4, motivated by the interaction of multi-source information and integration of more detailed information from earlier layers to top. The difference from our work is an ensemble-learning of the developed model which is a naive averaging after training the model with different initializations. In [60], the authors also engaged in a gradual training which does not guarantee improvements of final results. A similar architecture strategy was approached by Sherrah [187]. In contrast to many deep learning architectures, this work presented a novel no-downsampling network to maintain the full resolution of the imagery at every layer in the FCN. This was achieved by using the "à trous" algorithm [188] which removed the pooling layers that caused the down-sampling effect. The fusion was done in the

fully connected layers. However, the fusion at this point did not lead to significant improvement. Also, the authors did not up-sample the resulted output image from the network but used bilinear interpolation afterwards to achieve the same size as the input image. Both works Marmanis *et al.* [60] and Sherrah [187] advocated to use pre-trained networks for the spectral channels, but trained the network for height channel from scratch. Audebert *et al.* [189] investigated the hybrid encoder-decoder architecture from Badrinarayanan *et al.* [177] for dealing with diverse data sources by concatenating the intermediate feature maps of separately trained dual-stream architecture and feeding the merged results to a three-convolution layers network. They also introduced multi-kernel convolutional layers in the decoder part to aggregate multi-scale information while up-sampling. Although their fusion network is similar to the one presented in Chapter 4, the main difference is the additional combination of the output from the fusion network with average scores of the two independent branches. In our case, the fusion is supposed to correct errors within one fusion network without additional concatenations by giving more weight to the activations of the most suitable information among complementary sources. Moreover, the presented architecture, in contrast to ours, does not have any "skip" connections which allow the decoder to recover important details that are lost due to the down-sampling in the encoder. Another difference to our work is the data they used. The addition to spectral image in this work was a composite image consisting of DSM, nDSM, and NDVI information. Because NDVI is a good indicator for vegetation, the authors believed that this kind of auxiliary information helps to improve vegetation detection. However, because the components of the index calculation (the *infrared (IR)* and red channels) were already given to the network as input, the network was capable of distinguishing the vegetation itself. Another reason not to take NDVI into account, at least for the building detection, is that we do not need a precise vegetation prediction but only building discrimination from other above-ground objects.

Xu *et al.* [190] not only used the multiple data as input to the neural network, they also joined the UNet and the ResNet frameworks into one for building segmentation from VHR aerial images. The input to the network combines RGB and DSM as well as hand-crafted features including NDVI, nDSM and the first component of *Principal Component Analysis 1 (PCA1)* extracted from *color infrared (CIR)* imagery. After training the proposed architecture for the pixel-wise binary task, a guided filter is employed to refine the resulted prediction maps. This filter mainly emphases the pixels on building boundaries which at the same time eliminates the salt-and-pepper noise. In contrast to Wu *et al.* [180], Xu *et al.* [190] did not incorporate multi-constraints into UNet architecture.

In the context of urban scene understanding, the DSM is not only the data source that can provide extra knowledge. Recently, efforts have been made for joint edge detection and semantic classification. Marmanis *et al.* [191] present an end-to-end ensemble of CNNs for semantic segmentation with an explicit awareness of semantically meaningful class boundaries. The boundary detection significantly improves semantic segmentation results and the overall accuracy achieved more than $> 90\%$ on the ISPRS Vaihingen benchmark. Hu *et al.* [192] investigate the fusion of spectrum information of hyper-spectral image and the scattering mechanism of *Polarimetric Synthetic Aperture Radar*

*(PolSAR)* data. They propose a novel architecture which fuses two separated streams in a balanced manner. Because space-borne remote sensing videos are becoming essential resources for remote sensing applications Mou *et al.* [193] propose to fuse multi-spectral images and space videos for spatio-temporal analysis to achieve a fine-resolution spatial scene labeling map. In our recent work [194], we also investigated the deep learning-based data fusion in the context of building footprint extraction. We not only adapted the classical U-shape network, but also converted it into the multiple stream network which join the spectral images with different spatial resolution together with the height information through an end-to-end learning procedure. This work [194] is based on methodology presented in Chapter 4.

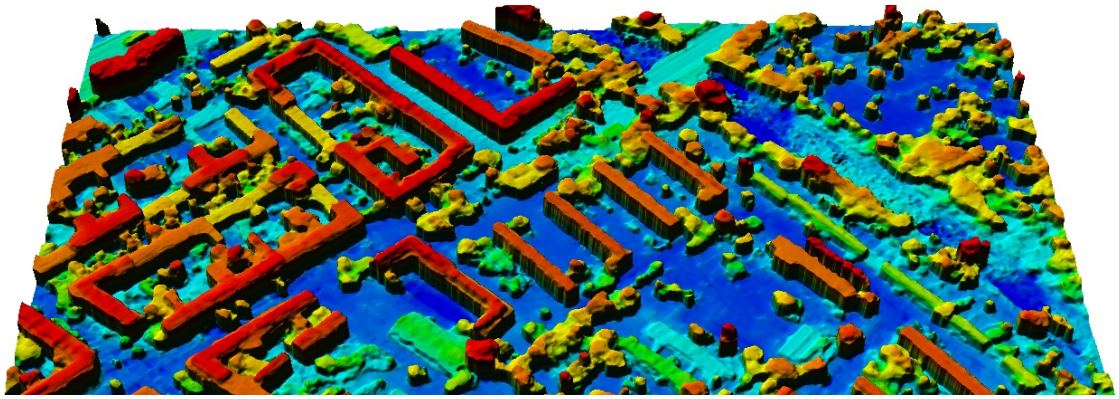
In Section 3.3.1, the motivation for our work regarding building footprint extraction from VHR satellite data published in paper [3] is given in detail.

## 3.2 Building Shape Refinement for Elevation Models

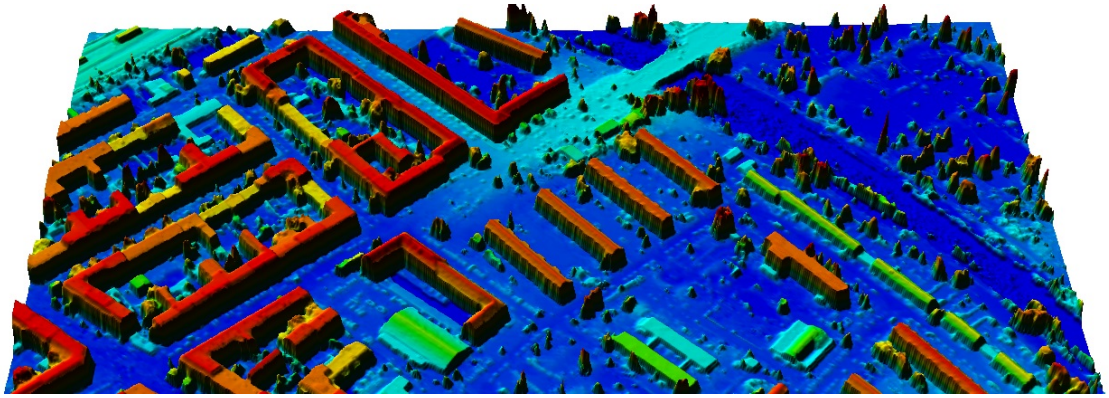
A DSM is a valuable data source for semantic information extraction as well as geometry reconstruction of various terrestrial targets from images. Building structures are important objects among them. Their reconstruction is one of the most challenging, but important tasks for many remote sensing applications. The task is complicated because during the DSM generation with stereo image matching approaches, some unwanted failures in building geometries may occur due to low-resolution, applied interpolation techniques, temporal changes or matching errors. Moreover, the urban environments with densely located buildings surrounded by vegetation may also cause uncertainty on building edges. Hence, these DSMs need to be refined either manually or automatically to be more useful for remote sensing applications. The examples of such low-quality space-borne DSMs together with desired high-quality DSMs from LiDAR and CityGML data, depicting a complex urban scene, are illustrated in Figures 3.2a, 3.2b and 3.2c, respectively.

### 3.2.1 Filter-Based Approaches

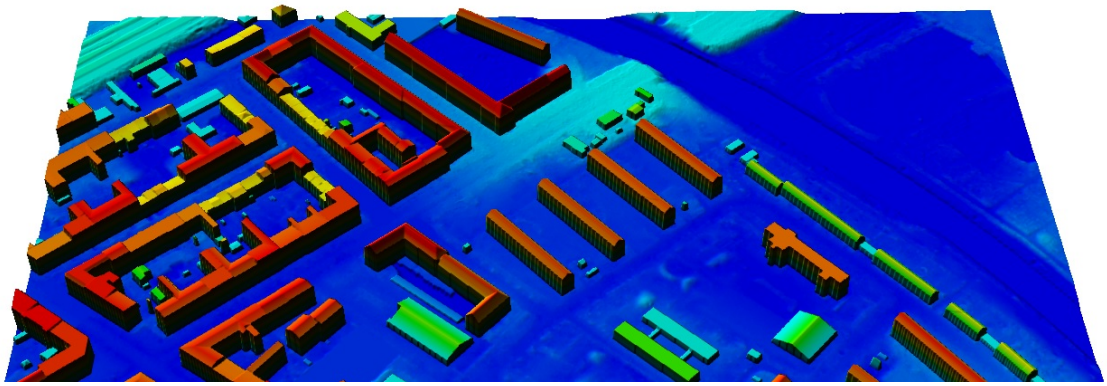
In the published literature, very few of the proposed approaches consider a photogrammetric DSM enhancement of urban areas. Generally, DSM outliers tend to be identified by assuming that elevation surfaces are continuous and regular, i.e., without sudden changes between neighboring points [195]. Thus, the analysis of values within nearest neighborhood may be a reasonable strategy for inconsistency detection. The pioneer methodologies investigated DSM refinements by applying filtering techniques. Felicísimo [196] and López [197] introduced the methodologies for blunder detection using different statistical criteria to identify anomalous values in DSMs. Felicísimo [196] utilized local deviations of height values within each pixiel's neighborhood. If an extreme value was detected, it was replaced by the elevation value estimated from its neighbors. However, the method of outlier correction, which is only based on the neighboring region, is inappropriate as it cannot be assumed that the elevations of neighboring pixels are



**(a)** Photogrammetric DSM



**(b)** LiDAR-DSM



**(c)** DSM from CityGML data

**Figure 3.2:** Visual comparison of DSMs generated with (a) *Semi-Global Matching (SGM)* [17] approach from six pan-chromatic Worldview-1 images, (b) LiDAR point cloud and (c) *City Geography Markup Language (CityGML)* data. All DSMs were produced with a resolution of 0.5 m. The depicted area is located in Berlin, Germany and represents a 0.25 $km^2$ coverage. DSMs are color-shaded for better visualization.

correct. López [197] proposed to improve the method of searching anomalous values within the pixels neighborhood by employing *Principal Component Analysis (PCA)* to locate uncorrelated patterns. The PCA approach controlled the spatial autocorrelation in the data, and as a result avoided systematic error. The obtained results demonstrated the reduction in *Root Mean Square Error (RMSE)* of up to 8% although the rate of the detected outlier comprised less than 1% of pixels in the DSM.

Wang [198] proposed a 2D Kalman filtering approach to reduce random errors in elevation surface by using a prediction model. The method mainly took into account the slope information of the previous pixel together with the distance between points to compute the slope value of the next pixels. Moreover, as the accuracy of the estimated elevation was improved with the number of processed points, the authors suggested a 2D Kalman filter to process the same elevation data twice with different orientations.

Walker *et al.* [199] later investigated the capability of a Gaussian filter to reduce a high frequency noise in photogrammetric DSM. It was observed that a $3 \times 3$ filter already improved the visual appearance of the elevation model. Increasing the size of the filter led to the loss of fine details.

Arrell *et al.* [200] applied a spectral filtering based on *Fast Fourier Transformation (FFT)* frequency data to detect artifacts associated with gross error. Transforming the elevation into the frequency domain led to the possibility of clearly distinguishing the noise because it was found to be heterogeneous through the surfaces. Spectral filtering demonstrated superior and more precise results compared to mean spatial filters as it avoids strong data smoothing.

Although the filtering approaches can reduce the noise in DSMs, their major drawback is the smoothing effect, which dramatically influences the correct steepness of building walls.

### 3.2.2 Interpolation-Based Approaches

An alternative method for DSM improvements is to first detect uncorrelated patterns from original sampling points and then, using remaining sampling points, apply interpolation techniques to construct the DSM. Examples of popular interpolation methods include *Inverse Distance Weighting (IDW)* and kriging interpolations [201, 202], which reduce point densities, but maintain a satisfactory DSM estimation; spline-based methods [203], which produce smooth surfaces, but with fewer recognizable characteristics; and hybrid methods that combine linear and non-linear interpolation [204].

Some investigations [205, 206] demonstrated that the multi-quadric method is one of the best methods for elevation model interpolation. Franke [206] made a comparison of classical interpolation methods and found that in terms of storage, speed, accuracy, complexity and the visual appeal of the resulting surface, the *multi-quadric (MQ)* method performs very well. The authors adapted the MQ method further and developed a *least squares (LS)* MQ method which employs less knots than the number of given sampling points. The LS MQ was able to recover the spatially coherent signal while simultaneously removing the noise by dividing the data into signals and noise components. However, in the instance of strong noise, the method does not work properly.

Shi *et al.* [204] proposed a hybrid interpolation method which combined bilinear and bicubic methods for DSM construction. The results demonstrated that the method is effective. However, in this study the assumption was made that the input data are error-free and that the main error source in the interpolated DSM was coming from the interpolation method itself.

Recently, Chen *et al.* [207] proposed a MQ-based interpolation method for DSMs construction with decreased impact of outliers. Namely, *Least Absolute Deviation (LAD)* MQ was constructed as the objective function based on the least sum of absolute deviation. It incorporated two independent procedures, i.e., the knot selection with a space-filling design and solution of the system with linear equations using LAD. Compared to classical MQ, the LAD MQ produce more accurate results when sample points are subject to strong noise. However, LAD MQ cannot entirely avoid the influence of outliers and it is difficult to define the optimal number of knots. It also requires vast memory storage.

In general, although interpolation-based approaches achieve some progress towards DSM refinement, they are not suitable for areas with strong surface discontinuities like urban areas, as this leads to lost sharpness in building edges [2].

### 3.2.3 Incorporation of Auxiliary Data Sources

The process of integrating multiple data sources in order to produce more accurate and consistent results is well known in the remote sensing area. It is functional, as data from multiple modalities, though they can be inhomogeneous in terms of spectral, spatial and temporal characteristics, still represent the same physical environment. As a result, the combination of modalities originating from different sources is oriented to their cooperative interactions that leads to an enhancement of data quality and decision-making. Therefore, data fusion technique can be also applicable for the DSM refinement task.

Several researchers proposed to use already existing DSMs to improve the quality of newly derived ones. For example, Milledge *et al.* [208] introduced the idea of using an old DSM to eliminate errors in stereo matching. The new DSM was mainly compared to old data to find the regions where the errors are higher than the pre-defined value. This knowledge was integrated into the stereo matching algorithm and helped improve the new DSM up to 50 %. Karkee *et al.* [209] proposed to register *Shuttle Radar Topographic Mission (SRTM)* DSMs and *Advanced Space-borne Thermal Emission and Reflection (ASTER)* DSMs for identifying voids in both DSMs. The identified voids were then filled by employing an erosion technique which used elevation, slope, and aspect information from these DSMs.

Some methodologies later proposed to additionally utilize spectral images for dealing with outliers, mismatches, and erroneously detected occlusions in DSMs, as for instance, the spectral data contains accurate information about object boundaries or texture. For instance, Krauß *et al.* [210] transferred segmentation information from stereo satellite imagery to the DSM and, from statistical analysis and spectral information, performed object detection and classification. As a further step, this information is used to refine

the DSM. The work of Poli *et al.* [211] also investigated surface model enhancement by fusing the DSMs and pan-chromatic VHR satellite scenes through a hierarchical image partitioning. The method consisted of several steps. The orthorectified VHR images were segmented with alpha-omega connectivity [212] and overlaid the initial DSM. The statistics of the heights of the points grouped in each segment were then calculated and used for a refined DSM construction. Heights of each segment for new DSM were mainly computed as the minimum, mean or maximum values of the initial DSM heights and assigned as a single constant value to each segmented region. In their further work, Poli *et al.* [213] proposed a more advanced approach which was based on the radiometric information and also took the information from the initial DSM into consideration. The initial DSM was partitioned in concave, convex and flat regions for computing the height information for a new DSM depending on the type of segments. The method is limited to the case of homogeneously textured complex buildings. Moreover, the method is able to refine building shapes only to the *Level of Detail (LoD)* 1.

The process of combining different knowledge towards DSM improvements demonstrates tremendous potential, since the missing or incorrect information can be revised by the complementary information. However, such auxiliary data sources are not always available for some areas. This is the main drawback of methodologies which are based on utilizing existing DSMs for improving the new ones. Approaches based on segmentation knowledge integration into the refinement task feature some problems when small segments vanish or too large segments fuse together different type of scene information, e.g., roofs, walls and shadow areas on the street [210]. Since an average height is calculated and assigned as the segments height, the incorrect surface segments appear on resulted DSMs. As a result, a better way of incorporating spectral information into the DSM refinements process needs to be found.

### 3.2.4 Object-Oriented Refinement

In the last 10 years, researchers have paid more attention to not only removing noise and inconsistency errors from DSMs, but also keeping the forms and shapes of building constructions more accurate.

For instance, Canu *et al.* [214] performed the depth discontinuities refinement by interpolating flat surfaces on areas segmented as buildings. For this purpose, they first segmented the DSM into homogeneous regions and attributed them to one of two classes: building and ground. After applying mathematical morphology on the segmented building regions to make them more compact, the final refined DSM was obtained by fitting the polygons to the predefined building regions.

In the work of Vinson *et al.* [215] and followed by Cohen *et al.* [216], the approaches for rectangular building detection in DSMs were further investigated. They first extracted the above ground segments from DSM. The rectangle forms were then fitted to the extracted above ground segments to model building shapes. Lastly, the estimated rectangular shapes were used to improve the DSM quality. Following this idea, Sirmacek *et al.* [2] extracted potential building segments through thresholding the nDSM and applying a box-fitting algorithm to extract 2D building shapes. The 3D building

forms were then refined by sharpening building walls using the information from the detected building shapes and smoothing the noise in building rooftops. Sirmacek *et al.* [217] considered Canny edge information from spectral images in the procedure of fitting a chain of active shape models to the input data to further improve the detection of complex buildings. Although they achieved better results in terms of building footprints compared to [2], only one single height value was assigned to each building shape. Both methodologies were only limited to the detection and enhancement of rectangular buildings.

Although these strategies of fitting the predefined shapes led to more detailed and sharper elevation models production with more accurate building shapes, they failed if the building structures were more complicated. Moreover, the refined building shapes were only reconstructed as flat roofs.

### 3.2.5 Deep Learning-Based Approaches

Most of the conventional approaches which investigate DSM improvements are still based on assumptions, such as a specific shape of man-made constructions, their equal height or spectral information within one object polygon. However, the urban city planning does not follow a specific pattern in building construction. Therefore, the generation of the approach which is able to reconstruct an accurate elevation with true silhouettes of terrestrial objects on it without taking into consideration the pre-defined knowledge or the assumption is of great interest. In this section, we review the deep learning-based methodologies which have been initiated and that are able to achieve promising results for the depth information reconstruction task.

#### 3.2.5.1 Depth Image Reconstruction from Single Image

As deep learning techniques have emerged over the past 10 years, new approaches for remote sensing image processing have achieved significant breakthroughs. However, most of these approaches work with spectral imagery, while depth image processing still has not been well investigated using these new techniques, especially, when it comes to satellite data.

In contrast with computer vision, several attempts have been made to generate, restore, and enhance depth images using CNNs. A first attempt at applying CNNs for depth estimation was done by Eigen *et al.* [218] and followed by Eigen *et al.* [115], where the authors performed coarse-to-fine learning of two and three convolutional networks in stages, respectively, to transform a monocular color input image into a geometrically meaningful output image at a higher resolution. Tian *et al.* [219] trained a CNN on patches cropped by a large window centered at each pixel of raw RGB image. The authors explained the purpose of large window as a requirement for each pixel to get a wide enough contextual information from the surrounding area. Li *et al.* [220] tackled the problem of depth prediction from single color image by regression in a CNN coupled with a CRF which played the role of post-processing refinement step. Applying the proposed CNN, the method learned the mapping from multi-scale image patches to

depth at the super-pixel level. The super-pixels were then refined to the pixel level by the hierarchical CRF. Unlike the above method, Liu *et al.* [221] explored the strength of an end-to-end deep structured CNN which learns the unary and pairwise potentials of a continuous CRF enforcing local consistency in the output image. In contrast to standard methods, it inputs an image consisting of small regions of homogeneous pixels to the network. The method can also work with single pixels, but it is computationally inefficient. It delivers predictions with sharper transitions compared to previous studies, but with a mosaic appearance.

Zhu *et al.* [222] trained a model for depth estimation consisting of two parts: a pre-trained VGG [36] and two fully connected layers of their own design. This network only allows a gradient descent optimization algorithm for five convolutional layers starting from the end. Although these methods are able to generate depth images relatively close to the ground truth, the sharpness of the object edges and their appearances in the image are very coarse. Jeon *et al.* [223] aimed at solving a problem similar to ours regarding depth image enhancement. They explored a multi-scale Laplacian pyramid-based neural network and structure preserving loss functions to progressively reduce the noise and holes from coarse to fine scales.

The development of GANs [65] helped to achieved impressive results in high-quality image generation tasks. There have already been many studies on the mapping of images between different domains, such as black and white images into color, or satellite images to maps [67]. Recently, some works proposed the learning of object representations in three-dimensional space based on different variations of GAN architecture. These methods typically use autoencoder networks [224, 225] combined with a generative adversarial approach to generate 3D objects. Wu *et al.* [226] modeled 3D shapes from a random input vector by using a variant of GAN with volumetric convolutions. Although the algorithm produces 3D objects with high quality and fine-grained details, the final grid has limited resolution. Rezende *et al.* [227] introduced a general framework to learn 3D structures from 2D observations with a 3D-2D projection mechanism. However, the proposed projection mechanism minimizes the discrepancy between the observed mask and the reprojected predictions either through a learned or fixed reprojection function. Recently, Yang *et al.* [228] proposed an automatic completion of 3D shapes from a single depth image using GANs. The architecture combines *conditional Generative Adversarial Networks (cGANs)* [66] with autoencoders to generate accurate 3D structures of objects. The method learns both local geometric details and the global 3D context of the scene to infer occluded objects from the scene layout. However, designing a network that can efficiently learn both components is a non-trivial task [229]. All of these studies learn a single object reconstruction based on existing libraries of individual objects and are able to produce a probability for occupancy at each discrete position in the 3D voxel space. Yet the computational and spatial complexities of such voxelized representations significantly limit the output resolution.

In contrast to the computer vision field, height image generation from single input data has so far been rarely addressed in the remote sensing community. Mou *et al.* [230] tackled a problem of height prediction from a single monocular remote sensing image using

an end-to-end fully convolutional-deconvolutional network architecture *IM2HEIGHT*, encompassing residual learning. The authors demonstrated that skip connections are very important for remote sensing tasks because they keep detailed boundaries and edges for miniscule objects representations in remote sensing images. The method was developed, different from our task, for processing aerial images which in general are more accurate and detailed compared to satellite data. Moreover, their approach is able to reconstruct only nDSM. Ghamisi *et al.* [231] proposed a cGAN-based *IMG2DSM* network for simulating the DSMs from single optical images consisting of near-infrared, red and green bands. Their generated output was a DSM with three channels representing the same DSM copied three times. The training was done on high-resolution aerial images with *Ground Sampling Distance (GSD)* below 10 cm. Their experiment showed that the presented network was able to generalize well on the test data resembling the spatial-spectral information to the training dataset, but produced relatively low results on a new region which was not covered by training data and generated by a different acquisition platform.

#### 3.2.5.2 Depth Image Reconstruction from Multiple Data Sources

As mentioned above, the problem of continuous values prediction in remote sensing based on learning techniques started to evolve only recently. As a result, the idea of classical remote sensing approaches which integrates multiple data sources to compensate the lack of knowledge from a single image also started to spread within deep learning-based methodologies for surface models reconstruction. However, only a few of them have been recently developed.

Costante *et al.* [232] developed a CNN-based method which reconstructed a monocular*Digital Elevation Model (DEM)* from intereferometric images. The amplitude and phase components of complex SAR images were mainly the inputs to the Encoder-Decoder architecture. The output was a DEM re-projected in the radar coordinates. The network was trained by minimizing the objective that was the pixel-wise linear RMSE. The network was able to estimate the elevation statistics resembling the ground truth. However, a significant smoothing effect was also present in the reconstructed elevation model.

Paschalidou *et al.* [233] explored multi-view geometry constraints from multi-view aerial images to correlate the physical process of perspective projection and occlusion based on a learning approach. More specifically, a CNN architecture responsible for estimating surface probabilities from correlated nearby images was integrated with MRF that aggregated the physics of perspective projection and occlusion across all viewpoints.

#### 3.2.5.3 Depth Image Reconstruction in a Multi-Task Context

Image analysis tasks, whether classification, semantic segmentation, or regression, are related to each other and can feature some aspects that are in common. As a result, one task can help to learn other tasks. As it has been already reviewed in Section 2.2.6,

the multi-task learning has been successfully integrated in many computer vision applications [104–107].

In remote sensing, the method proposed by Srivastava *et al.* [234] is the only known multi-task deep learning-based approach developed for semantic segmentation maps prediction, as well as nDSM generation from single monocular images. The authors used a joint loss function for CNN training, which is a linear combination of a dense image classification loss and a regression loss responsible for DSM error minimization. The model is trained by alternating over two losses. However, the major drawback of the proposed method is that in the training phase the network requires pixel-wise labeled segmentation masks as input, which are not widely available.

## 3.3 Contributions of this Dissertation

The envisioned objective of this thesis is enabling Earth scene understanding with focus on building information extraction and reconstruction problems by applying deep learning techniques. In achieving this objective, a comprehensive study is performed to explore the individual and cooperative influence of multiple VHR satellite images together with the impact of neural network architecture on the overall quality of declared task and decision-making.

The availability of accurate building footprint maps is vital for many remote sensing data analysis tasks. Different methodologies built on radiometric or geometric information have already been extensively explored in the past for the building extraction problem. However, most of them are limited to shape or color assumptions. The recent deep learning techniques overcame those constraints and proved to be robust for a large diversity of building forms. But there are still some open questions: what network structure is the best or what type of data benefits the solution the most? The building extraction task appears to be highly challenging because the building size from satellite perspectives can be miniscule, which leads to difficulties by extracting the building's contour details.

There is a growing interest in generating high-quality DSMs from satellite stereo data which cover large areas and feature detailed shapes of terrestrial objects. Many attempts have been made to achieve this goal: from manual corrections and filtering to object-oriented fitting algorithms (see Section 3.2). Unfortunately, these methods cannot achieve the desired results regarding object shapes due to smoothing effects or limited prior knowledge about the object form. With deep learning techniques introduction, a several of methodologies appeared which were able to reconstruct continuous values within an image, i.e., to perform a depth image regression task. However, none of them were focused on the generation of large-scale high-quality DSMs with detailed building shapes.

The contributions of this dissertation regarding the two problems addressed are summarized in the following:

### 3.3.1 Building Footprint Extraction

To achieve a reliable building mask extraction, a *Fused-FCN*4 network is proposed which automatically learns the complementary knowledge from multiple data sources, mainly RGB, PAN, and nDSM, and allows detailed boundary extraction by including additional comprehensive high-frequency information from earlier network layers. The late fusion of parallel streams related to each independent data source allows the network to first learn corresponding features from each input image and then decide at the end from which data source it benefits more or less for pixels classification. The process of learning from multiple data sources makes the neural network more confident in making decisions. Although both RGB and PAN images provide the intensity of solar radiation that is reflected from different surfaces, each of them assists the network differently, compensating the lack of information from an individual data source. The developed model is not limited to a special region and is able to generalize to scenes unknown for the model.

Chapter 4 is dedicated to these contributions.

### 3.3.2 Digital Surface Refinement with Focus on Building Shapes

To obtain a realistic ground surface model, a new height image formation neural network is proposed which is able to automatically reconstruct either LiDAR-like DSMs or LoD2-like DSMs, depending on the required task or data availability, from low-quality photogrammetric DSMs with a focus on exact building shapes. The implemented network is capable of not only improving the building ridge lines and roof surfaces, but also reducing or totally eliminating the vegetation if required.

In continuation, the influence of multi-task learning on several remote sensing tasks is investigated with the DSM refinement problem being among them. A two-stream cGAN-based network is developed, where the generative part consists of two parallel networks specific for height image generation and roof type classification joined via two frontal convolutional layers for transferring the knowledge from one domain to another.

Inspired by the advantages of multiple data sources fusion, which proved to be beneficial for building footprint extraction in Chapter 4, a *Hybrid-cGAN* model is introduced for further DSM improvement. The model merges the height information from photogrammetric DSMs together with the intensity from PAN images which provides much better knowledge about building boundaries. To decrease the number of parameters in the system and make the cooperation of intensity and height information even stronger, the strategy of earlier fusion is applied.

All proposed models are universal because they can generalize to new cities with different topological information compared to the training dataset. This ability is highly import for remote sensing data processing because good-quality data is not available for every region of the Earth.

Chapter 5 is dedicated to these contributions.

Chapter 4

# Building Footprint Extraction from VHR Remote Sensing Images Combined with Normalized DSMs using Fused Fully Convolutional Networks

This chapter describes a novel approach for the building footprint extraction problem from *Very High-Resolution (VHR)* satellite imagery by incorporating both the spectral and height information within one end-to-end deep learning based framework. It represents the following peer-reviewed journal paper:

[3]: **K. Bittner**, F. Adam, S. Cui, M. Körner, and P. Reinartz, "Building Footprint Extraction from VHR Remote Sensing Images Combined with Normalized DSMs using Fused Fully Convolutional Networks," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 8, pp. 2615–2629, 2018.

## 4.1 Problem Statement

Since the launch of the first satellite for Earth monitoring, the development of different sensors significantly increased the availability of high-resolution remote sensing imagery, providing a huge potential for meaningful and accurate terrestrial scene interpretation. The analysis of satellite imagery involves the identification of building rooftops as one of the most challenging, but important objects among various terrestrial targets in an image. This information is useful for many remote sensing applications, such as urban planning and reconstruction, disaster monitoring, 3D city modeling, etc. A vast amount of manual work is done on interpretation and identification of targets in remote sensing imagery by human interpreters. However, it is very time-consuming and expensive to distinguish buildings from other objects and delineate their contours manually. There-

fore, there was a great number of attempts to develop methodologies to extract buildings automatically.

Some algorithms for building detection on the basis of aerial [235] and high-resolution satellite imagery [152, 236] utilize specific criteria of building appearance like the uniform spectral reflectance values [135, 145]. The main problem to be encountered in these approaches is the confusion of the building with other objects with similar spectral reflectance. Many automatic building extraction methods from multi-spectral imagery or *Digital Surface Models (DSMs)*, providing height information for a scene, define the criteria such as the shapes of relatively homogeneous buildings follow a certain pattern [154, 237, 238]. However, these methodologies are very limited, because the defined criteria work only for certain types of buildings but fail to generalize to areas with complex and heterogeneous buildings. Different data sources can provide complementary information to each other. As a result, the integration of different data sources creates the opportunity for improving accuracy and robustness of the extraction results. Therefore, recently developed methodologies apply the use of fusing data sources, such as multi-spectral images with either stereo DSM or *Light Detection and Ranging (LiDAR)* DSM rather than the use of only a single data source [158, 239]. Although many approaches have been proposed for building footprint extraction, this topic remains a complex problem for scientists.

With the revolutionary development of deep learning techniques, the definition of task-specific features is not under demand anymore for learning-based image analysis tasks. Instead, the most suitable features can be discovered automatically during the training procedure on a big dataset through the organization of multi-layer neural networks. *Convolutional Neural Networks (CNNs)* [30, 240] are one of the most successful deep learning architectures. They achieved state-of-the-art results and became the dominant approach for image understanding in computer vision. The main objective of this work is to adapt the CNNs for remote sensing imagery understanding with high accuracy. This is a challenging task since the satellite imagery is very different from usual computer vision images in a sense of size, perspective view and semantic meaning of every pixel within the whole scene.

In summary, our main contributions of this chapter cover following aspects:

- We efficiently adapt the *Fully Convolutional Network (FCN)*8s architecture developed by Long *et al.* [59] from generic everyday images to satellite images and analyze it for three different data sources: *red, green, and blue (RGB)*, *normalized Digital Surface Model (nDSM)*, and *pan-chromatic (PAN)* images.

- We augment the FCN8s with additional "skip" connection, which combines the predictions at an earlier stage with the later one, for improving the segmentation results. We name the network FCN4s and inspect the improvements on RGB, nDSM, and PAN images in comparison to FCN8s.

- Inspired by the possibility to fuse multi-source data within one deep convolutional framework, we propose a Fused-FCN4s architecture which employes a late fusion approach of three identical parallel FCN4s networks, carrying information from

RGB, nDSM, and PAN images. To our knowledge, this is the first work which applies in a direct way a deep convolutional architecture on RGB, nDSM, and PAN satellite data for building footprint extraction. Code is available at https://gitlab.com/ksenia_bittner/fused-fcn4s.

- As generalization is a key point for remote sensing applications, we demonstrate the generalization capability of the proposed network by applying it to a different urban landscape, unseen by the model before.

The remainder of the chapter is arranged as follows. The background of CNNs, their transformation to FCNs, and details of our deep network architecture are described in Section 4.2. In Section 4.3, we introduce the dataset and present implementation details and training strategies. The experimental results on two different datasets applying the proposed deep network architecture, together with their quantitative evaluation are shown and discussed in Section 4.4. Section 4.5 summarizes the chapter.

## 4.2 Methodology

### 4.2.1 Convolutional Neural Networks

CNNs are a category of artificial neural networks that have successfully been applied to visual imagery understanding. They are commonly organized in a series of layers. This hierarchy allows the network to learn multiple levels of data representation, starting from low-level features at the bottom layers, such as edges and corners, proceeding to generate coarse feature maps with high-level semantic information at the top layers. CNNs take advantage of the 2D structure of an input image by applying on it learnable 2D convolutional filters

$$y_j^l = \sigma\Big(\sum_{k \in -\frac{W}{2} \times \frac{W}{2}} w_{jk} \cdot y_k^{l-1} + b_j^l\Big) \tag{4.1}$$

which connect each neuron at level $l$ with a specially localized region of fixed size $W \times W$ from previous layer $l-1$, and takes a weighted sum over all neurons followed by some activation function $\sigma$. The $b_j^l$ corresponds to a bias. Due to the weights $w_{jk}$ being shared across all neurons for each dimension per layer, the number of free parameters is significantly reduced in the model, compared to the standard *Multilayer Perceptron (MLP)*, which differs mainly by the fact that no weight sharing takes place in this type of neural networks. Additionally, the weight sharing introduces translation equivariance [241], another desirable attribute for the network. The bias can be considered yet another weight (with $y_{i=0} = 1$). The merit of the activation function is to introduce non-linearity into the network. The most common activation function applied after each convolutional layer in CNNs is the *Rectified Linear Unit (ReLU)*

$$y_{relu}^l = \max(0, y^l) \tag{4.2}$$

which sets all negative numbers in the convolution matrix to zero and keeps the positive values unchanged. The main advantages of using ReLU in neural networks are, first, it induces sparsity in the hidden units, second, it does not suffer from the gradient vanishing problem [242].

As CNNs were originally developed for image classification problems, their goal was to predict the correct class associated with the input image. Therefore, the top layers of the network are usually *Fully Connected (FC)* layers, which merge the information of the whole image. The final layer is a 1D array and consists then of as many output neurons as there are possible classes, representing class assignment as probabilities, most often using softmax normalization on each of the neurons.

The classifier computed by the network is determined by the weights and biases parameters. To generate an optimal network classifier means to find such weights and biases which will minimize the difference between predicted values and target values. The misclassifications are penalized by a loss function $\mathcal{L}(\mathbf{x}, \mathbf{t}, \mathbf{p})$. The commonly used cross-entropy loss function

$$\mathcal{L}(\mathbf{x}, \mathbf{t}, \mathbf{p}) = -\sum_i t_i \log p(x_i) \tag{4.3}$$

avoids the problem of slowing down the learning (in comparison to, for instance, the Euclidean distance loss function) and provides a more numerically stable gradient when paired with softmax normalization [41]. Here, $\mathbf{x} = \{x_1, \ldots, x_n\}$ is the set of input examples in the training dataset and $\mathbf{t} = \{t_1, \ldots, t_n\}$ is the corresponding set of target values for those input examples. The $p(x_i)$ represents the output of the neural network for given input $x_i$. We minimize the logistic loss of the softmax outputs over the whole patch.

A standard technique to minimize the loss function is *gradient descent* which computes the derivatives of the loss function with respect to parameters $\frac{\partial \mathcal{L}}{\partial w_i}$ and $\frac{\partial \mathcal{L}}{\partial b_i}$ and updates the parameters with learning rate $\alpha$ in the following way:

$$w_i \leftarrow w_i - \alpha \frac{\partial \mathcal{L}}{\partial w_i} \tag{4.4}$$

$$b_i \leftarrow b_i - \alpha \frac{\partial \mathcal{L}}{\partial b_i} \tag{4.5}$$

The derivatives $\frac{\partial \mathcal{L}}{\partial w_i}$ and $\frac{\partial \mathcal{L}}{\partial b_i}$ are calculated by the *backpropagation* algorithm [40] commonly used in the *Stochastic Gradient Descent (SGD)* optimization algorithm in small batches for efficiency. In this model, we used SGD with momentum, an extent to the vanilla SGD method. Additional methods have been suggested recently like *ADAM* [46] and *RMSProp* [243]. Although the optimization technique is very critical in the case of training from scratch, its role is muted in the case of pre-training, because the network is hindered from rapidly changing the weights, typically by using a very small learning rate. Therefore the technique itself plays finally a less important role in the convergence. A good overview of gradient descent optimization algorithms is given by Ruder [244].

### 4.2.2 Fully Convolutional Network Architecture

In this chapter, we address a full pixel-wise binary labeling problem for building vs. non-building classes. It means that we want to give the network an image and receive an output image of the same size, with meaningful shape and structure of building footprints. The original CNNs were constructed for recognition tasks where only one label is assigned to each image. The recently developed FCNs became the state-of-the-art methodology for semantic segmentation. They are the extensions of the traditional CNN architecture, where all FC layers are replaced with convolutional layers. The advantage of this transformation is the independence of the input image size. Additionally, in contrast to the basic CNNs, FCNs do not lose the spatial information in the top layers but allow to track it back. The per-class probability maps $cl_i(x, y)$, which the FCNs generate, have a coarse resolution due to the pooling and convolution with stride larger than one operations along the network. The number of probability maps $cl_i(x, y)$ in the last convolutional layer is equal to the number of classes of the task. So, for our binary classification problem, this number is equal to two. In order to up-sample the feature maps from the previous layer, the FCNs are augmented with *deconvolution* layers. This type of layer performs a learned interpolation from a set of nearby points. The construction of the network with several deconvolution layers at its top part allows obtaining the resulted class probability maps of the same size as the input image. In our network, we initialize the deconvolution weights with a set of bilinear interpolation parameters.

#### 4.2.2.1 FCN4s Network

Applying several up-sampling layers and, as a result, bringing the classification maps to the original size, does not guarantee very detailed and accurate object boundaries in the resulting images. Long *et al.* [59] were the first who suggested to use the high-frequency information from the feature representations of the shallow part of the network, bypassing several layers of non-linear processing, and combining it using an *element-wise addition* with the output from the deconvolution layers at the same resolution. This type of structure received the name of "skip" connection and is depicted in Figure 4.1 by a long arrow in violet color. In this way, the FCN8s network proposed by Long *et al.* [59] hierarchically includes the earlier layers *pool4* and *pool3* to the upper layers of the network, adding more detailed information.

However, the FCN8s was originally created for semantic labeling in the field of computer vision, where objects are big and well separated. Remote sensing imagery, in contrast to multi-media images, is very different. First of all, due to the big difference in the *Ground Sampling Distance (GSD)*, even if the resolution of remote sensing images is high, the containing information is still very heterogeneous. It consists of many objects like trees, buildings, roads, etc. Secondly, those objects can be represented only by a small number of pixels. Therefore, it is more challenging to extract very accurate boundaries and structures from such images. As a result, we modify the FCN8s network to an FCN4s by adding yet another "skip" connection from *pool2* layer, which incor-

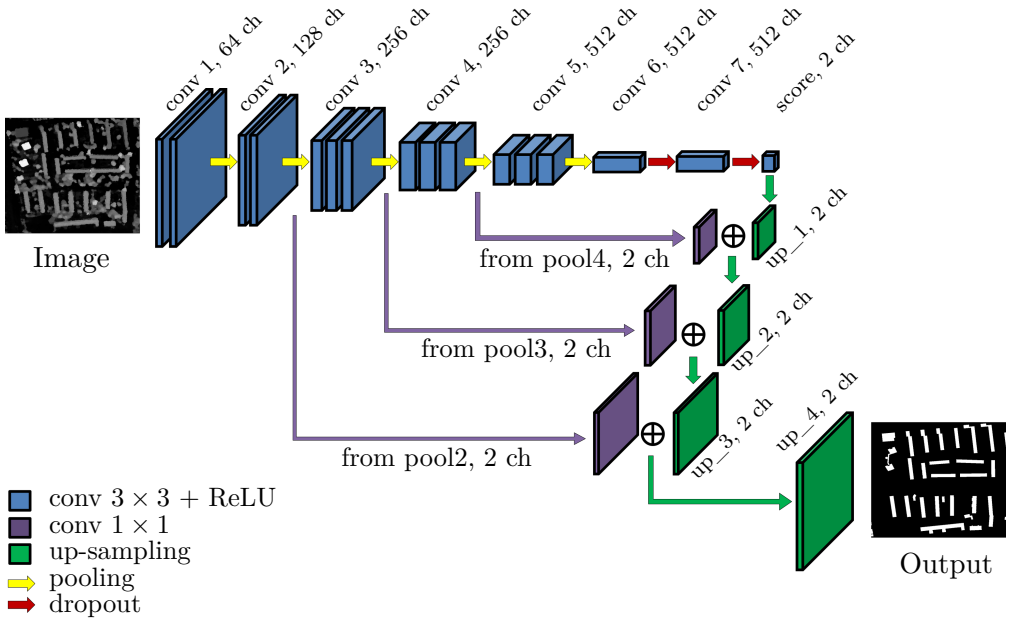



**Figure 4.1:** Schematic representation of our FCN4s architecture adapted from FCN8s [59] for building footprint extraction task from VHR remote sensing imagery. The encoder part of the network consists of $3 \times 3$ convolutional layers ( ) followed by ReLU activation functions and polling ( ) or dropout ( ) layers. The decoder part of the network includes gradually up-sampled ( ) feature maps to the desired output size concatenated with high-frequency information from the feature representations of the shallow part of the network after applying $1 \times 1$ convolution ( ).

porates even finer details, allowing more efficient building footprint reconstruction (see Figure 4.1). We also adapt the number of channel dimensions from 21 to 2. The training is done by fine-tuning the weights of the model, which is pre-trained on the large image collection of ImageNet.

#### 4.2.2.2 Fused-FCN4s Network

For the semantic segmentation task, the data used are often three-channel imagery. In this work, we propose a new network which integrates image information from RGB and PAN images, together with depth information from nDSM, as the latter provides geometrical silhouettes, which allow a better separation of buildings from the background. Besides, depth images are invariant to illumination and color variations. Since depth information and intensity have different physical meaning, we propose a hybrid network where three separate networks with the same architecture are used: We feed one part with the red, green, and blue spectral bands and initialize it with the weights pre-trained on ImageNet as mentioned in Section 4.2.2.1. The second part we feed with the PAN image converted to three-channel by copying it three times. The network is initialized the same way as the first part. The reason to use pre-trained weights for gray scale image is two-fold: First, the pre-trained networks demonstrate a strong ability to generalize

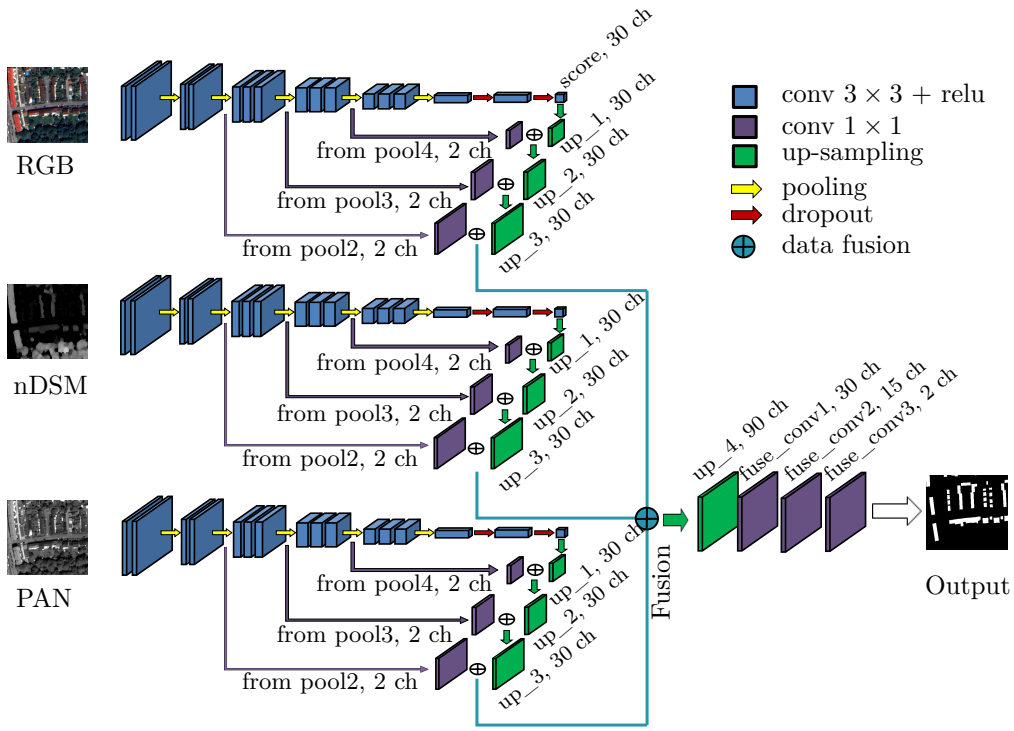



**Figure 4.2:** Schematic representation of the proposed Fused-FCN4s architecture for building footprint extraction task from multiple VHR remote sensing images. The architecture consists of three parallel networks merged at a late stage ( ), which helps propagating fine detailed information from earlier layers to higher-levels, in order to produce an output with more accurate building outlines. The inputs to the proposed Fused-FCN4s are RGB, PAN, and nDSM images. The encoder part of each single-stream network consists of $3 \times 3$ convolutional layers ( ) followed by ReLU activation functions and polling ( ) or dropout ( ) layers. The decoder part of each single-stream network includes gradually up-sampled ( ) feature maps to the desired output size concatenated with high-frequency information from the feature representations of the shallow part of the network after applying $1 \times 1$ convolution ( ).

to images outside the ImageNet dataset via transfer learning. Thus, we make modifications in the pre-existing model by fine-tuning it. Second, the PAN image has the same topology as our RGB image. So, as the visual filters from generic images can be built upon for RGB images, they are applicable for PAN images too. The third branch is fed with one-channel nDSM, initializing the convolutional layers randomly since elevation data and intensity data have different modalities and, as a result, require different feature representations. We examine two fusion strategies: a) a naïve averaging of three branches after softmax, and b) merging by the neural network itself.

The schematic diagram of the proposed network architecture is illustrated in Figure 4.2. First, it stacks the sets of spectral and height features from three streams at a very top level, but before the last *up_4* up-sampling layer as depicted in Figure 4.2. As a

result, the number of features increases three times. Second, the up-sampling is applied to bring the combined feature maps to the final size. Finally, the resulting intermediate features are sent as an input to three additional convolutional layers of size $1 \times 1$, which play the role of information fusion from different modalities, and can correct small deficiencies in the predictions, by automatically learning which stream of the network gives the best prediction result. This architecture is similar to the one presented by Marmanis *et al.* [60]. Although our implementation is based on the paper description, we made additional modifications which experimentally improve our final results. For example, the number of feature maps at the higher layers of the network is set to a larger number to allow the network to learn a wider range of features. However, we decreased the number of channels suggested by Marmanis *et al.* [60] from 60 to 30 and, experimentally, obtained better results. Besides, having a network with a huge number of parameters but rather small training set can lead to overfitting. Additionally, in contrast to Marmanis *et al.* [60], we did not find it necessary to introduce *Local Response Normalization (LRN)* to the last layer of three independent branches for spectral intensities and height before merging as the network is able itself to balance the activations between heterogeneous data. It also prevents from additional tuning of the hyper-parameters for LRN.

The network can only see a part of the image when it is centered at a pixel. This region in the input is the receptive field for that pixel and can be computed by the formula mentioned in Le *et al.* [245]

$$R_k = R_{k-1} + (f_k - 1) \prod_{i=1}^{k-1} s_i, \tag{4.6}$$

where $R_k$ is the current layer, $R_{k-1}$ is the previous layer, $f_k$ represents the filter size of layer k, $s_i$ is the stride of layer i. The receptive field of the output unit of the network that we use in this work is $404 \times 404$ pixels.

## 4.3 Study area and Experiments

We performed experiments on WorldView-2 data showing the city of Munich, Germany, consisting of a color image with red, green, and blue channels, a very high-resolution stereo PAN imagery and a DSM derived from it using the *Semi-Global Matching (SGM)* method [17]. The RGB and PAN images used in the experiment have been orthorectified, because it is important for building detection to have images where every pixel in the image is depicted as if viewed at nadir, so that occlusions do not pose a challenge.

As a ground-truth for our training, a building mask from the municipality of the city of Munich, covering the same region as the satellite imagery, is used for learning the parameters in the neural network.

In order to investigate the prediction model capacity over a different urban landscape, a second WorldView-2 dataset showing a small part of Istanbul city, Turkey, was considered. As the ground-truth for this area is not available, a building mask was extracted from *OpenStreetMap (OSM)*. However, only a few building footprints are available for

this area and the rest is missing. Therefore, a small area of around 0.5 $km^2$ was selected over the available building footprints. The rest was manually delineated.

### 4.3.1 Data Preprocessing

To perform a network training from the multiple data sources, first a PAN image with a GSD of 0.5 m was used to pan-sharpen the color image with a GSD of 2 m using the pan-sharpening method proposed by Krauß *et al.* [7].

Second, in order to obtain above-ground information only, namely to generate a nDSM, the topographical information was removed from DSM based on the methodology described by Qin *et al.* [246]. Additionally, by investigating the histogram of height data in the nDSM, it was found that there are about 0.05% outliers, which enlarge the distribution range dramatically (to 205 m height), although the majority of values lay within a much smaller range. The explanation to these outliers can be the presence of noise, due to the absence of information because of clouds. Therefore, the decision was made to remove this 0.05% of outliers and use linear spline interpolation to find the values of thresholded points. It should be mentioned that even if there are some buildings in the image higher than the selected threshold, for our binary classification task it is not very critical to loose the true height of very high buildings within the city area, since we are only interested in footprints. Another advantage of the suggested data pre-processing is the simplicity of the network training.

### 4.3.2 Implementation and Training Details

We developed our FCN4s and Fused-FCN4s models based on the FCN8s implemented in *Caffe* deep learning framework [247]. For learning process, we prepared the training data consisting of 22 057 pairs of patches, and validation data of 3358 pairs, selected from a different area. The patches cropped from the satellite image have a size of $300 \times 300$ pixels. Having a large receptive field size of the architecture leads to the question about the relative influence of boundary effects on the predictions. In our case, as the context information is available only within $300 \times 300$ pixels, each output unit of the network is influenced by the boundary effect. Therefore, to prevent artifacts and discontinuity at patch boundaries, we used an overlap of 200 pixels out of 300 (67 %) when sliding the window across the satellite image in both directions. To further improve the prediction on boundaries, all overlapping patches are stacked together first, then the final prediction is calculated as the average at each pixel. As a result, some pixels are predicted once, twice or four times like the ones at the corners. This is a commonly used approach for remote sensing problems [189].

As mentioned in Section 4.2.2.2, the two branches of the network corresponding to spectral images were initialized with a pre-trained model. This applies to the network before the fully convolutional layers. All layers above the fully convolutional layers were initialized within a range defined inversely proportional to the number of input neurons. For a layer with N neurons, the weights were initialized in the range $[-\frac{1}{N}, \frac{1}{N}]$ using uniform sampling. The network branch corresponding to nDSM data is trained from

scratch for the reasons explained in Section 4.2.2.2. We start the training process of our network with learning rate $\lambda = 0.01$ for all randomly initialized layers and $\lambda = 0.001$ for layers initialized with the pre-trained model, decreasing them by a factor of 10 for each 20000 iterations. The total number of iterations was set to 60000 with batch size of 1 on a single NVIDIA TITAN X (Pascal) *Graphics Processing Unit (GPU)* with 12 GB memory. A weight decay $\eta$ and momentum factor $m$ were set to $\eta = 0.0005$ and $m = 0.9$, respectively. All parameters were obtained empirically during investigation of the training process on the validation dataset. Within the training, random shuffling of the samples was performed before feeding them into the network.

### 4.3.3 Comparison with alternative methods

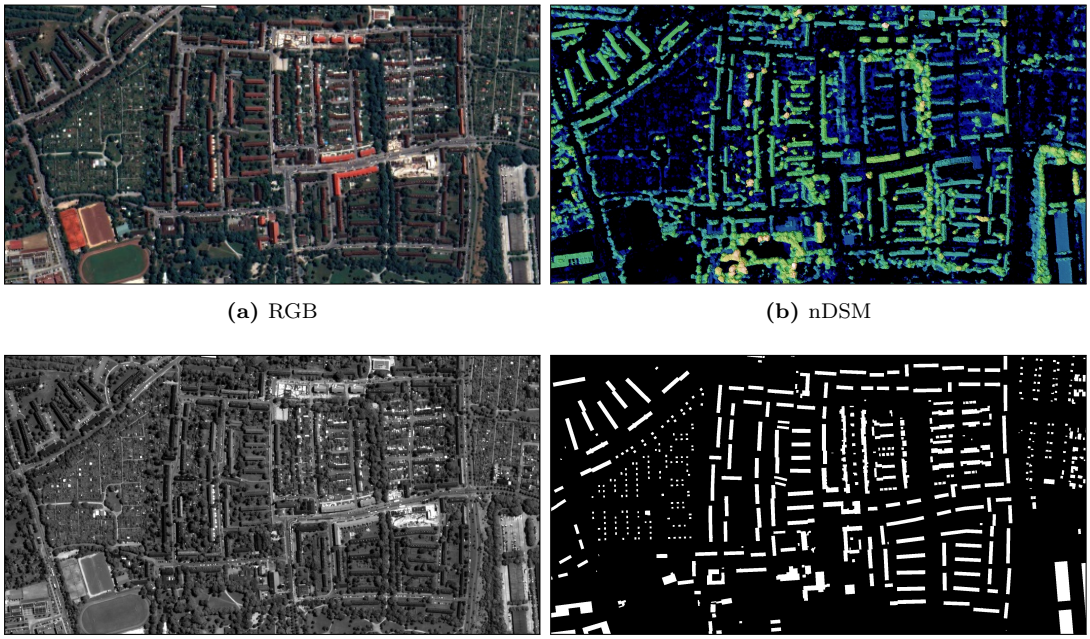
Apart from the developed FCN4s network, presented in Section 4.2.2.1, we also compare our approach with the FCN8s network proposed in [59]. We directly employed it for RGB and nDSM images, by changing only the number of outputs to 2 in order to be consistent with our binary classification task. During the fine-tuning of the FCN8s on RGB and PAN images using the pre-trained ImageNet model, the base learning rate was set to $\lambda = 0.0001$. For training the FCN8s from scratch for nDSM image, the base learning rate was set to $\lambda = 0.01$.

In order to demonstrate the advantage of end-to-end deep learning data fusion, we compare the designed architecture with naïve prediction fusion. Moreover, to indicate the influence of every data source we compare our approach with two-stream fusions: 1) RGB and nDSM; 2) PAN and nDSM.

Besides, we conduct a comparison on DSM-based building detection method proposed by Krauß *et al.* [7]. This method, first, generates a height map by distinguishing the above-ground objects from the ground level ones using nDSM. The extracted height map is used then for buildings delineation from the surroundings by applying the Advanced Rule-based Fuzzy Spectral Classification [7]. The implementation distributed by the authors is applied to the nDSM and 8-channel multi-spectral image covering the test area.

## 4.4 Results and Discussion

In the following section, the results of the considered experiments for FCN8s, FCN4s, and the proposed Fused-FCN4s, on different data sources, is presented. Their respective performance is discussed, in order to evaluate the introduced architecture for binary building mask generation, both qualitatively and quantitatively. To demonstrate the effectiveness of the models, we fed a new test dataset to the network, unseen before neither for the training nor for the validation. A test area from the city of Munich and its corresponding ground-truth image is depicted in Figure 4.3.



**(a)** RGB **(b)** nDSM

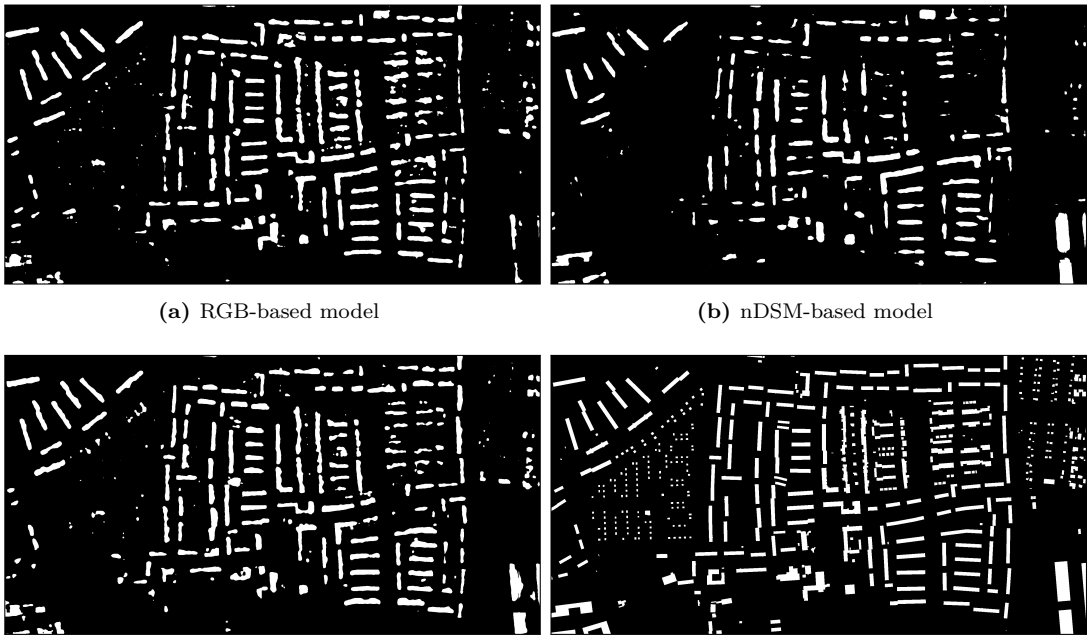
**(c)** PAN **(d)** Ground-truth

**Figure 4.3:** A test area from the city of Munich unseen neither for the training nor for the validation phases consisted of (a) RGB, (b) nDSM, (c) PAN and (d) Ground-truth building mask. The nDSM is color-shaded for better visualization.

### 4.4.1 Qualitative Evaluation

#### 4.4.1.1 FCN8s Network

The building masks generated by the FCN8s network separately on RGB, nDSM, and PAN images are presented in Figure 4.4. As can be seen from the results, the FCN8s model, generated for multi-media imagery semantic segmentation, is applicable to remote sensing data too. Moreover, not only intensity images but also the nDSM representing depth information can be used for building footprints extraction using FCNs. This has been also analyzed by Davydova *et al.* [184] and Bittner *et al.* [185]. As illustrated in the figures, the FCN8s model is able to extract the buildings from each given data source without any influence of other above-ground objects such as trees, cranes, etc. However, as it can be noticed, some footprints are better extracted from intensity images and some of them from the depth image. For example, there are two big buildings in the bottom right corner. Referring to the original RGB image in Figure 4.3a, one can see that the roofs of both constructions have a color similar to the asphalt. Therefore, we deduct that the network confuses these buildings with the road. From PAN images, the network could learn different features and as a result, enable the network to identify the area as buildings, but not optimally yet. On the other hand, from the height information provided by the nDSM, it was easier for the network to distinguish these buildings from the ground. As can be seen from the results, many buildings are missing in the building



**(a)** RGB-based model **(b)** nDSM-based model

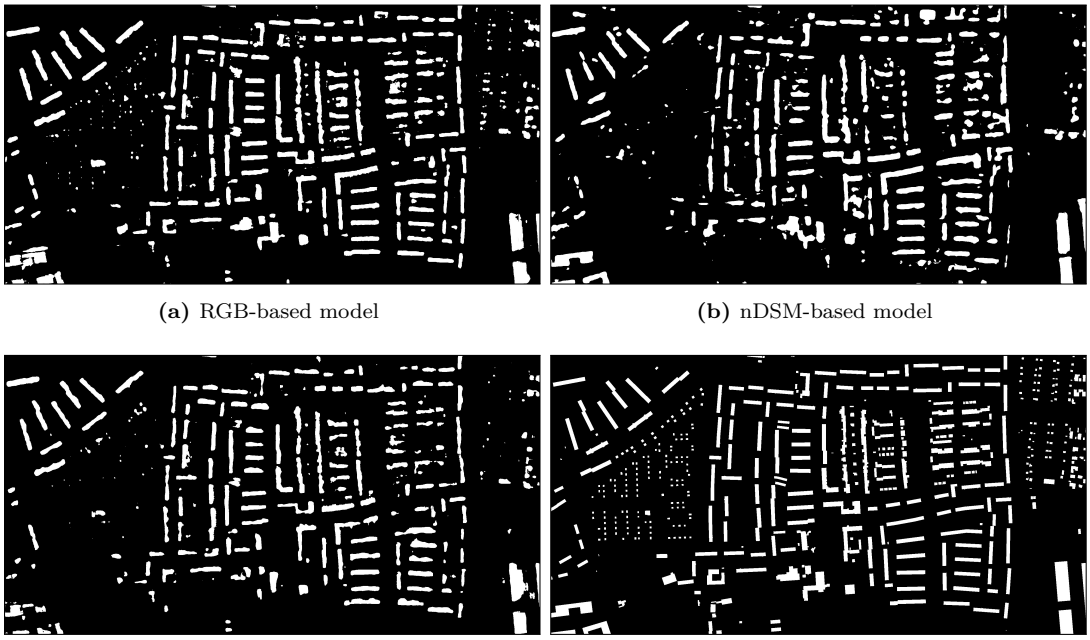
**(c)** PAN-based model **(d)** Ground-truth

**Figure 4.4:** The relative performance of the FCN8s model for building mask generation on individual data sources (a) RGB, (c) nDSM and (b) PAN images. Figure (d) illustrates the ground-truth building mask.

mask, even the one extracted from the nDSM. This can be caused by trees occluding some buildings, or inaccurate height data in these locations.

#### 4.4.1.2 FCN4s Network

It is always good to have additional information which can be added to the system, as it makes the system more powerful. CNNs are capable of extracting representative features for a classification task if enough information is present. Therefore, as we wanted to improve the building outlines without any post-processing steps, it was decided to enrich the system by adding more detailed information from earlier network layers. As a general rule, CNNs gradually abandon lower level features in the pursuit of higher levels, which leads to a more abstract description of the image. This strategy can be countered by passing lower level features up the hierarchy in a separate path (skip connection). In this way, the network itself automatically learns higher detailed building representations. The effectiveness of the suggested FCN4s approach is illustrated in Figure 4.5. First of all, in each resulting image, for every data source, one can notice that more buildings are extracted. Second, the shapes of the footprints, even for the complex building structures, are closer to the ground-truth and better in comparison to the one extracted by FCN8s architecture. Finally, the addition of the *pool2* skip connection, enable the network to recognize even the low-rise buildings.



**(a)** RGB-based model **(b)** nDSM-based model

**(c)** PAN-based model **(d)** Ground-truth

**Figure 4.5:** The relative performance of the FCN4s model for building mask generation on individual data sources (a) RGB, (c) nDSM and (b) PAN images. Figure (d) illustrates the ground-truth building mask.

#### 4.4.1.3 Fused-FCN4s Setup

In this section, we investigate different setups of Fused-FCN4s architecture. Setting the number of convolution layers to 2 for performing a fusion from different network streams and increasing the number of feature maps at the top layers lead to a tendency to improve the result (see Table 4.1). This happens due to the fact that the increase of the parameters number in the network raises its capacity and, thus, makes it possible to perform better generalization. However, at some point the network can reach too much complexity which comes with the risk of overfitting. This effect can be observed with a configuration of 60 feature maps and three convolutional layers. The results of generalization degrade in comparison to a fusion network with 30 feature maps and three convolutional layers. Growing the number of feature maps in the network increases the computation time respectively as depicted in Table 4.1. However, it helps to improve the results significantly. Hence, we choose the model with 30 feature maps and three convolutional layers as it provides the best results in this experiment.

#### 4.4.1.4 Fused-FCN4s Versus FCN8s and FCN4s

The Fused-FCN4s architecture, which combines the spectral information from RGB and PAN images, together with the height information from nDSM, delivers the best performance in discriminating buildings from background, in comparison to FCN8s and

| | Mean acc. | Mean IoU | Overall acc. | IoU | $F_{measure}$ | $n_p$ | $t_f$, ms | $t_b$, ms | $t_{f-b}$, ms |
|---|---|---|---|---|---|---|---|---|---|
| 2fmaps_2conv | 81.4 | 74.5 | 93.9 | 69.7 | 71.6 | 402 773 872 | 85.86 | 347.648 | 433.647 |
| 30fmaps_2conv | 91 | 85 | 96.5 | 74 | 85 | 403 205 352 | 93.1116 | 367.919 | 461.177 |
| 60fmaps_2conv | 91.4 | 85.9 | 96.7 | 75.4 | 86 | 403 672 872 | 102.791 | 376.647 | 479.616 |
| 2fmaps_3conv | 90.7 | 84.6 | 96.3 | 72.83 | 84.3 | 402 773 876 | 86.41 | 350.601 | 437.16 |
| 30fmaps_3conv | **91.5** | **86** | **96.8** | **75.7** | **86.1** | 403 205 772 | 93.5415 | 370.621 | 464.297 |
| 60fmaps_3conv | 90.9 | 85.6 | 96.7 | 74.9 | 85.7 | 403 647 612 | 102.826 | 377.569 | 480.529 |

**Table 4.1:** The results of detailed investigation on Fused-FCN4s model performance with respect to modifications in architecture. We vary the number of feature maps (fmaps) in the top layers together with the number of convolutional layers after merging the streams from three data sources. The $n_p$ indicates a number of parameters in the network, $t_f$ is the average time for one forward pass on a single NVIDIA Titan X (Pascal) GPU, $t_b$ is the average time for one backward pass and $t_{f-b}$ is the average time for one forward-backward pass.

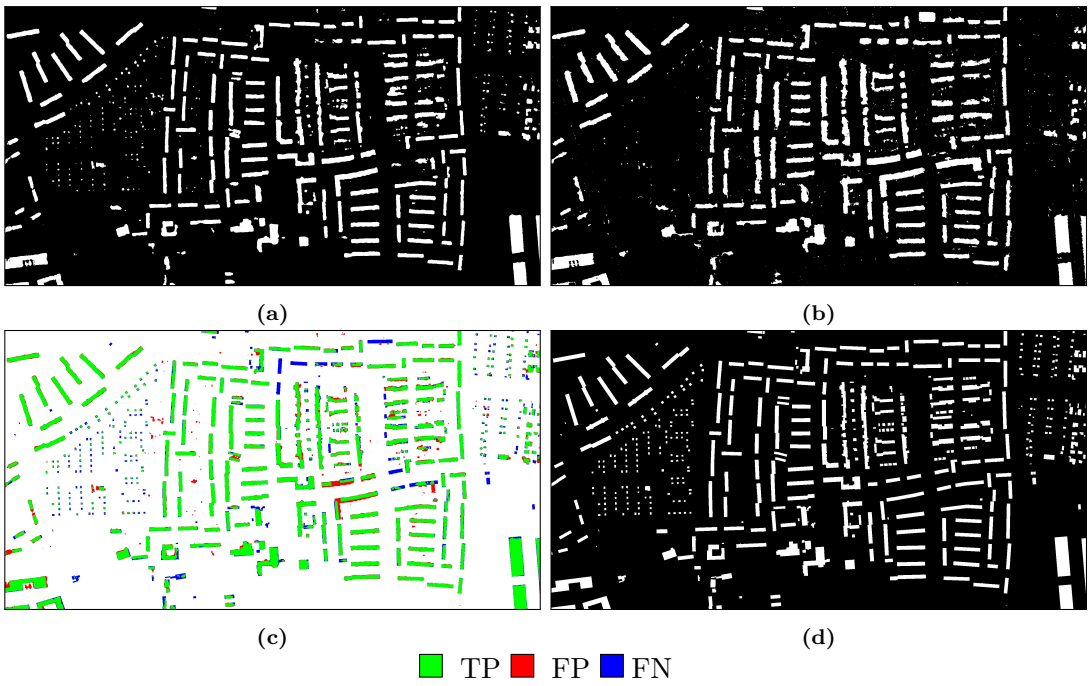



**Figure 4.6:** The comparison of generated building masks over test area obtained (a) directly form Fused-FCN4s and (b) from Krauß *et al.* [7]. Figure (c) depicts the extracted building footprints in respect to reference building footprints of Fused-FCN4s and Figure (d) is a ground-truth building mask.

FCN4s shown in Figures 4.4 and 4.5, respectively. The results obtained by Fused-FCN4s architecture are shown in Figure 4.6a. For visualization and better interpretation the extracted building footprints are also overlapped with the reference building footprints in Figure 4.6c. The significant improvement of the buildings outlines can be easily observed. The footprints are more accurate and their shapes are more complete without missing parts of the various structures. It also can be seen that the network really benefits from all data sources, which allow it to extract more detailed information of building construction compared to the reference image in Figure 4.6d. For example,

| | Mean acc.,% | Mean IoU,% | Overall acc.,% | IoU,% | $F_{measure}$,% |
|---|---|---|---|---|---|
| **FCN8s** | | | | | |
| RGB | 82.8 | 75.5 | 94 | 57.6 | 73.1 |
| nDSM | 74.5 | 69 | 92.8 | 45.7 | 62.7 |
| PAN | 84.6 | 77 | 94.3 | 60.2 | 75.1 |
| **FCN4s** | | | | | |
| RGB | 89.3 | 81 | 95.2 | 67.1 | 80.4 |
| nDSM | 83.3 | 73.3 | 92.9 | 54.3 | 70.4 |
| PAN | 84.4 | 77.5 | 94.6 | 60.9 | 75.7 |
| **Fused-FCN4s** | | | | | |
| RGB & nDSM | 90.9 | 84.7 | 96.1 | 73.5 | 84.7 |
| PAN & nDSM | 87.5 | 82.2 | 95.9 | 68.9 | 81.6 |
| RGB & nDSM & PAN | **91.5** | **86** | **96.8** | **75.7** | **86.1** |
| **Naïve fusion** | | | | | |
| RGB & nDSM & PAN | 87.6 | 81.7 | 95.7 | 68.1 | 81 |
| **DSM-based building detection method** | | | | | |
| MS image & nDSM | 89.1 | 78.2 | 94.6 | 62.4 | 76.8 |

**Table 4.2:** The quantitative evaluation of proposed Fused-FCN4s on three data sources in comparison to different methodologies and setups.

the building in the left bottom corner obviously has some additional structures in the middle, which can be easily identified on the nDSM image (see Figure 4.3b), but they are missing in the ground-truth. The extraction of low-rise buildings, on which the selected scene is rich, is more accurate now, and their pattern of placement is very close to the ground-truth. Some of them are still missing, but that is explainable due to their really small size, difficult to distinguish even for the human eye.

Besides, it is experimentally proven that the proposed network benefits from three remote sensing images used for training in comparison to two-stream networks of RGB and nDSM and PAN and nDSM (see Table 4.2). We can see that the use of the PAN image leads to improvements of 2.2% on *Intersection over Union (IoU)* and from 0.7% to 2% on the rest of the metrics.

#### 4.4.1.5 Fused-FCN4s Versus Naïve Fusion

The experimental results from Table 4.2 demonstrate that naïve fusion by averaging the predicted maps improves the IoU metrics only by 1% in comparison to the results, achieved by FCN4s model trained on RGB. But the proposed Fused-FCN4s boosts the IoU metrics by 8%. Thus, the shapes of generated building footprints are enhanced in comparison to those obtained by single FCN4s. Additionally, a significant improvement of other metrics is also achieved. This proves that the network learns by itself from which multi-source data the better prediction of the pixel can be gained.

#### 4.4.1.6 Fused-FCN4s Versus DSM-based building detection method

As can be seen from Figure 4.6b the DSM-based building detection method proposed by Krauß *et al.* [7] is able to extract a similar building mask as our proposed approach.

However, a close investigation shows that Fused-FCN4s is able to find more buildings than the DSM-based building detection method (see Figure 4.7). Additionally, one can notice that the extraction of low-rise buildings by our approach is significantly better. Besides, the footprints outlines are more accurate and rectilinear, that makes them look qualitatively more realistic and, as a result, similar to the ground-truth.

### 4.4.2 Quantitative Evaluation

For quantitative evaluation of the obtained results, we evaluated the metrics commonly used in semantic segmentation problem. The first group of metrics is described in [59]. They are *mean accuracy*, *mean IoU* and *overall accuracy*

$$\text{Mean accuracy} = \frac{1}{n_{cl}} \sum_i \frac{n_{ii}}{t_i}, \tag{4.7}$$

$$\text{Mean IoU} = \frac{1}{n_{cl}} \sum_i \frac{n_{ii}}{t_i + \sum_j n_{ji} - n_{ii}}, \tag{4.8}$$

$$\text{Overall accuracy} = \frac{\sum_i n_{ii}}{\sum_i t_i}, \tag{4.9}$$

where $n_{ij}$ is the number of pixels belong to class $i$, but predicted as class $j$, $n_{cl}$ is the number of different classes, and $t_i = \sum_j n_{ij}$ is the total number of pixels belong to class $i$.

The second group of selected metrics, suitable for binary classification evaluation, are based on predicted values represented by the total number of true positive (TP), false positive (FP) and false negative (FN). The *Precision* and *Recall*

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{4.10}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4.11}$$

are the number of predicted positives extracted precisely, and fraction of actual positives

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

**Figure 4.7:** The detailed comparison between (a) Fused-FCN4s and (b) DSM-based building detection method proposed by Krauß *et al.* [7]. Figure (c) depicts the ground-truth building mask.

to predicted positives, respectively. Based on these values the *F-measure* is defined as

$$F_{measure} = \frac{(1+\beta^2)TP}{(1+\beta)^2TP + \beta^2FN + FP} , \tag{4.12}$$

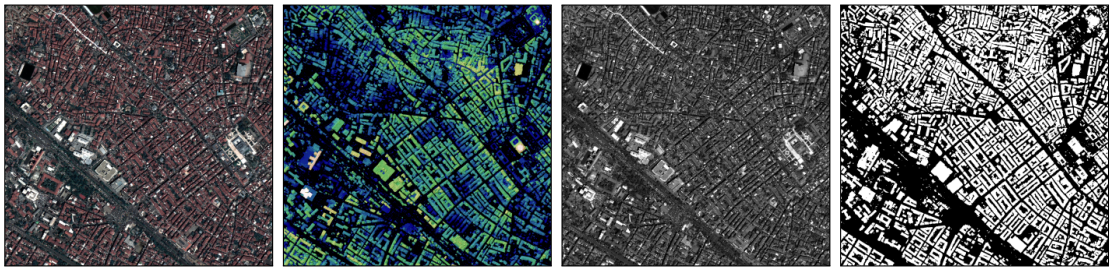
where for our work the parameter $\beta$ was set to 1. Additionally, we use the IoU metric

$$\text{IoU} = \frac{TP}{TP + FN + FP} \tag{4.13}$$

adapted for the task, where the amount of pixels belonging to the objects (buildings) are much smaller compared to those belonging to the background. This metric is represented by the proportion of the number of pixels classified as buildings, both in the predicted image and in the ground-truth, to the total number of pixels classified as buildings in each of them [41].

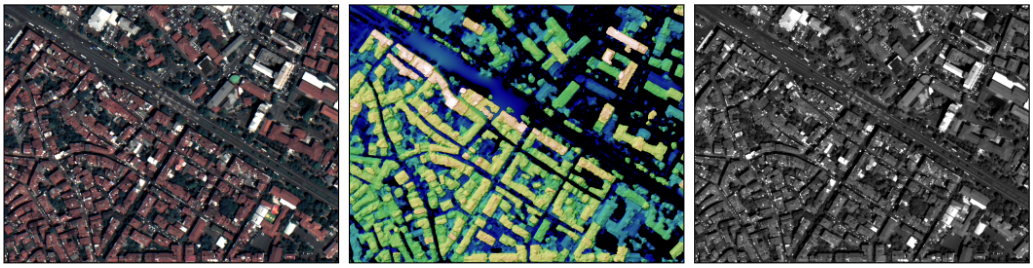
The summarized performances of FCN8s, FCN4s, Fused-FCN4s networks and DSM-based building detection method proposed by Krauß *et al.* [7] using above described metrics are grouped in Table 4.2. From the quantitative statistics we can see that, first, the performance of all networks on spectral images are better than on the image representing the height information. This is reasonable, as the DSM images themselves are obtained from the multi-view stereo PAN pairs and some information can be unavailable, due to occlusions by different objects or clouds within the scene. Second, by further augmenting the architecture with "skip" connection from the *pool2* layer, to generate FCN4s network, we gain improvements of performance on nDSM and RGB images. However, for PAN image the improvement is not very significant. This is due to the fact that the network became more complicated using the additional connection as a result of an enlarged number of parameters, but the extracted information comes only from the three times duplicated image and is not enough to provide the network with much more features. Finally, the proposed Fused-FCN4s network obtains the best performance for all metrics in comparison to other networks and the DSM-based building detection method. The overall accuracy gained 2% points in comparison to FCN8s for RGB and PAN images, and around 4 % points related to FCN8s for the nDSM image. It should be mentioned, that the IoU metric on Fused-FCN4s network increased over 15% and 30% in comparison with FCN8s on spectral and depth images, respectively. That indicates a significant improvement of the building footprint delineation accuracy. Besides, the difference of the IoU metric of 13.3% between the Fused-FCN4s and DSM-based building detection method, in favor of the first, points out that applying our approach there is no need for any post-processing steps for building outline refinement as it already provides very accurate building mask.

Processing a selected test area of $1300 \times 2500$ pixels with Fused-FCN4s network takes 25.89 seconds on a single NVIDIA Titan X Pascal GPU with a 100 pixels stride and around 2 minutes for stitching the overlapped patches for the final full image generation.



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

**Figure 4.8:** Generalization over Istanbul city, Turkey on WorldView-2 data consisted of (a) RGB, (b) nDSM and (c) PAN images. The nDSM is color-shaded for better visualization. Figure (d) illustrates the resulted mask derived from Fused-FCN4s model.



(a) (b) (c)

**Figure 4.9:** The selected area over Istanbul city for statistical evaluation depicted input (a) RGB, (b) nDSM and (c) PAN images.

### 4.4.3 Model Generalization Capability

In order to investigate the model capacity to capture the essential features separating buildings from non-buildings, Istanbul dataset was used (see Figure 4.8). This dataset is very different from the Munich dataset, and it is very challenging in itself due to the dense placement of buildings, and the vastly different construction and architecture style. Without re-training the model on the new dataset, the building footprint map was directly obtained by passing the WorldView-2 data through the FCN4s and Fused-FCN4s networks. From the resulting mask shown in Figure 4.8d, it can be seen that the proposed model managed to predict reasonable building mask even from a new and quite complicated dataset. As it was mentioned in Section 4.3, for quantitative evaluation a small area of around 0.5 $km^2$ was selected (see Figure 4.9). The predicted results and ground-truth of this area are presented in Figure 4.10.

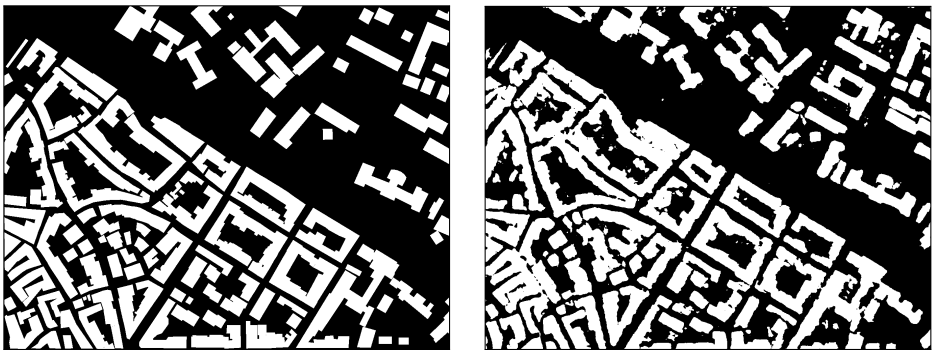
The statistical results of the experiment over the small area can be found in Table 4.3. We can see that the model achieves high performance on this dataset as well. Besides, the advantage of using fused data vs. only one is also demonstrated in Table 4.3.

It can be clearly seen that the model successfully extracts the shapes of building footprints, without missing any of them. The IoU metric confirms this statement by its high value of about $\sim 68.1\%$. Additionally, no influence of other above-ground objects such

| FCN4s | | | | | |
|---|---|---|---|---|---|
| | Mean acc.,% | Mean IoU,% | Overall acc.,% | IoU,% | $F_{measure}$,% |
| RGB | 84.3 | 72.8 | 85.1 | 66.9 | 80 |
| nDSM | 76.3 | 60 | 75.2 | 54 | 70.7 |
| PAN | 79.4 | 66.6 | 81.3 | 58.8 | 74.1 |
| **Fused-FCN4s** | | | | | |
| RGB & nDSM | 85 | 72.8 | 84.9 | 67.7 | 80.8 |
| PAN & nDSM | 84.9 | 72.6 | 84.8 | 66.6 | 79.3 |
| RGB & nDSM & PAN | **85.1** | **73.5** | **85.5** | **68.1** | **81** |

**Table 4.3:** Prediction accuracies of FCN4s and Fused-FCN4s models on all investigated metrics over Istanbul city area selected for statistical evaluation (*cf.* Figure 4.9).

as trees is observed. However, one can notice a small improvement between using one spectral image or two together with an nDSM. Both RGB & nDSM and PAN & nDSM models already gave good results using the advantages from spectral and height information. Inserting additional spectral information only helps to improve minor errors, especially on building outline as the IoU shows high values. But it is still a significant progress as commonly used methodologies for building extraction are not very flexible and can not be easily generalized on different city areas. Moreover, it can be identified that the quantitative results are lower than the ones from Munich dataset. This can be explained by scene complexity: The network did not experience such types of constructions, their close placement to each other and the narrow streets. Besides, the maximum height within Munich nDSM area is 58.37 m and for Istanbul is 24.66 m which also can influence the performance. Another reason is that the manually generated ground-truth is far from ideal, due to the subjective interpretation of human. The probable solution to those small problems can be a fine-tuning of the proposed model on some small areas of different cities, which will contribute to the model performance by introducing a new dataset for model learning, even if it is only a small part of the area.



**(a)** **(b)**

**Figure 4.10:** Generalization results over Istanbul city area selected for statistical evaluation (*cf.* Figure 4.9). Image (a) shows the ground-truth, partially obtained from OSM and partially completed by manually drawing the footprints. Image (b) illustrates the predicted map by Fused-FCN4s model.

## 4.5 Summary

We presented a novel method to segment buildings in complex urban areas using multiple remote sensing data on the basis of fully convolutional networks. The designed end-to-end Fused-FCN4s framework integrates the automatically learned relevant contextual features from spectral and height information from *red, green, and blue (RGB)*, *normalized Digital Surface Models (nDSMs)*, and *pan-chromatic (PAN)* images respectively, within one architecture for pixel-wise classification, and produces a unique binary building mask. Both, spectral images and nDSMs, have their strong and weak sides, but they can complement each other significantly, as, for example, the nDSMs provides elevation information of the objects, but spectral images provide texture information and more accurate boundaries. The trained system was tested on two unseen areas of Munich city, Germany, and Istanbul city, Turkey, and achieved accurate results. Experimental results have shown that even small objects with tiny details in their building footprint can be successfully extracted from satellite images by applying the deep neural network framework. The proposed architecture can be generalized over diverse urban and industrial building shapes, without any difficulties due to their complexity and orientation. Additionally, we show that the designed model does not need any post-processing. Some noise or still present inaccuracies in the resulting building mask can be a result of buildings totally covered by trees, or very complex areas which are difficult to recognize even for the human eye, for accurately extracting the building outlines. Besides, a noisy nDSMs can influence the results to a great extend, as the height information is crucial to identify buildings. We believe that the presented technique has a great potential to provide a robust solution to the problem of building footprint extraction from remote sensing imagery at a large scale.

Chapter 5

# DSM-to-LoD2: Spaceborne Stereo Digital Surface Model Refinement

This chapter describes novel concepts towards improved *Digital Surface Models (DSMs)* generation with realistic building geometries from low-quality half-meter resolution photogrammetric DSMs from satellite data. Mainly involving *conditional Generative Adversarial Networks (cGANs)* with an objective function based on negative log likelihood, improved DSMs are generated with enhanced building forms close to the *Level of Detail (LoD)* 2 according to the *City Geography Markup Language (CityGML)* standard directly from noisy inputs. Focusing on the further improvement of low-quality satellite DSMs, potentials of multi-task learning dedicated to the joint end-to-end training of regression and pixel-wise classification tasks, and the fusion of multiple modalities representing spectral and height information at different stages are demonstrated. This chapter describes combined findings of the following two peer-reviewed journal papers

[4]: **K. Bittner**, P. d'Angelo, M. Körner, and P. Reinartz, "DSM-to-LoD2: Space-borne Stereo Digital Surface Model Refinement," *Remote Sensing*, vol. 10, no. 12, p. 1926, 2018

[5]: **K. Bittner**, M. Körner, F. Fraundorfer, and P. Reinartz, "Multi-Task cGAN for Simultaneous Space-borne DSM Refinement and Roof-Type Classification," *Remote Sensing*, vol. 11, no. 11, p. 1262, 2019

as well as the following double-blind peer-reviewed workshop full paper:

[6]: **K. Bittner**, M. Körner, and P. Reinartz, "Late or Earlier Information Fusion from Depth and Spectral Data? Large-Scale Digital Surface Model Refinement by Hybrid-cGAN," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019

the experiment section of which is extended with additional comparison results from methodology published in [5]. Further examinations towards applicability of refined

DSMs to different remote sensing applications are also investigated in Section 5.4.4, which were not included in any of the aforementioned papers.

## 5.1 Problem Statement

Worldwide urbanization has transformed vast farmlands and forests into urban landscapes, resulting in the appearance of new houses and infrastructures. Due to these rapid changes, accurate and timely updated cadastral 3D building model data are not often available, but which are valuable for urban planning and city management, navigation, virtual environment generation, disaster analysis, tourism, civil engineering, etc. The methodologies for realizing these applications are based mainly on 3D elevation information. Therefore the automatic generation of 3D elevation models with highly accurate building shapes, including the recovery of disturbed boundaries and robust reconstruction of high-quality rooftop geometries, is in demand.

Remote sensing technology provides several ways to measure the 3D urban morphology. Conventional ground surveying, stereo airborne or satellite photogrammetry, *Interferometric Synthetic Aperture Radar (InSAR)*, and *Light Detection and Ranging (LiDAR)* are the main data sources used to obtain high-resolution elevation information [248]. The main advantage of DSMs generated using ground surveying and LiDAR is their high-quality and detailed object representation. However, their production is costly and time consuming, and covers relatively small areas compared to images produced from space-borne remote sensing [249]. *Synthetic Aperture Radar (SAR)* imagery is operational in all seasons under different weather conditions. Nevertheless, it has a side-looking sensor principle that is not very useful for building recognition and reconstruction compared with optical imagery. The DSMs generated from space-borne data using image matching currently show relatively high spatial resolution and wide coverage which is preferable for large-scale remote sensing applications, particularly when using sub-meter multi-view stereo data from, e.g., WorldView or Pleiades satellites. However, due to low-resolution and shortcomings in automatic DSM generation, some unwanted failures in building geometries may occur which influence their later reconstruction and modeling. These originate from applied interpolation techniques, temporal changes, or matching errors. For example, low-textured, homogenous, or shadowed areas caused by a combination of Sun to satellite viewing geometries and surface properties often lead to blunders or not-sharp rooftop contours, as the automatic matching of the homogenous points fails. Specific radiometric effects, such as spilling and saturation on roof planes due to the acquisition geometry, the surface type, or the inclination, may also lead to local blunders [213]. Moreover, densely located buildings in city areas cause uncertainty about building edges. As a result, the applied interpolation techniques lead to low sharpness. Another common problem in urban areas is the occurrence of occlusions due to tall buildings or trees; these also depend on the acquisition viewing angles. Hence, these DSMs need to be refined either manually or automatically to make them more useful for remote sensing applications.

Because manual refinement is costly and time consuming, there has been a considerable amount of research done regarding automatic surface model refinement by conventional as well as deep learning-based methodologies reviewed in Section 3.2. Despite the trials of depth image estimation with deep *Convolutional Neural Networks (CNNs)* and 3D object generation with variational *Generative Adversarial Networks (GANs)* architectures, there is no direct similarity to the problem addressed in this chapter. In the past, there have been no attempts to generate a remote sensing elevation model with an accurate building using CNN-based methodologies. Additionally, among the variety of methodologies that have been developed to refine the 3D urban structure, only a few of them use DSMs generated from stereo satellite imagery, because this type of data features strong noise, inconsistencies, or absence of data due to occlusions between the objects. Urban surface reconstruction based on stereo satellite imagery is still a complex problem.

The ability of CNN-based approaches to reconstruct depth images is strongly correlated with our area of interest. As a result, the potential of cGANs to reconstruct depth images from a bird's-eye view perspective is analyzed in this chapter. In our case, depth images represent the urban 3D structure with elevation information in the form of continuous values. In continuation of previous investigations [250, 251], where the first attempt to generate a LiDAR-like quality DSM out of a given photogrammetric DSM is made using a cGAN with an objective function based on negative log likelihood, the potential of cGANs to generate DSMs with a refined form of buildings close to the LoD2 model according to the CityGML standard is explored in this chapter, without any limitations on their geometry or space scale. We also mention that although the so-called LoD2-DSM does not contain any above-ground objects except buildings, it is still useful for many remote sensing applications, like navigation, 3D city modeling, and cadastral database updates.

Our contributions regarding application of cGAN model for the LoD2-to-DSM approach can be summarized according to the paper [4] as follows:

- We efficiently adapt the cGAN architecture developed by Isola *et al.* [67] from generic images to satellite images and analyze it for different data sources: LoD2-DSM from CityGML and LiDAR-DSM.

- We investigate the potential of using the objective function with least squares instead of negative log likelihood through which we gain more accurate building structures.

- The proposed framework generates images with continuous values representing the elevation models and at the same time, enhances the building geometries.

- Our approach is not limited to the libraries of predefined building models and as a result, can be generalized to large-scale scenes.

- We propose a methodology to convert CityGML vector data into a LoD2-DSM.

- We develop a universal network which is able to generalize over different urban landscapes that have not been previously seen by the model, as generalization is an important aspect for remote sensing applications.

Surface topography is not the only data that contains useful information for DSM improvement. The 2D information in the form of pixel-wise semantic segmentation is also crucial for many remote sensing applications, because it provides additional knowledge about object boundaries or categories to which the object belongs. In most cases, each task, e.g., depth image generation and pixel-wise image classification, is tackled independently although they are closely connected. Jointly solving these multiple tasks can enhance the performance of each independent task, as well as speed up computation time. This observation leads to the advantages of *multi-task (MT)* learning. The approach of simultaneously improving the generalization performance of multiple outputs from a single input was applied to numerous machine learning techniques. As a promising concept for CNNs, multi-task learning has been successfully proven to leverage a variety of problems, like classification and semantic segmentation [109] or classification and object detection [110]. Due to the fact that different tasks may conflict, multi-task learning is regarded as the optimization of a multi-task loss which minimizes a linear combination of contributed single-task loss functions. As a result, in the second approach good-quality LoD2-like DSMs with realistic building geometries are produced together with dense pixel-wise rooftop classification masks. An auxiliary *normal vector loss* term is also added to the objective function enforcing the model to produce more planar and flat roof surfaces, similar to the desired LoD2-DSM derived from CityGML data.

Our contributions regarding the multi-task approach can be summarized based on the paper [5] as follows:

- We efficiently adapt the cGAN architecture developed by Isola *et al.* [67] for multi-task learning.

- The proposed framework generates images with continuous values representing elevation models with enhanced building geometries and at the same time, images with discrete values depicting the label information designating to which class out of three (flat roof, non-flat roof, and background) every single pixel belongs to.

- We investigate the potential of different network architectures for each task and select the combination of models that provides the best results for both pixel-wise classification and depth map generation. We show that joint training of multiple tasks within the end-to-end framework is beneficial. Moreover, the obtained roof classification information can be used later in a post-processing step for the final building modeling task.

- We investigate the potential of using a normal vector loss, which is included as an additional term to the objective function with least squares, thereby gaining more accurate and planar roof structures.

As it is common in the field of remote sensing to fuse data of different modalities to complement missing evidence, in the third approach cGAN-based networks are investigated which merge height and intensity information within end-to-end frameworks for further DSMs enhancement. The influence of the fusion concept at different network stages is mainly explored which merge two separate networks—fed with *pan-chromatic (PAN)* images and DSMs—either at a later stage right before producing the final output (WNet-cGAN architecture) or at an earlier stage (Hybrid-cGAN architecture) concatenating two encoders at the top layer and integrating information from two different modalities through a common decoder. In both cases, the networks automatically couple the advantages of PAN imagery containing sharp information about building boundaries and ridge lines and photogrammetric DSM with information about building silhouette and height. An auxiliary normal vector loss term is also added to the final objective function to influence the planarity and flatness of roof surfaces.

Our contributions regarding the multi-modal approach can be summarized according to the paper [6] as follows:

- We implement two architectures, namely WNet-cGAN and Hybrid-cGAN, capable of blending the intensity and height information from PAN images and DSMs, respectively, within a cGAN framework for generating accurate high-quality DSMs with refined building geometries.
- We show that the involvement of intensity information in the form of PAN images immediately improves the accuracy and appearance of the generated LoD2-like DSMs compared to the LoD2-like DSMs generated from the single photogrammetric DSM.
- We investigate the potential of late and earlier fusion stages within the network architecture and demonstrate that the earlier fusion not only reduces the number of network parameters, it also integrates the information from different modalities better.

To our knowledge, this is the first study to carry out DSM refinement using deep learning techniques.

The remainder of the chapter is arranged as follows. Section 5.2 is based on paper [4] and investigates the first methodology for accurate DSMs reconstruction using cGANs. The proposed deep network architecture, the background of GANs, and the objective functions are described in more detail in Section 5.2.1. Furthermore, the description of necessary ground-truth data generation, which is required for the training process is presented in this section. In Section 5.2.2, we introduce the dataset and present implementation details and training strategies of the first methodology. The experimental results for two different datasets applying the proposed deep network architecture together with qualitative and quantitative evaluations are shown and discussed in Section 5.2.3. Section 5.3 is based on paper [5] and explores the possible improvement of the methodology presented in Section 5.2 by applying the MT learning strategy for simultaneous good-quality DSMs generation and building roof type classification maps

production. Section 5.3.1 proposes different network architectures, which we aim to study together with the final objective function. In Section 5.3.2, the used dataset and applied training setups are described. The results of the proposed MT learning strategy are interpreted in Section 5.3.3. Section 5.4 is based on paper [6] and examines another strategy dedicated to DSM refinement. It mainly investigates intensity and height data fusion at different stages in the neural network to further improve building shapes and overall quality of DSMs. Section 5.4.1 is based on paper [6] and describes the proposed network architectures, objective functions, and used dataset. The detailed comparison of results produced by the data fusion network with results of two earlier methodologies (including multi-task methodology results additionally processed for comparison with the fusion methodology) are given in Section 5.4.3. Moreover, Section 5.4.1 includes additional investigation towards the applicability of resulting DSMs after the refinement procedure to different remote sensing applications which were not included in the original paper [6]. Section 5.5 presents a comparison discussion of the results obtained by the three proposed methodologies.

## 5.2 Conditional GAN for LoD2-like DSM Generation

Recently, fast emerging CNNs have been applied to depth estimation tasks. But within the remote sensing community, the height image generation problem using CNNs has rarely been addressed. Therefore, in this section the applicability of CNNs for the enhancement of DSMs is investigated.

### 5.2.1 Methodology

#### 5.2.1.1 Network Architecture

The proposed architectures are adapted from those presented by Isola *et al.* [67]. The *generator* $G$ is an *UNet* [179] accepting a single-channel depth image with continuous values as input. The *hyperbolic tangent* activation function

$$\sigma_{\tanh}(z) = \tanh(z) \tag{5.1}$$

is applied on the top layer of the $G$ network. The UNet is an encoder-decoder type of network that progressively down-samples the input through a series of layers until a bottleneck layer and codes back the process from this point. In order to recover important details that are lost while down-sampling in the encoder, skip connections are added to the network to combine the encoder layer $i$ with the up-sampled decoder layer $n - i$ at every stage. In our work, the encoder part of UNet is constructed with 8 layers and 7 skip connections. The photogrammetric DSM images are accepted as an input image of this network. The detailed UNet network architecture is illustrated in Figure 5.1.

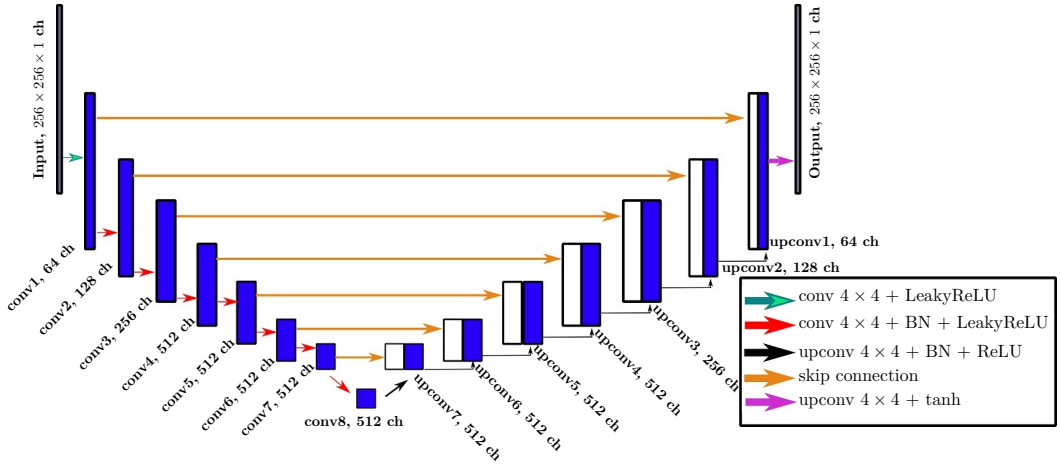



**Figure 5.1:** Schematic overview of the proposed *UNet* architecture. Each convolution operation has a kernel of size $4 \times 4$ with stride 2. For up-sampling, the transposed convolution operations with kernels of size $4 \times 4$ and stride 2 are used. The Leaky ReLU activation function in the encoder part of the network has a negative slope of 0.2.

The *discriminator* $D$ is a binary classification network consisting in our case of 5 convolutional layers. The input to the discriminator $D$ is a concatenation of a photogrammetric DSM with either a UNet-generated fake DSM or a ground-truth DSM. The $D$ has a *sigmoid* activation function

$$\sigma_{\text{sigm}}(z) = \frac{1}{1 + \mathrm{e}^{-z}} \tag{5.2}$$

on the top layer, because it is meant to output the probability that the input image belongs either to class 1 ("real") or class 0 ("generated").

A schematic diagram of the proposed network architecture is illustrated in Figure 5.2.

#### 5.2.1.2 Objective

As proposed by Goodfellow *et al.* [65] in 2014, GAN techniques are characterized by training a pair of networks, namely a generator $G$ and a discriminator $D$, which are trained in an adversarial manner to compete against each other. The aim of $G(\boldsymbol{z}) = \boldsymbol{y}$ is to implement a differentiable function to map a latent vector $\boldsymbol{z} \sim p_z(\cdot)$ drawn from any distribution $p_z(\cdot)$, e.g., a uniform distribution $p_z(\cdot) = \text{Unif}(a, b)$, to an element $\boldsymbol{y} \sim p_{\text{real}}(\cdot)$ that is approximately distributed according to $p_{\text{real}}$, i.e., into the form of the data we are interested in imitating. In contrast, $D(\boldsymbol{y}) \in [0, 1]$ attempts to differentiate between the generated data $\boldsymbol{y}$ and the genuine sample $\boldsymbol{y}^*$. The objective function for GANs can be expressed through a two-player minimax game

$$\min_G \max_D \mathcal{L}_{\text{GAN}}(G, D) = \mathbb{E}_{\boldsymbol{y} \sim p_{\text{real}}(\boldsymbol{y})}[\log D(\boldsymbol{y})] + \mathbb{E}_{\boldsymbol{z} \sim p_z(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))], \tag{5.3}$$

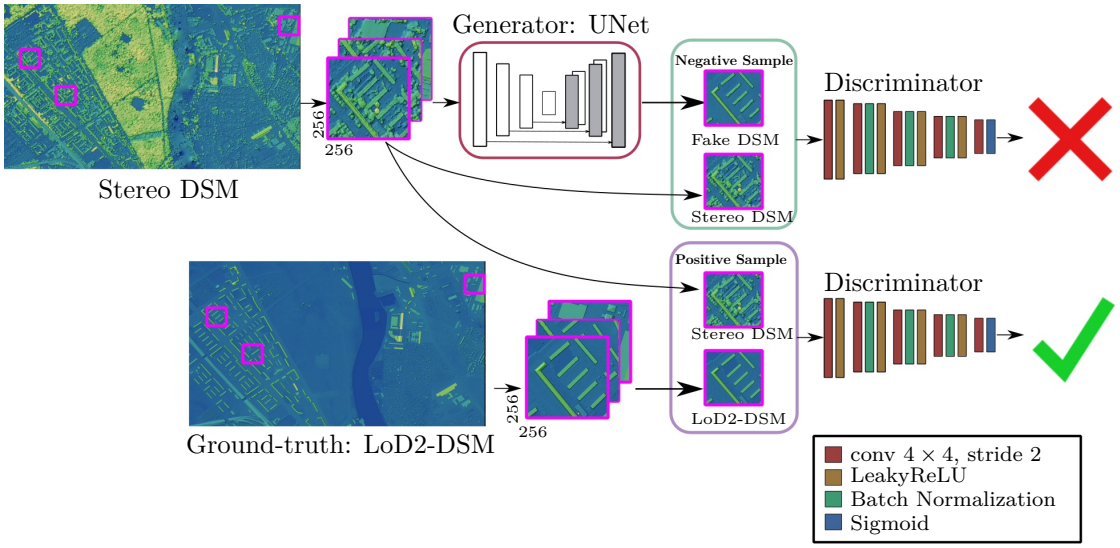



**Figure 5.2:** Schematic overview of the proposed method for the 3D building shape improvement in photogrammetric DSMs by cGAN. The DSM images are color-shaded for better visualization.

where $\mathbb{E}[\cdot]$ denotes the expectation value. The discriminator $D$ is realized as a binary classification network that outputs the probability that an input image belongs either to class 0 ("generated") or to class 1 ("real"). During training, $G$ aims to create samples that look more and more real, while $D$ intends to always correctly classify where a sample comes from.

In this chapter, we address the generation of better-quality DSMs featuring refined building shapes at LoD2 according to the definitions of CityGML [252]. In other words, the aim is to generate synthetic LoD2-like height images with a similar appearance to the given DSMs from stereo satellite imagery, but with an improved building appearance. The conditioning of the model on external information was first introduced by Mirza *et al.* [66]. The cGANs restrict both the generator in its output and the discriminator in its expected input. As a result, cGANs allow the generation of synthetic images similar to some given input image $\boldsymbol{x}$. In contrast to Equation (5.3), the cGAN objective function

$$\min_G \max_D \mathcal{L}_{\text{cGAN}}(G, D) = \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}\sim p_{\text{real}}(\boldsymbol{y})}[\log D(\boldsymbol{y}|\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x},\boldsymbol{z}\sim p_z(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z}|\boldsymbol{x})|\boldsymbol{x}))] \tag{5.4}$$

now involves some conditional data $\boldsymbol{x}$.

Since the appearance of GAN, many extensions to its architecture have been proposed. The architecture of Isola *et al.* [67] gained the most popularity in cases involving the image-to-image translation problem and is currently used as a basic model for image generation tasks. Our method also builds upon this adversarial system to generate images with continuous values representing the elevation information.

It is common to blend the GANs objective with traditional losses, such as $L_1$ or $L_2$ distances, because this helps the generator to make the created image as close as possible

to the given ground-truth in an $L_1$ or $L_2$ sense. As we are interested in images where the buildings have steep walls and sharp ridge lines, we use the $L_1$ distance

$$\mathcal{L}_{L_1}(G) = \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}\sim p_{\text{real}}(\boldsymbol{y}),\boldsymbol{z}\sim p_z(\boldsymbol{z})}[\|\mathbf{y} - G(\boldsymbol{z}|\boldsymbol{x})\|_1], \tag{5.5}$$

because it encourages less blurring. Adding this term leads to our final objective:

$$G^\star = \arg\min_G \max_D \mathcal{L}_{\text{cGAN}}(G, D) + \lambda \cdot \mathcal{L}_{L_1}(G), \tag{5.6}$$

where $0 \leq \lambda \in \mathbb{R}$ is a weighting hyper-parameter.

Moreover, to overcome the common problem of unstable training when the objective function of GANs is based on the negative log-likelihood, a technique that was recently proposed by Mao *et al.* [253] is applied which replaces the negative log-likelihood in Equation (5.4) by a least square loss $L_2$, yielding the *conditional Least Square Generative Adversarial Network (cLSGAN)* objective

$$\mathcal{L}_{\text{cLSGAN}}(G, D) = \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}\sim p_{\text{real}}(\boldsymbol{y})}[(D(\boldsymbol{y}|\boldsymbol{x}) - 1)^2] + \mathbb{E}_{\boldsymbol{x},\boldsymbol{z}\sim p_z(\boldsymbol{z})}[D(G(\boldsymbol{z}|\boldsymbol{x})|\boldsymbol{x})^2]. \tag{5.7}$$

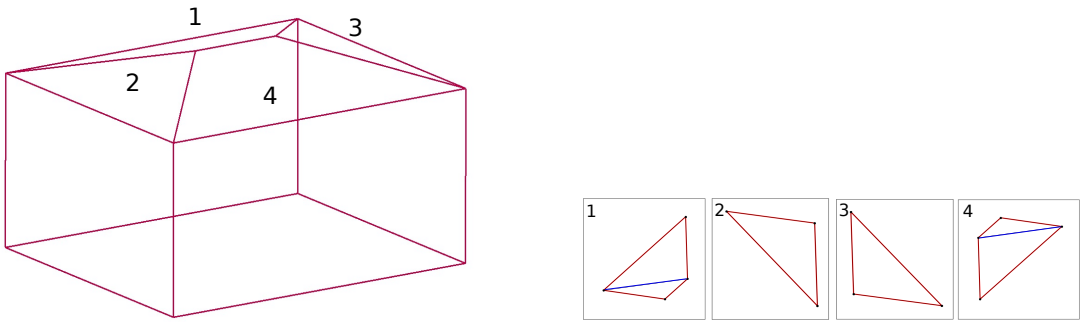
This makes it possible to stabilize the training process and to improve the quality of the generated image. Therefore, the influence of this alternative training procedure using cLSGAN

$$G^\star = \arg\min_G \max_D \mathcal{L}_{\text{cLSGAN}}(G, D) + \lambda\mathcal{L}_{L1}(G), \tag{5.8}$$

is also investigated for the proposed generation of better-quality DSMs.

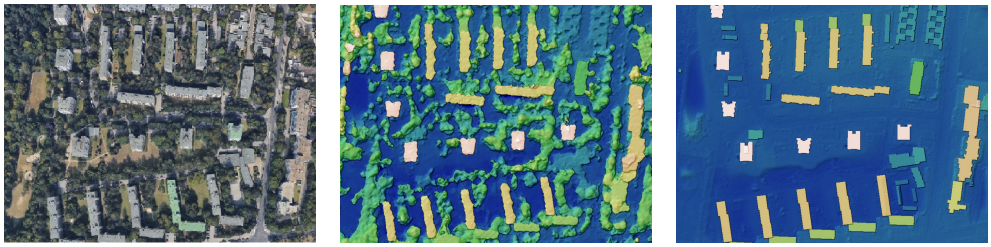
#### 5.2.1.3 LoD2-DSM Ground-Truth Data Generation

CityGML encodes a standard model and mechanism for describing different types of 3D city objects with respect to their geometry, topology, semantics, and appearance. It also provides specific relationships between different objects, e.g., a building is decomposed into roof, wall, and ground surfaces, as seen in Figure 5.3a. For the creation of LoD2-DSM, which is assumed to be a ground-truth data for the experiments, the roof polygons of each building from the database consisting of points with location and height information are selected. Each polygon is triangulated afterwards (*cf.* Figure 5.3b) using the algorithm introduced by Shewchuk [254] based on *Delaunay triangulation* [255]. The software is publicly available. It should be noted that the triangular surfaces are left as they are. In order to generate a raster height image, a DLR software is used to calculate a unique height value of pixels lying inside each triangle using Barycentric interpolation. The pixels outside buildings are filled with a *Digital Terrain Model (DTM)*, i.e., a mathematical representation of the ground surface without above-ground objects. As a result, the synthetically generated so-called LoD2-DSM does not contain any vegetation, only building information. This can be clearly seen in Figure 5.4.




(a) (b)

**Figure 5.3:** An example of CityGML building model representation and triangulation of its roof surfaces. Figure (a) illustrates CityGML building model representation; Figure (b) depicts roof surface triangulation.



**(a)** Google Earth View **(b)** Photogrammetric DSM **(c)** LoD2-DSM

**Figure 5.4:** Illustration of differences in vegetation representation between a photogrammetric DSM from the WorldView-1 satellite and an synthetically generated LoD2-DSM.

## 5.2.2 Study Area and Experiments

### 5.2.2.1 Data

Two different types of ground-truth datasets were used for network training and evaluation of the results.

The first dataset consisted of a space-borne photogrammetric DSM and a LoD2-DSM. The LoD2-DSM was generated with a resolution of 0.5 m from a CityGML data model that is freely available on the download portal Berlin 3D (`http://www.businesslocationcenter.de/downloadportal`). The process of this type of DSM generation is given in Section 5.2.1.3.

The second dataset consisted of a space-borne photogrammetric DSM and a LiDAR-DSM. This experiment was run to demonstrate the improvements in the results from our previous work [250] and additionally, to produce a trained network to perform a model generalization test on (see Section 5.2.3.2), as only LiDAR-DSM data was available for another region. *Semi-Global Matching (SGM)* [17] was used to generate a photogrammetric DSM with a resolution of 0.5 m from six pan-chromatic Worldview-1 images acquired on two different days. The LiDAR-DSM considered as a ground-truth was provided by

the *Senate Department for Urban Development and Housing, Berlin*. It was generated from airborne laser scanning data with a resolution of 1 m and up-sampled to the resolution of 0.5 m to establish consistency with the available photogrammetric DSM. The last pulse laser scanning data was used which contained much less vegetation compared to the photogrammetric DSM. Both datasets used for this experiments show the city of Berlin, Germany within a total area of 410 $\mathrm{km}^2$.

The LiDAR-DSM was used as vertical and horizontal reference during the automatic image orientation as part of the LoD2-DSM generation process. However, a systematic deformation of up to 3.5 m between the LoD2-DSM and the photogrammetric DSM was noticed. Thus, the LoD2-DSM was co-registered to the photogrammetric DSM using an affine transformation based on 19 manually selected points, leading to a fit with a standard deviation of less than 1 m.

To investigate the capacity of the prediction models over a different urban landscape, a third photogrammetric DSM and LiDAR-DSM dataset showing a section of Munich, Germany, was considered. For this test, the photogrammetric DSM from WorldView-2 satellite imagery with a resolution of 0.5 m derived by the same methodology [17] was used. The last pulse airborne laser scanning data with a resolution of 1 m was provided by *Bavarian Agency for Digitisation, High-Speed Internet and Surveying*. The data was rasterized with a resolution of 0.5 m. The Munich dataset covering 9 $\mathrm{km}^2$ was used only in the inference phase.

#### 5.2.2.2 Implementation and Training Details

The DSM-to-LoD2 network was based on the cGAN architecture developed by Isola *et al.* [67] on the *PyTorch* Python package with a slight extension. To organize the training data, the satellite images were tiled into patches of size $256 \times 256$ px which fit into the available *Graphics Processing Unit (GPU)* memory and and were large enough to capture a constellation of building structures and their surroundings. This led to the production of sufficient context information required by the network about building shapes, positions, and orientations. The prepared training data for the learning process consisted of 21 480 pairs of patches covering an area of 353 $\mathrm{km}^2$. To tune the hyper-parameters, validation data covering 6 $\mathrm{km}^2$ was used. The DSM-to-LoD2 network was trained with mini-batch *Stochastic Gradient Descent (SGD)* using the ADAM optimizer [46] with an initial learning rate of $\alpha = 0.0002$ and momentum parameters of $\beta_1 = 0.5$ and $\beta_2 = 0.999$ for both setups, cGAN and cLSGAN. The weighting hyper-parameter $\lambda = 1000$ was chosen after performing the experimental training and examining the resulting generated images and their profiles. During the training phase, two networks $G$ and $D$ were trained at the same time by alternating one gradient descent step of $D$ and one gradient descent step of $G$. To achieve a better optimization behavior when training cGANs, it is common to change $G$ to maximize the $\log D(\boldsymbol{x}, G(\boldsymbol{z}|\boldsymbol{x}))$ instead of minimizing $\log(1 - D(\boldsymbol{x}, G(\boldsymbol{z}|\boldsymbol{x})))$. The total number of epochs was set to 200 with a batch size of 5 on a single NVIDIA TITAN X (Pascal) GPU with 12 GB of memory. During training, a random cropping of the tiles up to one tile size was used instead of the up-scale and random crop data augmentation from the original code. The idea behind this was that the network may observe only some parts of a building in

one patch for one cropping and the whole building in the next time period. Although different configurations were observed in different moments, the same building featuring the same properties was used. This made the network more general, robust, and flexible for a variety of building types.

#### 5.2.2.3 Inference Process

During the inference process, only the trained generator $G$ of the DSM-to-LoD2 network was involved. It generated LoD2-like height images covering 50 km$^2$ after stitching the overlapping patches for the final full image generation. The overlap for the test data was fixed at 128 px in both the horizontal and vertical directions. The test dataset consisted of photogrammetric DSM patches that were never shown to the networks during the training phase.

#### 5.2.2.4 Evaluation Metrics

The quantitative evaluation of generative models is a challenging task, especially if the generated images do not contain continuous values rather than discrete values. Common metrics used to measure accuracy for continuous variables are the *Mean Absolute Error (MAE)*

$$\varepsilon_{\mathrm{MAE}}(\boldsymbol{h}, \hat{\boldsymbol{h}}) = \frac{1}{n} \sum_{j=1}^{n} |\hat{h}_j - h_j| \tag{5.9}$$

and the *Root Mean Square Error (RMSE)*

$$\varepsilon_{\mathrm{RMSE}}(\boldsymbol{h}, \hat{\boldsymbol{h}}) = \sqrt{\frac{1}{n} \sum_{j=1}^{n} (\hat{h}_j - h_j)^2}, \tag{5.10}$$

where $\hat{\boldsymbol{h}} = (\hat{h}_j)_j, 1 \leq j \leq n$, denotes the predicted heights and $\boldsymbol{h} = (h_j)_j$ the actual observed ones. The specifications of these accuracy metrics are usually based on the assumptions that the errors follow a Gaussian distribution and that no outliers exist [256]. However, DSMs derived by digital photogrammetry seldom features a normal error distribution due to the presence of outliers and filtering or interpolation errors. Therefore, Höhle *et al.* [256] proposed the use of a robust scale estimator, such as the *Normalized Median Absolute Deviation (NMAD)*

$$\varepsilon_{\mathrm{NMAD}}(\boldsymbol{h}, \hat{\boldsymbol{h}}) = 1.4826 \cdot \underset{j}{\mathrm{median}}(|\Delta h_j - m_{\Delta \boldsymbol{h}}|) \tag{5.11}$$

which is suitable for non-normal error distributions. It is proportional to the median of the absolute difference between height errors, denoted as $\Delta h_j$, and the median error $m_{\Delta \boldsymbol{h}}$. The constant 1.4826 was included so that NMAD is comparable to the standard deviation when the data are distributed normally. This estimator can be considered more robust to outliers in the dataset.

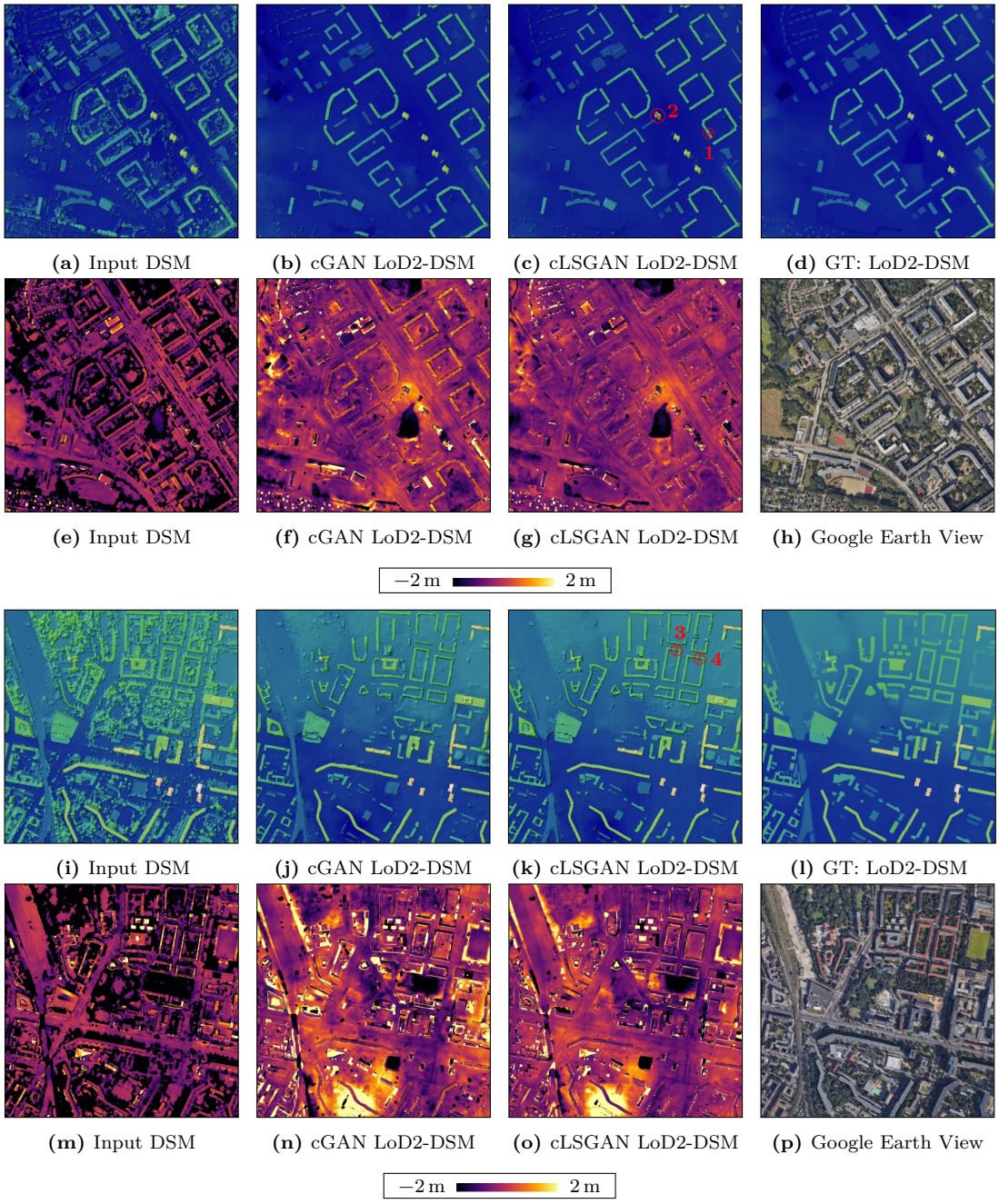
It should be mentioned that because we were interested in quantifying the improvements of the building shapes on DSMs, the above mentioned metrics were measured only in the area where buildings were situated. This was achieved by extracting useful information from the binary building mask generated from the same CityGML data model. We also extend the building footprints by a three-pixel dilation on the boundaries to make sure that the 3D information of building walls was included because we were interested in its improvement.

### 5.2.3 Results

#### 5.2.3.1 cGAN Versus cLSGAN

The examples of DSMs generated by the DSM-to-LoD2 network for both LoD2-DSM and LiDAR-DSM datasets are depicted in Figure 5.5 and Figure 5.6, respectively. By investigating the obtained DSMs, we can see that both cGAN and cLSGAN networks are able to generate the elevation models close to the given ground-truths. In the case of LoD2-like DSM, the model manages to learn that there is no vegetation in the synthetically created LoD2-DSM. This also can be observed by referencing the computed height variation maps in Figures 5.5f, 5.5g, 5.5n and 5.5o, of the generated DSMs in comparison to the LoD2-DSM from CityGML data. The high lightness denotes areas where the photogrammetric or generated DSMs are higher than the ground-truth DSM. As a result, Figures 5.5e and 5.5m demonstrate that the DSM from stereo satellite imagery, on the other hand, contains many trees. Other objects, e.g., cranes or electrical power poles, also vanish because they do not exist in the synthetic LoD2-DSM generated from CityGML data. Regarding the LiDAR-DSM dataset, there is only a small amount of vegetation on the ground-truth DSM because the LiDAR point cloud we use is from the last pulse. In addition, it should be highlighted that the network is trained to only manipulate existing buildings and does not generate new buildings or its parts if there is no building in the input data. Good examples can be seen in highlighted areas in Figures 5.6m, 5.6n and 5.6o or in the zoomed part of the first LiDAR-DSM area illustrated in Figure 5.7. This building exists on the given LiDAR-DSM. However, as there is no sign of the building within the photogrammetric DSM, the two models cGAN and cLSGAN do not reconstruct it.

The height difference maps, presented for LoD2- and LiDAR-like DSMs in the second and the fourth rows of Figures 5.5 and 5.6, respectively, demonstrate that there are less or no residuals within building areas on DSMs from cLSGAN models. In addition, it can be seen that the building structures are fully reconstructed without missing parts. This can be clearly observed in the next example. By zooming into two selected buildings (see Figure 5.8) from the LoD2-like first and second DSM areas, it can be noticed that the buildings generated by the cLSGAN model are more detailed and complete than the one from the cGAN model. The right side of the first building highlighted with 1 in Figure 5.8a is almost missing and the inside construction highlighted with 2 is not reconstructed at all. The cLSGAN produced better results here because it managed to generate those parts (see Figure 5.8b). Regarding the second building, we can see that the upper part of the construction generated by cGAN is not detailed compared to the



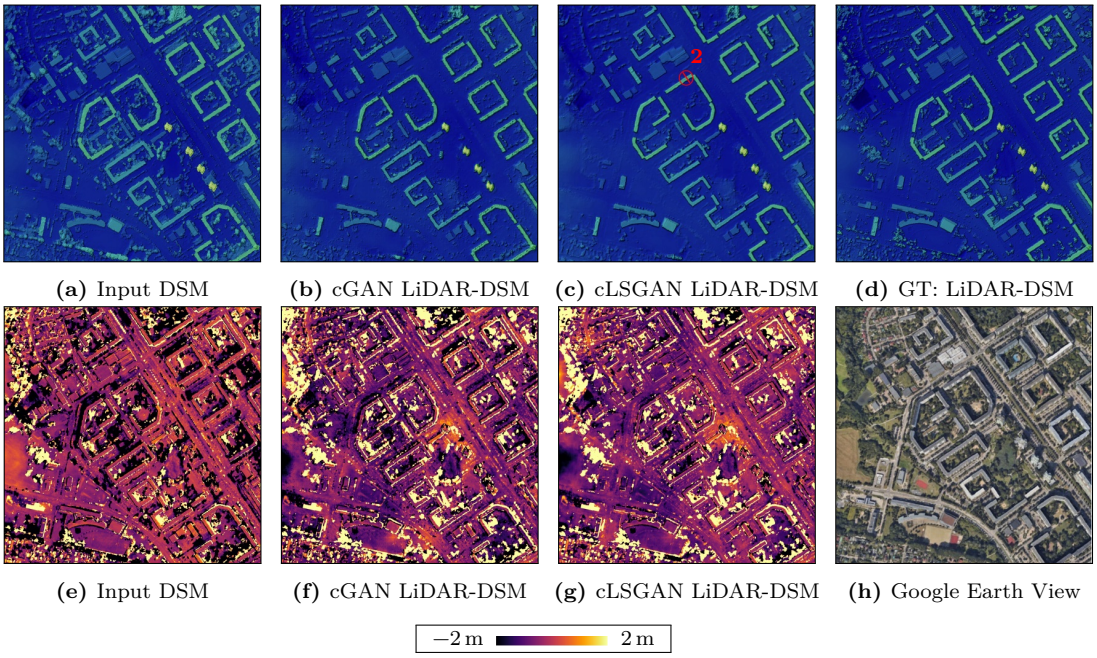

**(a)** Input DSM **(b)** cGAN LoD2-DSM **(c)** cLSGAN LoD2-DSM **(d)** GT: LoD2-DSM

**(e)** Input DSM **(f)** cGAN LoD2-DSM **(g)** cLSGAN LoD2-DSM **(h)** Google Earth View

**(i)** Input DSM **(j)** cGAN LoD2-DSM **(k)** cLSGAN LoD2-DSM **(l)** GT: LoD2-DSM

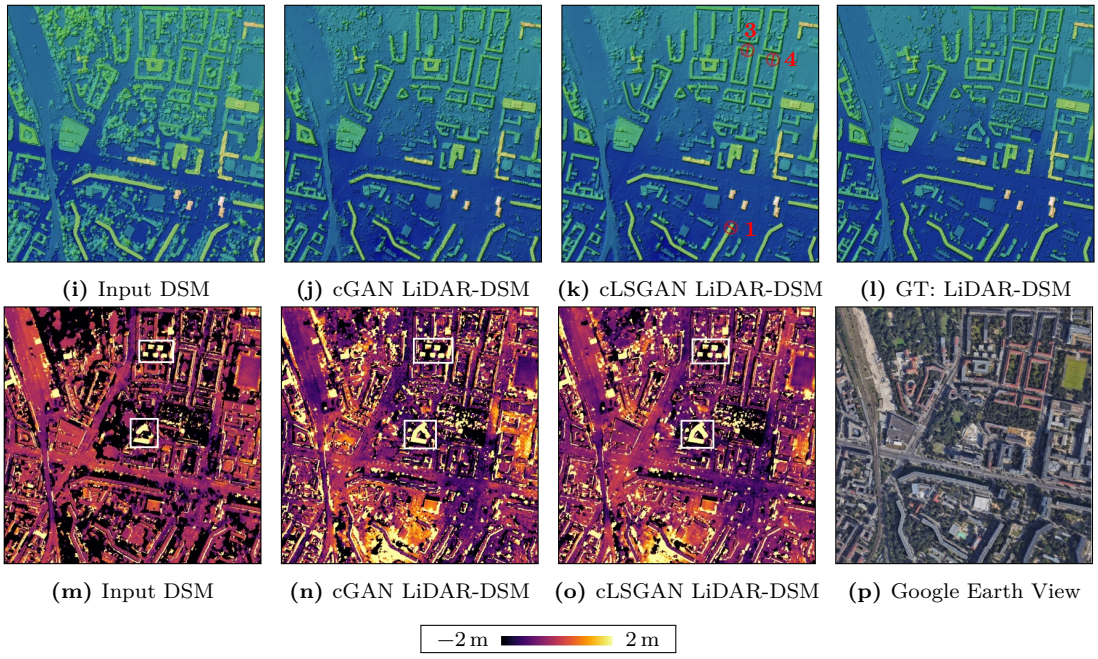
**(m)** Input DSM **(n)** cGAN LoD2-DSM **(o)** cLSGAN LoD2-DSM **(p)** Google Earth View

**Figure 5.5:** Visual analysis of DSMs, generated by cGAN and cLSGAN architectures, over selected urban areas. The DSM images are color-shaded for better visualization. Difference maps in meters of stereo and generated DSMs with respect to ground-truth LoD2-DSM of selected regions are in the second and fourth lines, respectively.




**(a)** Input DSM **(b)** cGAN LiDAR-DSM **(c)** cLSGAN LiDAR-DSM **(d)** GT: LiDAR-DSM

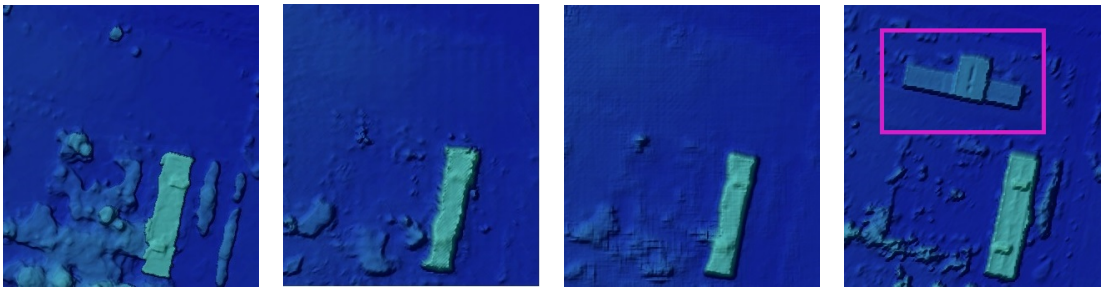
**(e)** Input DSM **(f)** cGAN LiDAR-DSM **(g)** cLSGAN LiDAR-DSM **(h)** Google Earth View




**(i)** Input DSM **(j)** cGAN LiDAR-DSM **(k)** cLSGAN LiDAR-DSM **(l)** GT: LiDAR-DSM

**(m)** Input DSM **(n)** cGAN LiDAR-DSM **(o)** cLSGAN LiDAR-DSM **(p)** Google Earth View

**Figure 5.6:** Visual analysis of DSMs, generated by cGAN and cLSGAN architectures, over selected urban areas. The DSM images are color-shaded for better visualization. Difference maps in meters of stereo and generated DSMs with respect to ground-truth LiDAR-DSM of selected regions are in the second and fourth lines, respectively.



**(a)** Input DSM **(b)** cGAN DSM **(c)** cLSGAN DSM **(d)** GT: LiDAR-DSM

**Figure 5.7:** Demonstration of generalization over existed buildings on input DSM using both cGAN and cLSGAN methodologies trained on LiDAR ground-truth data. (a) illustrates the input photogrammetric DSM, (b) is a generated DSM using cGAN, (c) is a generated DSM using cLSGAN and (d) is a LiDAR ground-truth.

one from cLSGAN. The same problem occurs with the bridge that connects two parts of the building. A specific pattern of holes on some buildings generated by cGAN models is also discovered on both datasets LoD2-DSM and LiDAR-DSM. An example of these distortions can be observed in detailed view in Figure 5.8a.

By investigating the profiles of selected buildings highlighted by red lines in Figures 5.5c and 5.5k and Figures 5.6c and 5.6k for LoD2-DSM and LiDAR-DSM datasets, respectively, we can confirm that in general, both cGAN and cLSGAN models can successfully learn 3D building representations that are close to the ground-truths (green profiles). However, as mentioned before, some artifacts exist on the height images generated by the cGAN model. This is not dependent on the data type as both obtained results from LoD2-DSM and LiDAR-DSM datasets have this problem. Examples are demonstrated in Figure 5.9b for LoD2-DSM and in Figures 5.10a and 5.10b for LiDAR-DSM. Fortunately, the application of the cLSGAN model helps to smooth the artifacts and brings the shape of building even closer to the ground-truth shapes. This achievement is clearly seen in Figure 5.9f and Figures 5.10e and 5.10f. Of course, not every building in the generated results exhibited holes. The poor examples demonstrated are chosen for the visual notion. Moreover, not only the quality of flat roofs is improved. The gable, hip, or any other roof type consisting of inclined planes can be improved by applying our methodology. As can be seen from the profiles (Figures 5.9g and 5.9h and Figures 5.10g and 5.10h), the cLSGAN model provides much better results. The planes of roof surfaces are much smoother and more symmetrical regarding the central ridge line. The ridge lines are primarily much sharper in comparison to ridge lines from the photogrammetric DSM and are at the central position which gives a more realistic view and is geometrically more correct. Additionally, all profiles show a very close resemblance to the ground-truth shapes, especially regarding the width and borders of the buildings.

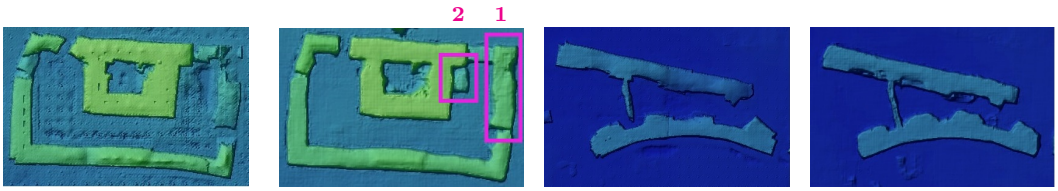
To quantify the generated DSMs, the proposed metrics are evaluated on all setups and their performances are reported in Table 5.1. In general, the resulting DSMs from cGAN model reveal the same or slightly worse results compared even to the normal

**Table 5.1:** Prediction accuracies of cGAN and cLSGAN models on all investigated metrics for LoD2-DSM and LiDAR-DSM datasets over the Berlin area.

| Method | LoD2-DSM Dataset | | |
|---|---|---|---|
| | MAE (m) | RMSE (m) | NMAD (m) |
| Photogrammetric DSM | 3.21 | 6.69 | 1.51 |
| cGAN | 3.05 | 6.66 | 1.30 |
| cLSGAN | 1.63 | 5.72 | 1.22 |
| | LiDAR-DSM Dataset | | |
| | MAE (m) | RMSE (m) | NMAD (m) |
| Photogrammetric DSM | 2.55 | 4.90 | 1.35 |
| cGAN | 2.80 | 5.15 | 1.75 |
| cLSGAN | 2.22 | 4.32 | 1.29 |

photogrammetric DSM as some parts of the buildings are still badly reconstructed, completely missed, or feature holes as shown in the example in Figure 5.8. The relatively high values of RMSE for the LoD2-DSM setup compared to the LiDAR-DSM could be due to the much bigger time difference in data acquisition which leads to cases like those depicted in Figure 5.7a. This significantly influences the computed metrics.

The results obtained by the cLSGAN model on both datasets quantitatively outperformed the photogrammetric DSMs and the DSMs generated by the cGAN model because they are much smoother, and able to reconstruct even small parts of buildings, and do not contain any artifacts. Correspondingly, values of evaluated metrics are lower compared to cGAN-based generated DSMs for both datasets that confirm our statements about the network independence from data type as well as the power of least squares loss function to generate the outputs much closer to the ground-truth. The quantitative evaluation supports the visual examination. Processing one patch of size $256 \times 256$ pixels with the proposed network takes 0.2 seconds on a single NVIDIA Titan X Pascal GPU.




**(a)** cGAN DSM **(b)** cLSGAN DSM **(c)** cGAN DSM **(d)** cLSGAN DSM

**Figure 5.8:** Comparison of generalization over DSM between cGAN and cLSGAN methodologies for two selected buildings. (a),(c) are the generated buildings by cGAN and (b),(d) are the generated buildings by cLSGAN.

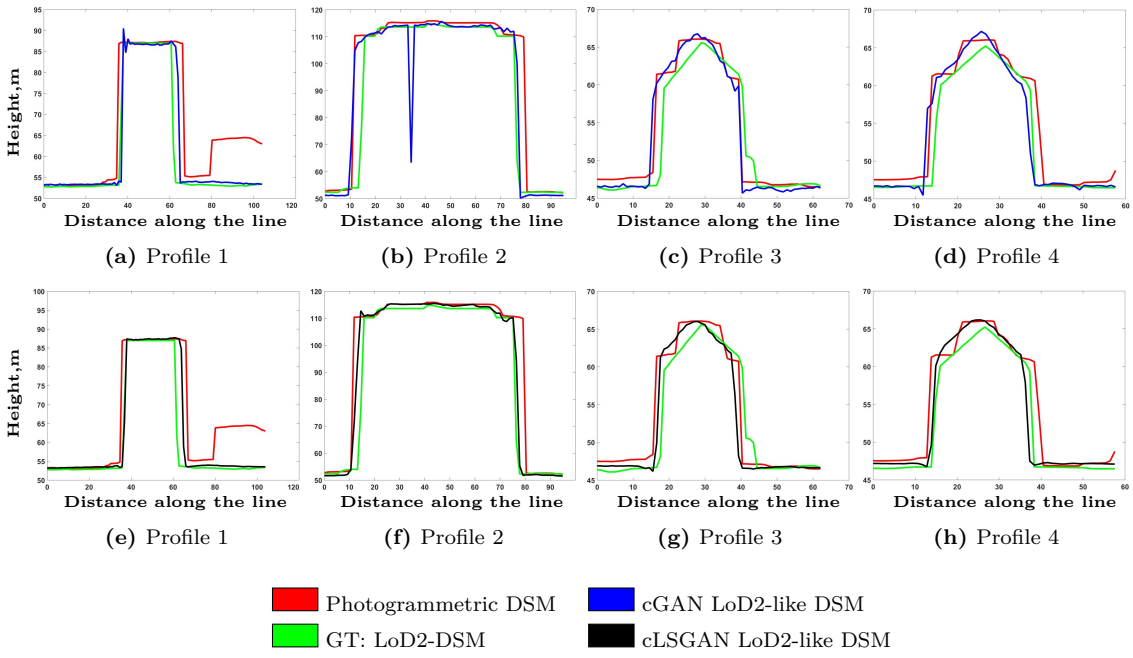



**Figure 5.9:** Visual analysis of selected building profiles (*cf.* Figure 5.5) from DSMs generated by cGAN (first line) and cLSGAN (second line) models in comparison to input poor quality DSMs and ground-truth LoD2-DSMs.

#### 5.2.3.2 Model Generalization Capability

To demonstrate the capability of the network to generalize on diverse urban landscapes, a 3D reconstruction of Munich city, Germany is performed. This dataset is different from the Berlin dataset. At first sight, the building architectural styles are similar to those in the city of Berlin, as both of them belong to the same city. However, the vast amount of construction within the cities excludes the possibility of meeting identical buildings. Moreover, the Munich and Berlin datasets have different absolute height values above sea level. Without re-training the model on the new dataset, we directly generate the Munich elevation image by passing WorldView-2 data through the DSM-to-LoD2 network trained on the LiDAR-DSM and LoD2-DSM. The $1000 \times 1000$ pixel examples of generated height images from both data types are illustrated in Figures 5.11b, 5.11c, 5.11f and 5.11g, respectively. From the resulting 3D height images, it can be seen that the proposed models successfully manage to generate the 3D elevation constructions over a new area. As expected, no new buildings are generated. However, some buildings are only partially reconstructed using both models. In example 1 of LiDAR-DSM in Figure 5.11b and LoD2-DSM in Figure 5.11c, the highlighted building is only partially reconstructed on both datasets. From the zoomed version of the highlighted buildings in Figure 5.12, we can clearly see that the quality of both input buildings from the photogrammetric DSM is quite low—not regular and very noisy. The shape of the first building (see Figure 5.12a) is not consistent, and the ridge line has a form close to a zigzag. The second building

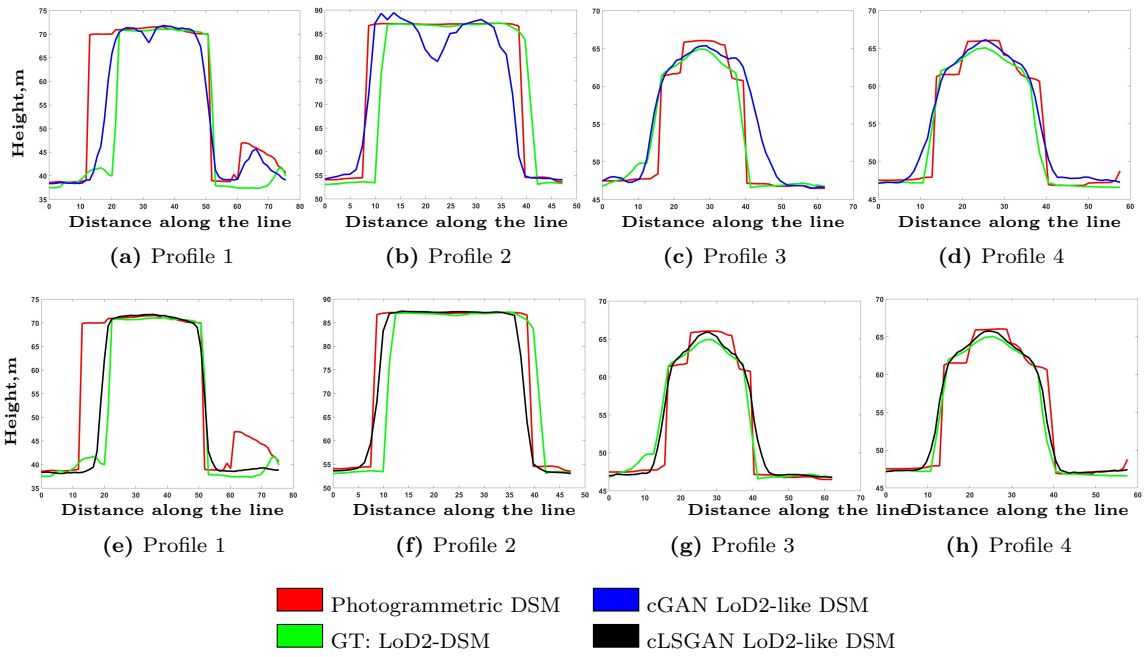
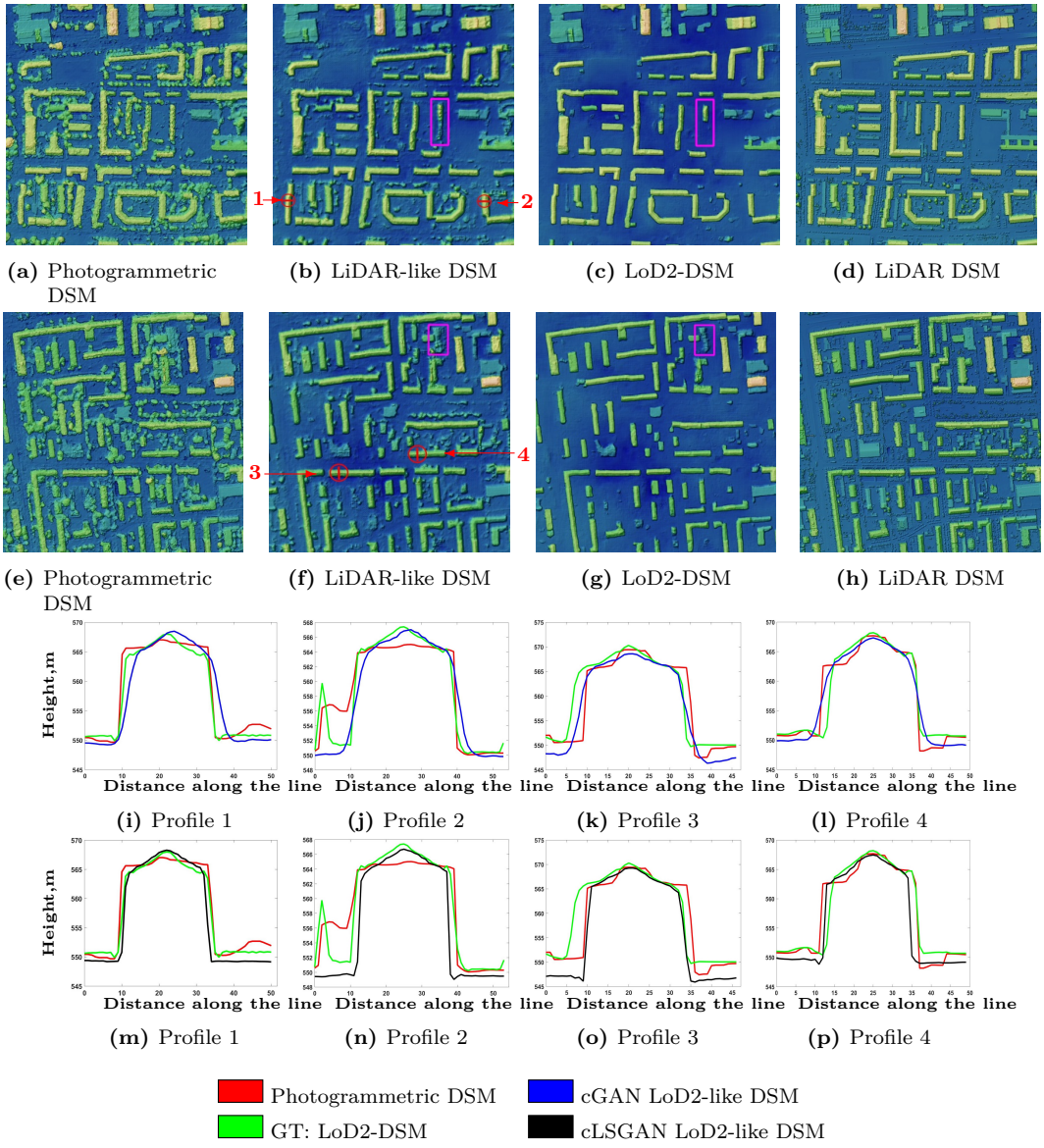



**Figure 5.10:** Visual analysis of selected building profiles (*cf.* Figure 5.6) from DSMs generated by cGAN (first line) and cLSGAN (second line) models in comparison to input poor quality DSMs and ground-truth LiDAR-DSMs.

in Figure 5.12e is most probably surrounded with vegetation and, due to interpolation processes, results in an object with irregular form. Even for the human eye, it is difficult to say if this object is a building. Therefore, both models experience problems with these buildings, which confirms our theory of poor quality data influence.

By investigating the presented profiles, it can be noticed that the roof shapes are improved by applying both models—the ridge lines are sharper and appear much better compared to the photogrammetric DSM. Comparing the roof profiles between the LiDAR-DSMs (third line in Figure 5.11) and LoD2-DSMs (fourth line in Figure 5.11), a small improvement in ridge line sharpness is produced with the LoD2-DSM model. Regarding the wall steepness, the buildings generated by the LoD2-DSM model are close to perpendicular wall planes in contrast to the one generated by applying the LiDAR-DSM model. This is reasonable, because the LiDAR-DSM is rasterized from a point cloud using interpolation and as a result features smooth transitions from the roof to the ground.

For the quantitative evaluation, it is assumed that the available LiDAR-DSM is to be our ground-truth, even when comparing the generated LoD2-like DSM, as no CityGML data is available for the Munich area. The statistical results of the experiment can be found in Table 5.2. The analysis of the results for the Munich area shows that the photogrammetric DSM produces closer results to the LiDAR-DSM, considered as the ground-truth, than both generated DSMs. This is due to the described effect of




(a) Photogrammetric DSM (b) LiDAR-like DSM (c) LoD2-DSM (d) LiDAR DSM

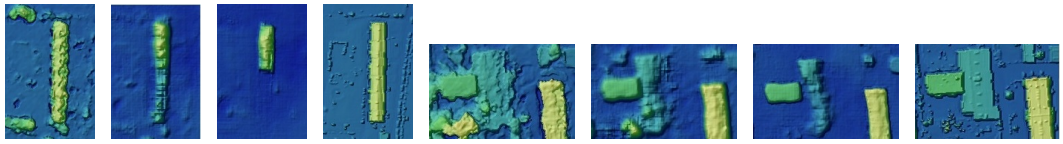
(e) Photogrammetric DSM (f) LiDAR-like DSM (g) LoD2-DSM (h) LiDAR DSM

(i) Profile 1 (j) Profile 2 (k) Profile 3 (l) Profile 4

(m) Profile 1 (n) Profile 2 (o) Profile 3 (p) Profile 4

**Figure 5.11:** Visual analysis of generalization by cLSGAN architecture over selected urban areas of the city of Munich using both LoD2-DSM and LiDAR-DSM setups. The DSM images are color-shaded for better visualization. Figures (a), (e) depict the input photogrammetric DSM data, (b), (f) is the generated DSM from LiDAR-DSM, (c), (g) is the generated DSM from LoD2-DSM and (d), (h) is the LiDAR ground-truth data depict the LiDAR ground-truth. The profiles of selected buildings from DSMs generated by the LiDAR-DSM setup are illustrated in the third line and the ones from the LoD2-DSM setup are in the fourth line.

unreconstructed buildings or their parts by both models from the poor quality input image. Additionally, the generated LiDAR-like DSM shows better results in comparison



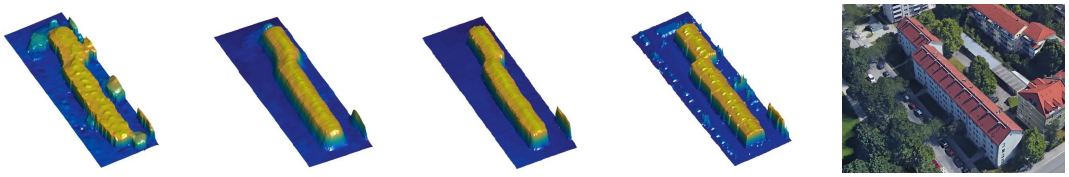
(a) (b) (c) (d) (e) (f) (g) (h)

**Figure 5.12:** Example of two buildings generated by the LiDAR-DSM setup (**b**,**f**) and the LoD2-DSM setup (**c**,**g**), respectively. (**a**,**e**) show the buildings on photogrammetric DSM and (**d**,**h**) on LiDAR DSM ground-truth data. The depicted examples are from the Munich area.

to the LoD2-like DSM due to the fact that the wall planes experience a smooth transition from the roof to the ground, which is similar to what occurred in the ground-truth data. In the LoD2-like DSM, this transition is more perpendicular and, as a result, farther from the considered ground-truth. This also influences the values of the metrics. However, as big improvements in shapes from analyzing the profiles are observed, it is also decided to evaluate the selected metric on a single building for which profile 1 is investigated. The obtained results are summarized in the second part of Table 5.2. In addition, the LoD2-like DSM even outperformed the LiDAR-like DSM, which was expected due to the fact that the CityGML data provides more regular and better quality building shapes than the LiDAR-DSMs and the network is able to learn these features. Therefore, it is proved that the proposed network is able to improve the low-quality building shapes. We also state that this kind of accuracy analysis is not generally suitable for a large area, but due to the lack of other potential evaluations, it is still used here. The improvements can be also seen in the three-dimensional visualization of the building geometry in Figure 5.13.

**Table 5.2:** Prediction accuracies of cGAN and cLSGAN models for all investigated metrics for LoD2-DSM and LiDAR-DSM datasets over the Munich area.

| Method | Munich area | | |
|---|---|---|---|
| | MAE (m) | RMSE (m) | NMAD (m) |
| Photogrammetric DSM | 2.10 | 4.68 | 0.92 |
| Generated LiDAR-like DSM | 2.53 | 4.88 | 1.41 |
| Generated LoD2-like DSM | 3.27 | 5.81 | 1.78 |
| | One Building | | |
| | MAE (m) | RMSE (m) | NMAD (m) |
| Photogrammetric DSM | 2.0 | 3.81 | 1.18 |
| Generated LiDAR-like DSM | 2.10 | 3.35 | 1.60 |
| Generated LoD2-like DSM | 1.87 | 3.39 | 1.48 |



**(a)** Input DSM **(b)** LiDAR-like DSM **(c)** LoD2-like DSM **(d)** LiDAR DSM **(e)** Google Earth View

**Figure 5.13:** Example of the generated building with a refined 3D shape for the city of Munich.

## 5.3 Multi-Task Network for Building Shape Improvements and Roof Type Understanding

Multi-task learning introduces the problem of optimizing the neural network model with respect to multiple objectives, because it requires modeling the trade-off between competing tasks [257]. In this work, the extraction of two tasks, mainly building shape improvement and roof surface classification, is performed within one model.

### 5.3.1 Methodology

The building shape refinement problem can be considered as a generative task, which has been recently successfully solved by GANs in other applications. To understand building geometry and semantics, we introduce architectures that learn to predict the pixel level depth and semantic classes from an input image. The network architecture presented by Isola *et al.* [67] is adapted for this purpose. The generator $G$ part of the network is modified from experiment to experiment within the entire study, while the discriminator $D$ part stays unchanged.

#### 5.3.1.1 One Generator, Two Outputs

As a continuation of the previous methodology introduced in Section 5.2, it is first considered that the generator $G$ of the cGAN consists of a one-stream *UNet* [179] with two outputs producing a single-channel depth image with continuous values and a three-channel roof class probability map. The detailed description of the UNet architecture used in this work has been depicted already in Figure 5.1 of Chapter 5 and discussed in Section 5.2.1.1. The tanh activation function (Equation (5.1)) is applied to the top layer of the generator $G$ responsible for depth image generation and the *softmax* normalization

$$\text{softmax}(\boldsymbol{z}) = \frac{e^{z_k}}{\sum_j e^{z_j}} \tag{5.12}$$

is applied in the training phase on top of the layer producing class probability maps. The encoder part of UNet progressively down-samples the given low-quality DSM through

eight layers and codes back the process with eight up-sampled decoder layers. To recover important details that are lost in down-sampling in the encoder, seven skip connections are added to the network.

Following the same strategy, a pre-trained *residual network (ResNet)* [61] consisting of 34 layers as a basis for the encoder part of the $G$ network is investigated. Architectures based on ResNets have already achieved state-of-the-art results and were successful in several competitions for image recognition tasks. One of the problems ResNets solve is an effect known as the vanishing gradient. When the network is too deep, the gradients from where the loss function is calculated can easily shrink to zero after several applications of the chain rule. This can lead to the problem that the weights never update their values and therefore, no learning is being performed. With ResNets, the gradients can flow directly through identity shortcut connections backward from later layers to initial filters. To complete the decoder part of the $G$ network, the feature maps that have been down-sampled by ResNet34 are up-sampled to obtain the resulting two outputs of the same size as the input image.

Lastly, the recently published *DeepLabv3+* [34] architecture is adapted to the multiple output $G$ network. A re-implementation of this architecture with a pre-trained ResNet101 is used. It utilizes a ResNet as a feature extractor to provide rich semantic information and uses *À trous Spatial Pyramid Pooling (ASPP)* [258] to preserve the spatial resolution at different rates. Thus, it provides the opportunity to refine the segmentation results, especially along object boundaries. The advantage of using à trous convolutions is that they allow one to expand the receptive field of filters to incorporate a larger context without increasing the number of weights. As a result, it offers an efficient mechanism to control the field-of-view and finds the best trade-off between accurate localization (small field-of-view) and context relation (large field-of-view).

The schematic representations of the proposed architectures are depicted in Figure 5.14a.

#### 5.3.1.2 Two Generators, Two Outputs

Depth regression and semantic segmentation representations are not the same, but can complement each other. Therefore, training only one common $G$ network for different problems is critical, especially because depth regression is a more complicated task and can negatively influence the final building outline results, while the segmentation has to follow the pattern of the building structure. Moreover, the intermediate features of 3D representations are different from 2D. This methodology aims at improving the roof forms and building shapes, specifically along the building borders by integrating the information about building outlines and roof types into the system. A cGAN model with two generators $G_1$ and $G_2$ is proposed which is responsible for better-quality DSM generation and building roof type pixel-wise classification mask production, respectively. At the beginning of the proposed architecture, two generators $G_1$ and $G_2$ are joined through two $1 \times 1$ convolutional layers with 8 and 32 channels, respectively. Coupling two tasks into a single model ensures that the model agrees between the independent task outputs while reducing computation time [114]. The schematic representations of

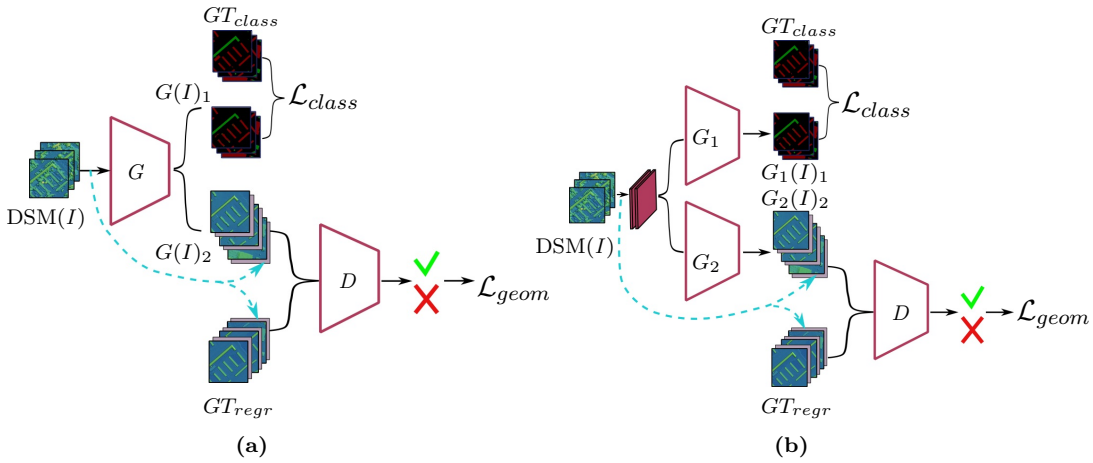



**Figure 5.14:** Schematic overview of two investigated architectures with (**a**) a one-stream generator $G$ and (**b**) a two-stream generator $G_1$ and $G_2$ for simultaneous building shape refinement $G(I)_2$ and roof classification map $G(I)_1$ generation. The input to both networks is a single photogrammetric *Digital Surface Model (DSM)* ($I$). The discriminator $D$ is identical for both models. The ground-truth for the regression task ($GT_{regr}$) is represented by *Level of Detail (LoD)*2-DSM generated from *City Geography Markup Language (CityGML)* data. The ground-truth for classification task ($GT_{class}$) is obtained from the orientation of the computed slope for each pixel. Each architecture is a *conditional Generative Adversarial Network (cGAN)* which conditions ( - - - - ) the model on side information such as input photogrammetric DSM. It is concatenated with either generated depth image $G(I)_2$ or ground-truth ($GT_{regr}$) as an additional channel (■) before going to the $D$ network. Although the multi-task problems $G(I)_2$ and $G(I)_1$ of the two-stream network are depicted as independent networks, in reality they are connected through the two $1 \times 1$ convolutional layers (■) with 8 and 32 channels, respectively. As a result, the joint loss function, which sums losses responsible for geometry reconstruction ($\mathcal{L}_{geom}$) and classification ($\mathcal{L}_{class}$), propagates back through the task-dependent layers, as well as the shared ones.

the proposed architectures are depicted in Figure 5.14b.

In this method, the constellations of (a) $G_1$: UNet and $G_2$: ResNet and (b) $G_1$: UNet and $G_2$: DeepLabv3+ are investigated. The reason for these combinations is that the UNet, in general, produces better 3D building representations than ResNet, and based on it DeepLabv3+, gives more accurate and complete semantic segmentation maps.

#### 5.3.1.3 Loss Function

The objective function of this method is based on *conditional Least Square Generative Adversarial Network (cLSGAN)* already discussed and presented by Equation (5.8) combined with $L_1$ distance loss function introduced in Equation (5.5) responsible for generating the synthesized image close to the given ground-truth.

To further refine the surface of roof planes, a normal vector loss

$$\mathcal{L}_{\text{normal}}(\mathcal{N}^{\text{t}}, \mathcal{N}^{\text{p}}) = \frac{1}{m} \sum_{i=1}^{m} \left( 1 - \frac{\langle \boldsymbol{n}_i^{\text{t}}, \boldsymbol{n}_i^{\text{p}} \rangle}{\|\boldsymbol{n}_i^{\text{t}}\| \, \|\boldsymbol{n}_i^{\text{p}}\|} \right), \tag{5.13}$$

is considered for the training, as proposed by Hu *et al.* [259], to measure the angle between the normal to the surface of an estimated DSM with respect to a target DSM. Here, $\mathcal{N}^{\mathrm{t}} = \{\boldsymbol{n}_1^{\mathrm{t}}, \ldots, \boldsymbol{n}_m^{\mathrm{t}}\}$ and $\mathcal{N}^{\mathrm{p}} = \{\boldsymbol{n}_1^{\mathrm{p}}, \ldots, \boldsymbol{n}_m^{\mathrm{p}}\}$ are normal vectors of the target and predicted DSMs, respectively, and $\langle \cdot, \cdot \rangle$ denotes the scalar product of two vectors. The loss is only computed within the building segments using an available binary building mask.

To learn pixel-wise roof type class probabilities, the cross-entropy loss function

$$\mathcal{L}_{\mathrm{CE}}(\boldsymbol{x}, \boldsymbol{t}, \boldsymbol{p}) = -\sum_i t_i \log p(x_i) \tag{5.14}$$

is used paired with softmax normalization (see Equation (5.12)) applied to the neural network outputs $z_k$ before the cross-entropy loss computation. Here, $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ is the set of input examples in the training dataset and $\boldsymbol{t} = \{t_1, \ldots, t_n\}$ is the corresponding set of ground-truth values for the input examples.

To train the multi-task jointly, the losses of each individual task are summed in a weighted linear manner together with losses responsible for image appearance refinement. This leads to the final combined objective function:

$$G^{\star} = \arg\min_G \max_D \underbrace{\mathcal{L}_{\mathrm{cLSGAN}}(G, D) + \lambda \cdot \mathcal{L}_{L_1}(G) + \gamma \cdot \mathcal{L}_{\mathrm{normal}}(G)}_{\mathcal{L}_{\mathrm{geom}}} + \underbrace{\beta \cdot \mathcal{L}_{\mathrm{CE}}(G)}_{\mathcal{L}_{\mathrm{class}}}, \tag{5.15}$$

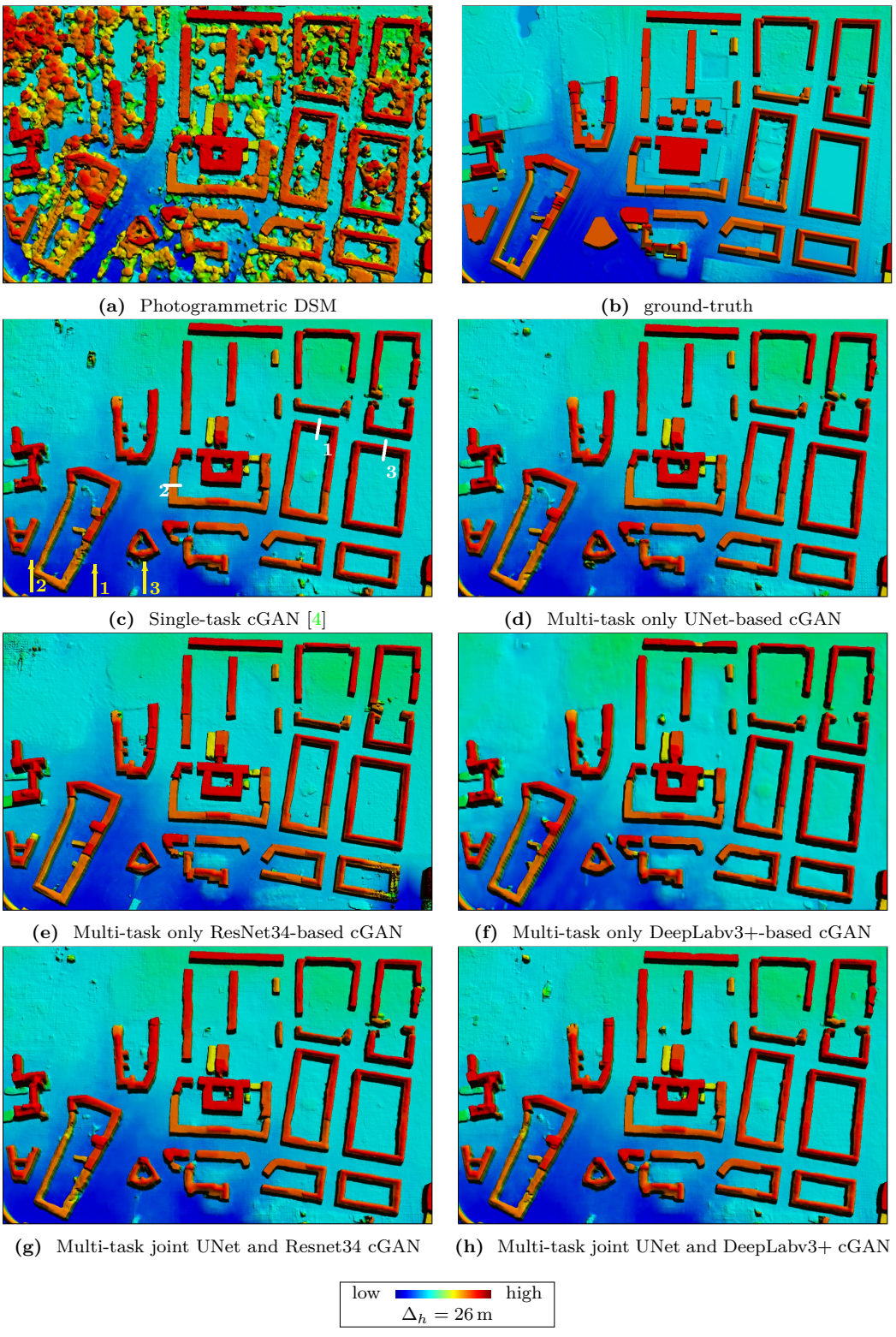
where $0 \leq \lambda, \beta, \gamma \in \mathbb{R}$ are the weighting hyper-parameters.

### 5.3.2 Study Area and Experiments

Experiments have been carried on the photogrammetric DSMs as input images and LoD2-DSMs as ground-truth images showing the city of Berlin, Germany, described in Section 5.2.2.1.

To generate the ground-truth for the pixel-wise classification task, the obtained LoD2-DSM is used to compute the slope for each pixel within the whole image as the maximum rate of change of elevation between that pixel and its surroundings. The aspect was then defined as the orientation of the computed slope, which was measured clockwise in degrees from 0° to 360°, where 0° is north-facing, 90° is east-facing, 180° is south-facing, and 270° is west-facing. The area that did not correspond to buildings was set to Class 0, background, and 90 degrees to Class 1, flat roofs, and the rest of the degree values were set to Class 2, sloped roofs.

The proposed multi-task cGAN implementation is an extension of the DSM-to-LoD2 network described in Section 5.2.2.2 which was developed on the *PyTorch* Python package based on the *pix2pix* software introduced by Isola *et al.* [67]. We kept the training and inference setups the same as in Sections 5.2.2.2 and 5.2.2.3 to be able to compare the obtained results of proposed methodologies. The weighting hyper-parameters $0 \leq \lambda, \beta, \gamma \in \mathbb{R}$ were set to $\lambda = 1000$, $\beta = 10$ and $\gamma = 10$ after performing training and examining the resulting generated images from the validation dataset.




**(a)** Photogrammetric DSM **(b)** ground-truth

**(c)** Single-task cGAN [4] **(d)** Multi-task only UNet-based cGAN

**(e)** Multi-task only ResNet34-based cGAN **(f)** Multi-task only DeepLabv3+-based cGAN

**(g)** Multi-task joint UNet and Resnet34 cGAN **(h)** Multi-task joint UNet and DeepLabv3+ cGAN

**Figure 5.15:** Visual comparison of DSMs over selected urban area, generated by a cGAN with least squares residuals using (c) the one-stream generator network for a single task [4], (d) the one-stream generator based on the UNet network for multiple tasks, (e) the one-stream generator based on ResNet34 network for multiple tasks, (f) the one-stream generator based on the DeepLabv3+ network for multiple tasks, (g) the two-stream generator network with jointly trained UNet and ResNet34 architectures for multiple tasks, and (h) the two-stream generator network with jointly trained UNet and DeepLab architectures for multiple tasks. (a) illustrates the input photogrammetric DSM, and (b) demonstrates the ground-truth data. The DSMs images are color-shaded for better visualization.

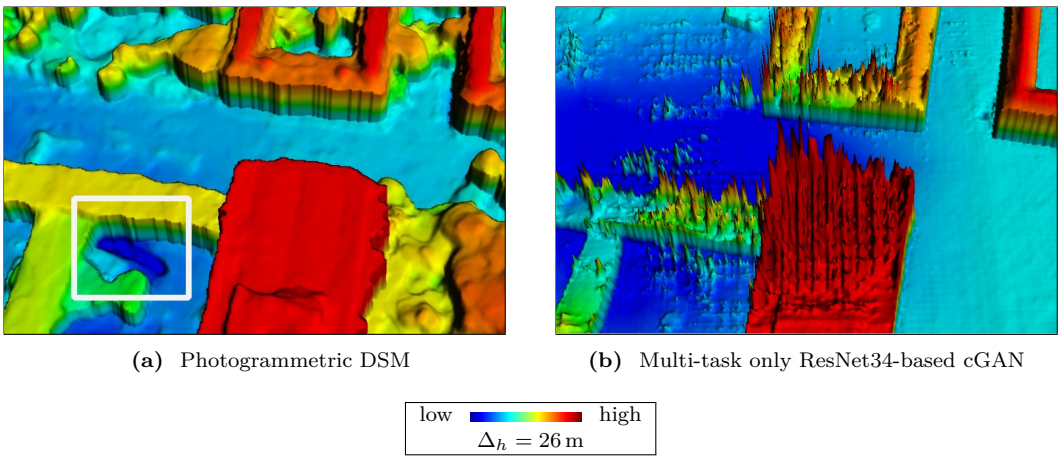
### 5.3.3 Results

In this section, several experiments are performed on simultaneous depth image generation with good-quality building shapes and pixel-wise building roof type classification map extraction. First, in comparing a single-task result to multi-task results in Figure 5.15, it can be can seen that the integration of the semantic segmentation task, even for the single-stream network, already improves the results, although two outputs are directly produced from the single-stream network illustrated in Figure 5.14b. One of the examples highlighted by the number "1" in Figure 5.15c do not have the complete structure by using a single output network from Bittner *et al.* [4]. However, the results obtained by multi-task networks (see Figures 5.15d, 5.15e, 5.15f, 5.15g and 5.15h) are able to further improve its shape. Moreover, it can be noticed that A-shaped building "2" and building "3", highlighted in Figure 5.15c, are also more accurate and very close to the corresponding building in the ground-truth.

At first sight, the ResNet34-based network demonstrates better results (see Figure 5.15e). The outlines of the buildings are more rectilinear, and the ridge lines are more distinguishable. However, one can notice the inability of the model to reconstruct the building in the lower-right corner correctly. This is due to the incorrect height information presented in the input photogrammetric DSM. In the detailed view illustrated in Figure 5.16a, the highlighted area depicts a recess in the ground. This incorrectly reconstructed part of the photogrammetric DSM happens usually due to occlusion of this area by the building walls or trees. As a result, while reconstructing this area with the ResNet34-based network, this error propagated within network layers as the receptive field grew, and influenced the reconstruction of the whole patch. This can be identified by the dark blue area in Figure 5.16b. The same phenomenon happens with other architectures as well, but with less strength. The examples are depicted in Figure 5.17. All generated results undergo the failure influence highlighted in Figure 5.17a. However, the UNet in Figure 5.17b and the DeepLabv3+ in Figure 5.17d generate better results compared to the ResNet34 in Figure 5.17c. The propagation of incorrect values is less intensive and wide and in the case of DeepLabv3+, even narrower than the UNet results.

Examining the profiles in Figures 5.18g, 5.18h and 5.18i of three selected buildings (highlighted with white color in Figure 5.15c), it can be observed that the roof plane reconstruction results are far from acceptable compared to the ones produced by UNet-based architectures. It can be concluded that for such complicated tasks as the 3D reconstruction of miniscule objects from satellite images, compared to the big size of objects in media images commonly used in the computer vision field, that the skip connections are needed. The skip connections carry detailed and fine information from lower layers, which is added to the up-sampled feature maps and helps to refine the final output [59].

Investigating Figure 5.15f, it can be said that the model based on the DeepLabv3+ architecture is able to generate depth maps with reasonable building forms regarding building boundaries. However, one can also notice that the walls of the buildings look different compared to the results obtained from the rest of the models. Going deeper and examining the profiles of selected buildings in Figures 5.18j, 5.18k and 5.18l, it can




**(a)** Photogrammetric DSM **(b)** Multi-task only ResNet34-based cGAN

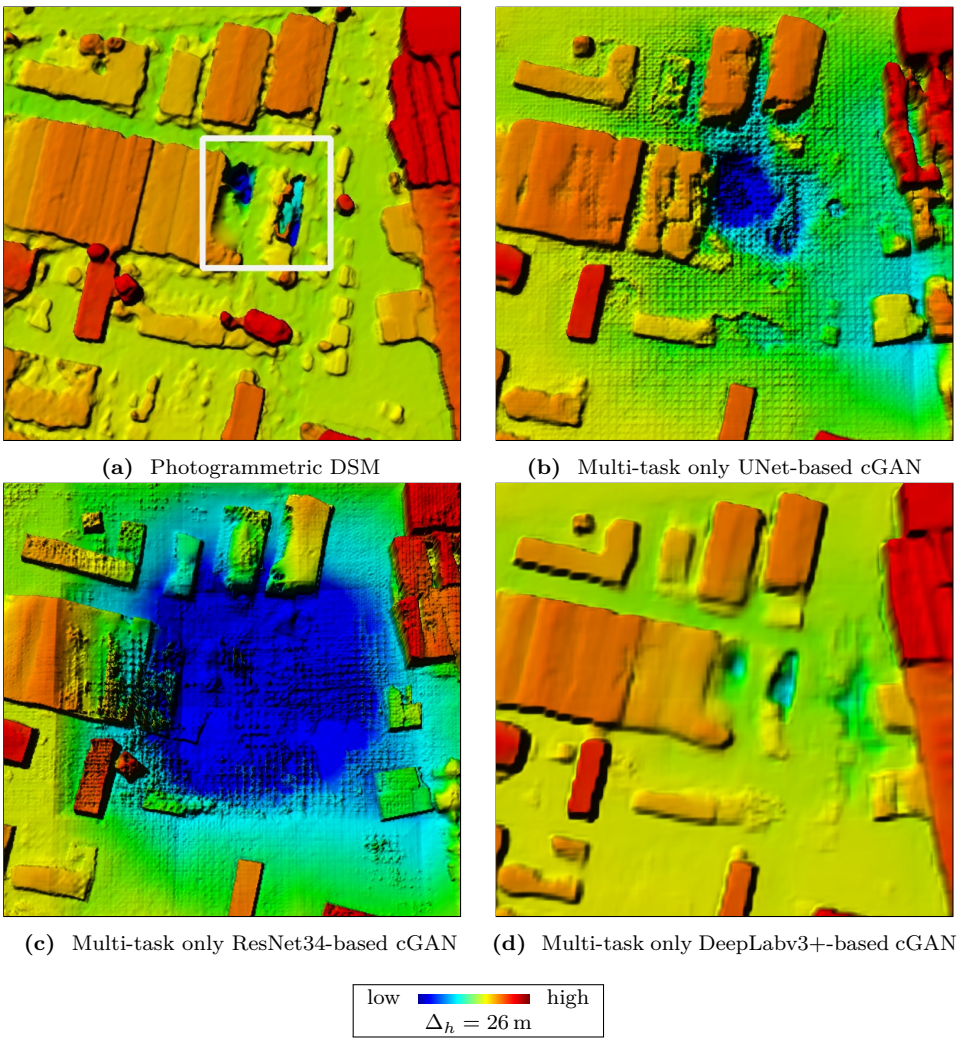
**Figure 5.16:** A detailed demonstration of a failure case on the generated LoD2-like DSM obtained by the ResNet34-based network Figure 5.14a architecture. (a) depicts the input photogrammetric DSM, and (b) shows the resulted ResNet34-based DSM from Figure 5.15e.

be seen that these buildings have a smooth transition from roof to ground, similar to a Gaussian form, compared to the DSMs generated by other networks. This effect can also be seen in Figure 5.19f.

Adding the normal vector loss function to the model helps in further refining the roof surfaces, making the roof planes flatter and more realistic. This beneficial effect can be seen comparing the profiles of selected buildings from the cGAN DSM [4] with profiles from the presented multi-task cGAN models. Moreover, the zoom-in view in Figure 5.19 confirms this statement, as the roof planes look smoother but at the same time keep the right form and sharp ridge lines. This is reasonable, as the models are pushed to learn roof surface representations where the normal vectors, which belong to the same plane, look in the same direction and close to the ground-truth.

From visual comparison between the ground-truth and results from all proposed networks, one can also conclude that the networks do not generate new buildings where no buildings are placed on the low-quality input image, but rather try to improve the available ones, even though some inconsistency occurs between existing and no longer existing buildings during the training. A good example can be seen in Figure 5.15b, where five small buildings in the middle of the scene are located. One can notice that in Figure 5.15a, only two different buildings are existing. This problem is not unique and can be faced in many cases during the training due to the time difference between the given photogrammetric DSM and the ground-truth LoD2-DSM. However, the proposed models are robust to such differences and do not produce "ghost" constructions.

The results of the roof classification task, simultaneously obtained in parallel with good-quality LoD2-like DSM generation, are presented in Figures 5.20 and 5.21. In the first region in Figure 5.20, it can be seen that masks, generated by ResNet34 and DeepLabv3+, provide better results for building boundaries, as well as for class separation compared to UNet-based results. Analyzing the ResNet34 and DeepLabv3+




**(a)** Photogrammetric DSM **(b)** Multi-task only UNet-based cGAN

**(c)** Multi-task only ResNet34-based cGAN **(d)** Multi-task only DeepLabv3+-based cGAN

**Figure 5.17:** A detailed demonstration of a failure case example on generated LoD2-like DSMs obtained by the UNet-, ResNet34-, and DeepLabv3+-based network Figure 5.14a architectures. (a) depicts the input photogrammetric DSM with the area highlighting the presented incorrect height information and its influence on the reconstructed LoD2-like DSMs from (b) multi-task only UNet-based cGAN, (c) multi-task only ResNet-based cGAN, and (d) multi-task only DeepLabv3+-based cGAN. The area that undergoes the influence is presented as a darker blue shade around the location where the failure is originated in (a).

results, one can observe that the DeepLabv3+ network provides more accurate classification. Good examples are Buildings "1", "2", "3", "4" depicted in Figure 5.20f. The ResNet34 model is only able to partially classify the building roof correctly, while the DeepLabv3+ set the right classes to them. Moreover, investigating the buildings highlighted as "2" and "3" in Figure 5.20f, one can conclude that the separate training of task-specific problems from some point in the network positively influences the final results compared to the multi-task network, which has a common body for several

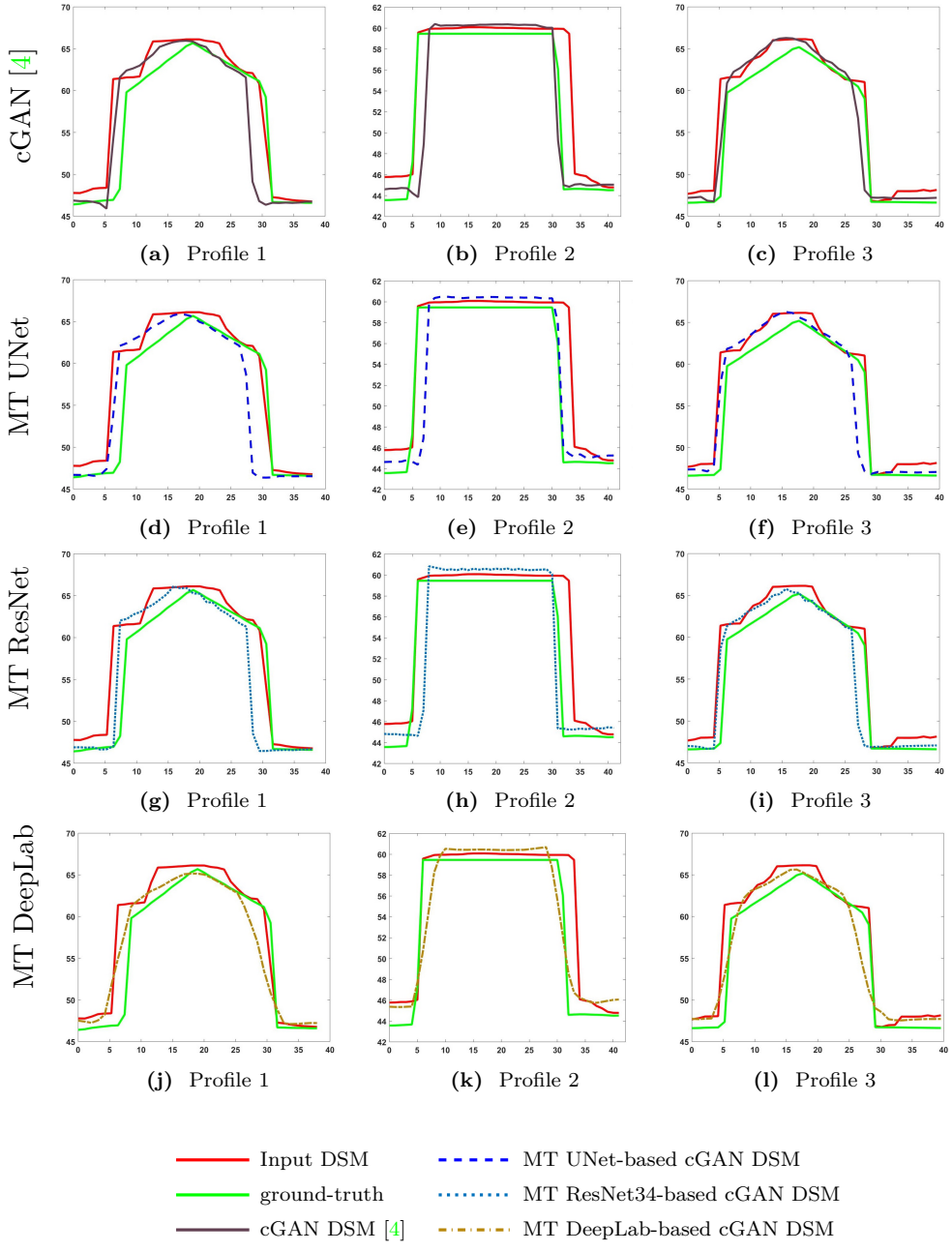



**Figure 5.18:** Illustration of the profiles for three selected buildings (*cf.* Figure 5.15c) from DSMs generated by (a)–(c) the cGAN model [4], (d)–(f) the multi-task only UNet-based cGAN, (g)–(i) the multi-task only ResNet34-based cGAN, and (j)–(l) the multi-task only DeepLabv3+-based cGAN. The results from the second, third, and fourth lines are generated by a one-generator, two-output network, depicted in Figure 5.14a.

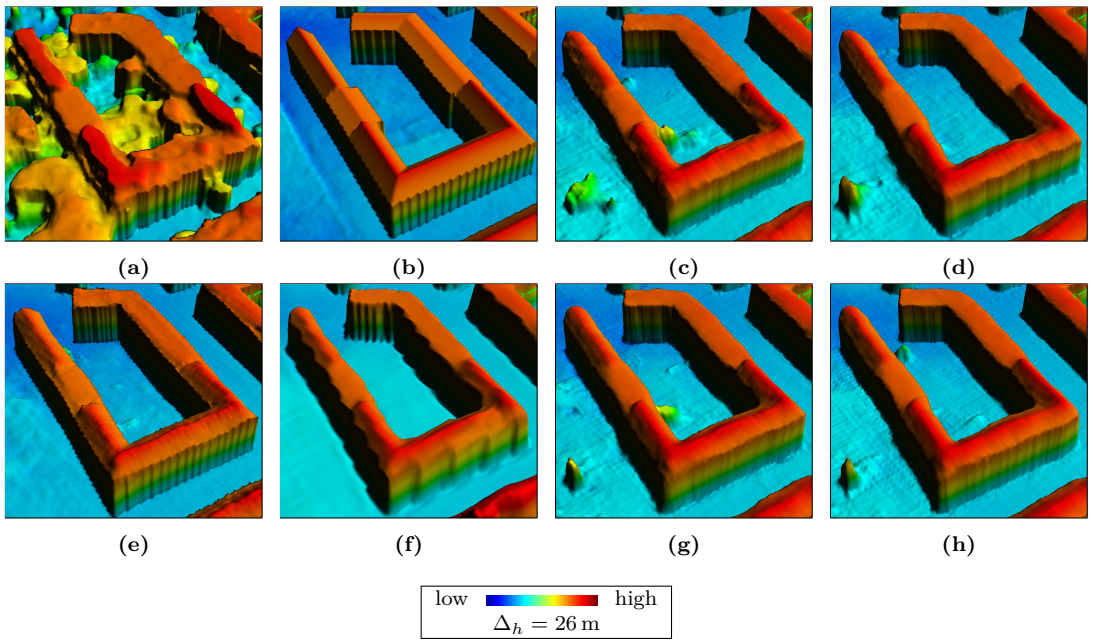



**Figure 5.19:** Comparison of the generalization over DSMs from (c) the one-stream generator network for a single task [4], (d) the one-stream generator based on the UNet network for multiple tasks, (e) the one-stream generator based on the ResNet34 network for multiple tasks, (f) the one-stream generator based on the DeepLabv3+ network for multiple tasks, (g) the two-stream generator network jointly trained UNet and ResNet34 architectures for multiple tasks, and (h) the two-stream generator network jointly trained UNet and DeepLabv3+ architectures for multiple tasks. (a) illustrates the input photogrammetric DSM, and (b) demonstrates the ground-truth data.

task-specific outputs (see Figures 5.20b, 5.20c and 5.20d). A larger difference shows up when investigating the residential area with small single-family houses. Looking at Figure 5.21, it can be observed that more small buildings are extracted by the DeepLabv3+ network, compared to other networks.

Besides visual inspections of refined depth images and pixel-wise classification maps, the following error metrics commonly used in relevant publications [256, 260–262] are investigated. The $\varepsilon_{MAE}$, $\varepsilon_{RMSE}$ and $\varepsilon_{NMAD}$ are used for depth evaluation and are presented in Section 5.2.2.4 by Equations (5.9), (5.10) and (5.11), respectively.

For the semantic segmentation task evaluation, the common *Intersection over Union (IoU)*

$$IoU = \frac{target \cap prediction}{target \cup prediction}, \tag{5.16}$$

the *Precision*

$$Precision = \frac{TP}{TP + FP}, \tag{5.17}$$

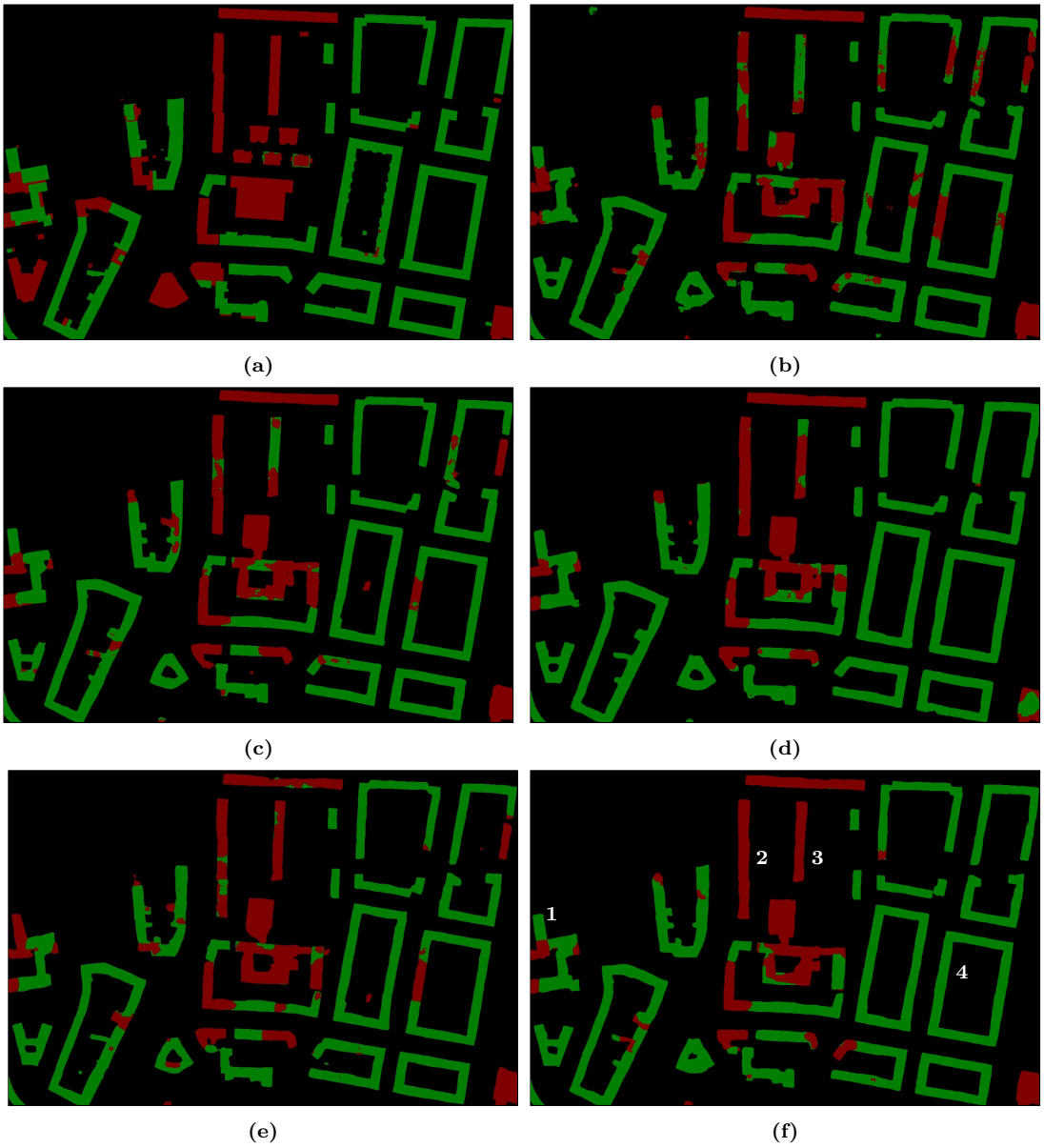



**Figure 5.20:** Visual comparison of roof classification maps over selected urban areas, generated by cGAN with least squares residuals using (b) the one-stream generator based on the UNet network for multiple tasks, (c) the one-stream generator based on the ResNet34 network for multiple tasks, (d) the one-stream generator based on the DeepLabv3+ network for multiple tasks, (e) the two-stream generator network jointly trained UNet and ResNet34 architectures for multiple tasks, (f) the two-stream generator network jointly trained UNet and DeepLab architectures for multiple tasks. (a) illustrates the ground-truth label mask.

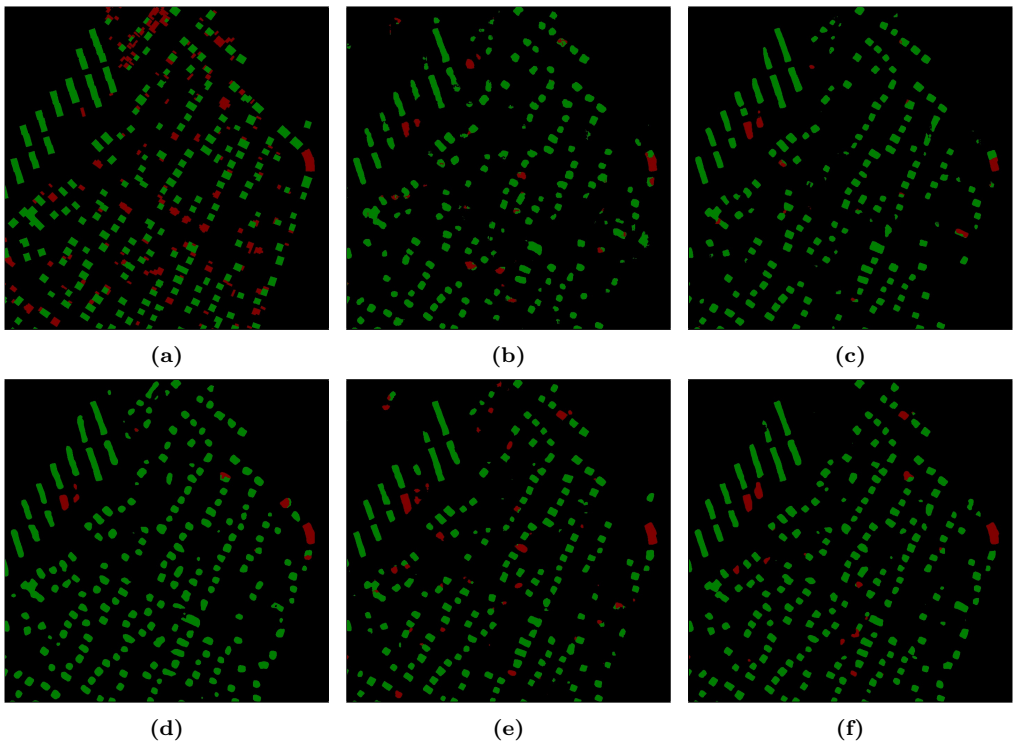



**Figure 5.21:** Visual comparison of roof classification maps over selected urban areas, generated by cGAN with least squares residuals using (b) the one-stream generator based on the UNet network for multiple tasks, (c) the one-stream generator based on the ResNet34 network for multiple tasks, (d) the one-stream generator based on the DeepLabv3+ network for multiple tasks, (e) the two-stream generator network jointly trained UNet and ResNet34 architectures for multiple tasks, and (f) the two-stream generator network jointly trained UNet and DeepLab architectures for multiple tasks. (a) depicts ground-truth label mask.

the *Recall*

$$Recall = \frac{TP}{TP + FN}, \tag{5.18}$$

and the *F1-score*

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{5.19}$$

are used. The IoU metric measures how much overlap exists between two regions. It is calculated separately for each class and then averaged over all classes to provide a global statistic. Precision answers the question about the correct proportion of positive identifications. Recall shows how well all the positives are found. For the semantic segmentation task, an evaluation over the whole test area is performed.

The evaluation results for depth map generation and roof classification tasks are pre-

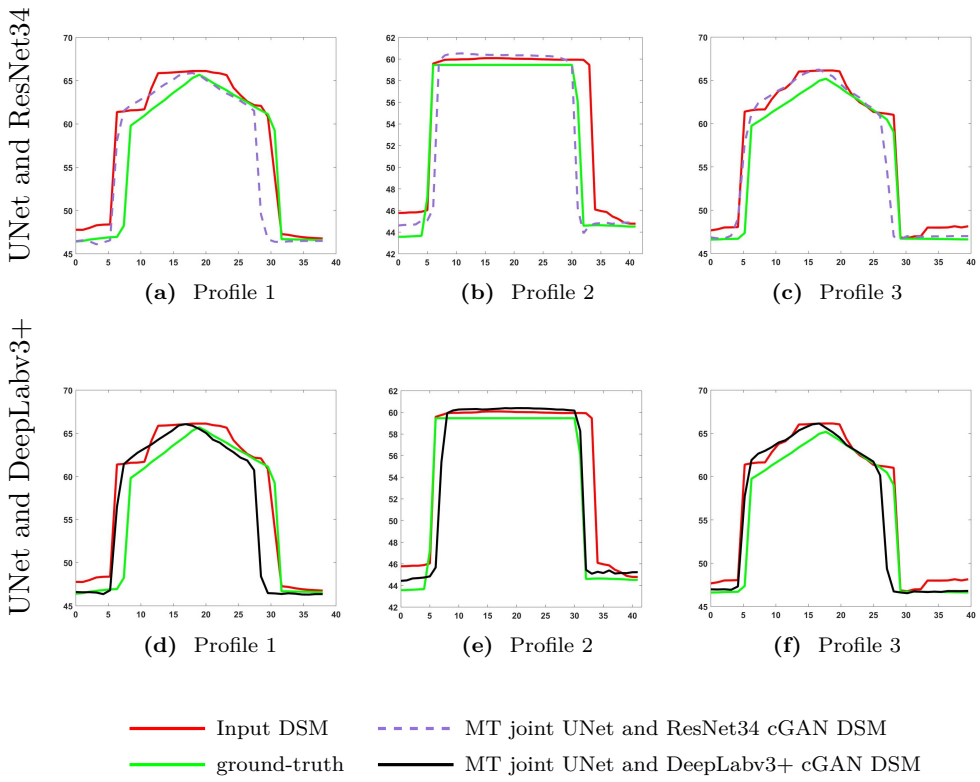



**Figure 5.22:** Illustration of the profiles for three selected buildings (*cf.* Figure 5.15c) from DSMs generated by (a)–(c) the two-stream generator network jointly trained UNet and ResNet34 architectures for multiple tasks and (c)–(f) the two-stream generator network jointly trained UNet and DeepLabv3+ architectures for multiple tasks. The results are generated by the two-generator, two-output network depicted in Figure 5.14b.

sented in Tables 5.3 and 5.4, respectively. In terms of $\varepsilon_{RMSE}$, the DeepLab-based cGAN network demonstrates the best result. As already seen from the profiles in Figure 5.18(j)–(l), the surfaces of the roofs present the smoothest results compared to other profiles. Moreover, from the semantic segmentation results shown in Table 5.4, the DeepLab-based network demonstrates the best results in terms of classification and outlines of buildings. As a result, those facts are reflected in the *RMSE* error.

Regarding the $\varepsilon_{\text{NMAD}}$ metric, one can notice that the evaluation result has the lowest value of 0.88 m and is the same for both the cGAN [4] and the joint UNet and the DeepLabv3+ network, although the $\varepsilon_{\text{RMSE}}$ is different. This can be due to only selecting the median height error value, which does not reflect the true error distribution. It can be clearly observed from the qualitative results shown in Figures 5.15f, 5.19h and 5.22f that the proposed joint UNet and DeepLabv3+ network can offer significantly improved roof surface qualities with noticeably smoother planes than that provided by the other tested methods. This observation suggests the need to develop new evaluation metrics that can assess the roof surface planarity better than the existing metrics.

**Table 5.3:** Quantitative results for the *Root Mean Square Error (RMSE)*, *Normalized Median Absolute Deviation (NMAD)*, and *Mean Absolute Error (MAE)* metrics evaluated on 12 selected buildings existing for both the photogrammetric DSM and the ground-truth LoD2-DSM of the area depicted in Figure 5.15.

| Method | Error | | |
|---|---|---|---|
| | RMSE (m) | NMAD (m) | MAE (m) |
| cGAN [4] | 3.29 | 0.88 | 1.78 |
| only UNet based | 3.20 | 0.91 | 1.71 |
| only ResNet34 based | 3.23 | 0.96 | 1.71 |
| only DeepLabv3+ based | 2.51 | 1.07 | 1.51 |
| joint UNet and ResNet34 | 3.21 | 0.89 | 1.72 |
| joint UNet and DeepLabv3+ | 3.12 | 0.90 | 1.69 |

**Table 5.4:** Quantitative results for the IoU, F1-score, precision, and recall metrics evaluated on the test area covering 50 $km^2$.

| Method | Error | | | |
|---|---|---|---|---|
| | IoU (%) | F1-Score (%) | Precision (%) | Recall (%) |
| only UNet based | 59.78 | 72.07 | 77.05 | 48.43 |
| only ResNet34 based | 61.05 | 73.28 | 79.55 | 51.64 |
| only DeepLabv3+ based | 62.73 | 74.83 | 78.59 | 52.18 |
| joint UNet and ResNet | 61.54 | 73.73 | 79.28 | 51.80 |
| joint UNet and DeepLabv3+ | **64.44** | **76.34** | **80.03** | **55.2** |

In general, it can be concluded from both qualitative and quantitative evaluation that an improved building boundary reconstruction, together with correct class label assignments, is positively influencing the whole elevation model generation. This is also visually confirmed by investigating the profiles from joint UNet and DeepLab networks illustrated in Figure 5.22(d)–(f).

## 5.4 Information Fusion from Depth and Intensity Data for Large-Scale Digital Surface Model Refinement

It is common in the field of remote sensing to fuse data of different modalities to complement missing knowledge. In our earlier work [263], a WNet-cGAN network which merges depth and spectra information within an end-to-end framework was introduced. Fusing data from separate networks—which are fed with *pan-chromatic (PAN)* images and DSMs—is performed at a *later* stage right before producing the final output.

Following up to the aforementioned approach, the fusion of spectral (Figure 5.23a) and height (Figure 5.23b) information at an *earlier* stage within an end-to-end Hybrid-cGAN network is investigated to further improve small and simple residential buildings, as well as complex industrial ones. An auxiliary *normal vector loss* term is also added to

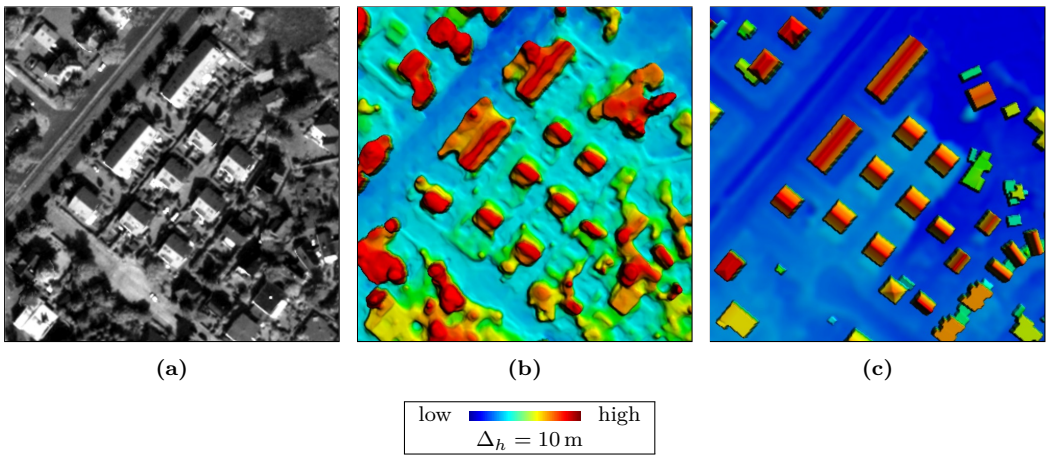



**Figure 5.23:** Sample of area from our dataset illustrated both inputs to the network (a) PAN image and (b) photogrammetric DSM, and (c) the ground-truth LoD-2-DSM. The DSM images are color-shaded for better visualization.

the objective function enforcing the model to produce more planar and flat roof surfaces, similar to the desired LoD2-DSM (Figure 5.23c) synthetically produced from CityGML data.

## 5.4.1 Methodology

### 5.4.1.1 Network Architecture

Because each photogrammetric DSM is a product obtained from pan-chromatic image pairs, it is reasonable to integrate depth and spectral data into one single network, as the latter provides sharper information about building silhouettes, which creates a better reconstruction of missing building parts and also the refinement of building outlines.

#### 5.4.1.1.1 Late Fusion

The first investigation was done by us in work [263], where two separate but identical U-form networks are fused at the later end within the $G$ part of a cGAN, where the first stream is fed with the PAN image and the second stream with the photogrammetric DSM, creating a so-called WNet architecture illustrated in Figure 5.24. The encoder and the decoder of each separate U-form architecture consists of 8 and 7 convolutional layers, respectively, with 7 skip connections, where the intermediate features from both streams are concatenated right before the last up-sampling which brings the features to the final output size. At the end, the WNet is increased with an additional convolutional layer of size $1 \times 1$ after the last up-sampling for better information fusion of different modalities. Each convolutional layer is followed by a *Leaky Rectified Linear Unit (LReLU)* [264]

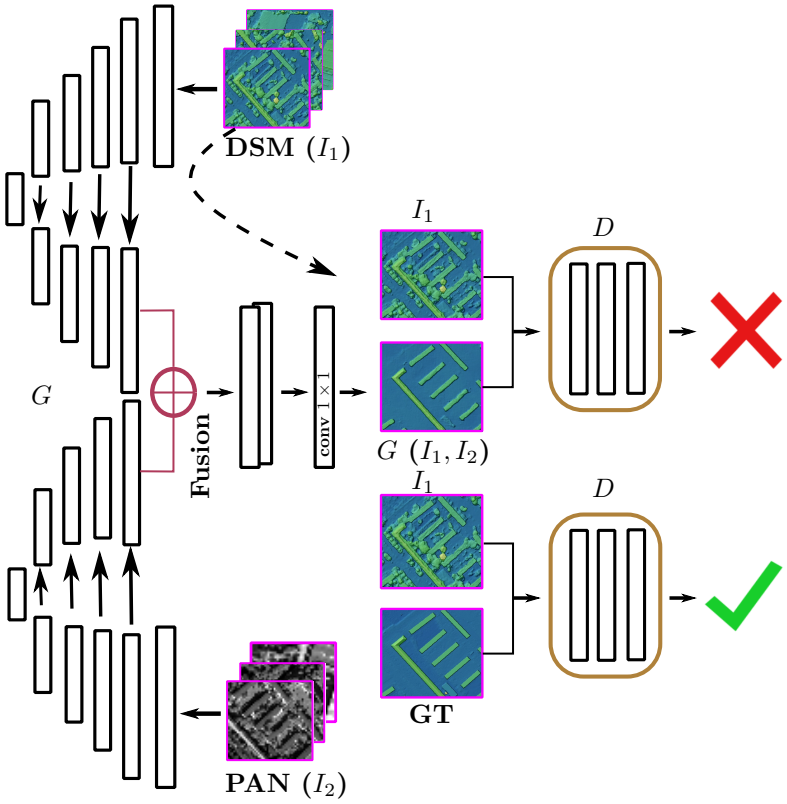



**Figure 5.24:** Schematic overview of the proposed late fusion architecture for the building shape refinement in the 3D surface model by WNet-cGAN using depth and spectral information. The illustration is adapted from Bittner *et al.* [263].

activation function

$$\sigma_{\mathrm{LReLU}}(z) = \begin{cases} z, & \text{if } z > 0 \\ az, & \text{otherwise} \end{cases}, a \in \mathbb{R}^+, \tag{5.20}$$

and *Batch Normalization (BN)* in case of the encoder, and a *Rectified Linear Unit (ReLU)* activation function

$$\sigma_{\mathrm{ReLU}}(z) = \begin{cases} z, & \text{if } z > 0 \\ 0, & \text{otherwise} \end{cases} \tag{5.21}$$

and BN in case of the decoder. On top of the generator network $G$, the tanh activation function (Equation (5.1)) is applied.

#### 5.4.1.1.2 Earlier Fusion

In the second investigation, the potential of an earlier fusion of data from different

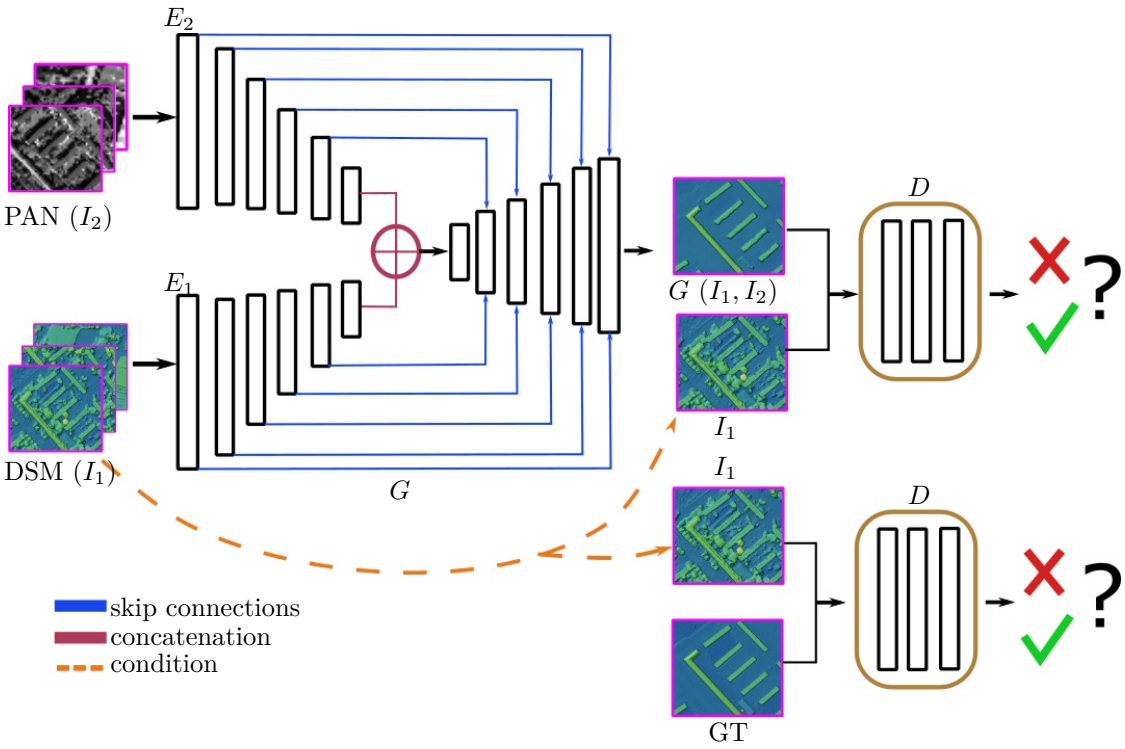



**Figure 5.25:** Schematic overview of the proposed earlier fusion architecture for the building shape refinement on photogrammetric DSMs by Hybrid-cGAN using both depth and spectral information.

modalities is examined, because it could potentially improve blending together the depth and spectral information. The generator $G$ of the proposed Hybrid-cGAN network depicted in Figure 5.25 mainly consists of two encoders $E_1$ and $E_2$, concatenated at the top layer, and a common decoder, which integrates information from two different modalities and generates an LoD2-like DSM with refined building shapes. The inputs to $E_1$ are the single-channel orthorectified PAN images, while $E_2$ receives the single-channel photogrammetric DSMs with continuous values. Since intensity and depth information have different physical meanings, it is unlikely that jointly propagating them at the beginning will be effective. It is reasonable to separate them first and allow the network to learn the most valuable features from each modality itself. The generator $G$ is constructed by a U-form network with 14 skip connections from both the spectral stream and the depth stream allowing the decoder to learn back detailed features that were lost by pooling in the encoders. In particular, the encoder and decoder consist of 8 convolutional layers each with LReLU activation function presented by Equation (5.20) and BN in case of the encoder, and a ReLU activation function presented by Equation (5.21) and BN in case of the decoder. The tanh activation function is placed on top of the generator network $G$.

The discriminator network $D$ for both WNet [263] and Hybrid-cGAN is a binary clas-

sification network constructed with 5 convolutional layers, followed by LReLU activation function and a BN layer. It had a *sigmoid* activation function Equation (5.2) at the top layer to output the likelihood that the input image belongs either to class 1 ("real") or class 0 ("generated").

#### 5.4.1.2 Objective function

The objective function of both WNet [263] and Hybrid-cGAN models is based on cLSGAN combined with $L_1$ distance loss function and represented by Equation (5.8). However, the objective function

$$G^{\star} = \arg\min_{G}\max_{D} \mathcal{L}_{\text{cLSGAN}}(G,D) + \lambda\mathcal{L}_{L1}(G) + \gamma\mathcal{L}_{\text{ normal}}\,, \tag{5.22}$$

for Hybrid-cGAN network was additionally extended by the normal vector loss function (*cf.* Equation (5.13)) responsible for the surface of roof planes refinement.

### 5.4.2 Study Area and Experiments

Experiments have been carried out on the photogrammetric DSMs as input images and LoD2-DSMs as ground-truth images showing the city of Berlin, Germany, described in Section 5.2.2.1. The intensity information has been introduced to the network from half-meter resolution orthorectified PAN images showing the closest nadir view among six pan-chromatic Worldview-1 images acquired on two different days.

The networks $G$ and $D$ were trained at the same time all along the training phase by alternating one gradient descent step of $D$ and one gradient descent step of $G$. During the inference process, only the trained generator model $G$ of the Hybrid-cGAN network or WNet [263] network was involved.

The training and inference setups were kept the same as in Sections 5.2.2.2, 5.2.2.3 and 5.3.2 in order to compare the obtained results of the proposed methodologies. The weighting hyper-parameters $0 \leq \lambda, \gamma \in \mathbb{R}$ were set to $\lambda = 1000$ and $\gamma = 10$ after performing training and examining the resulting generated images from the validation dataset.

### 5.4.3 Results

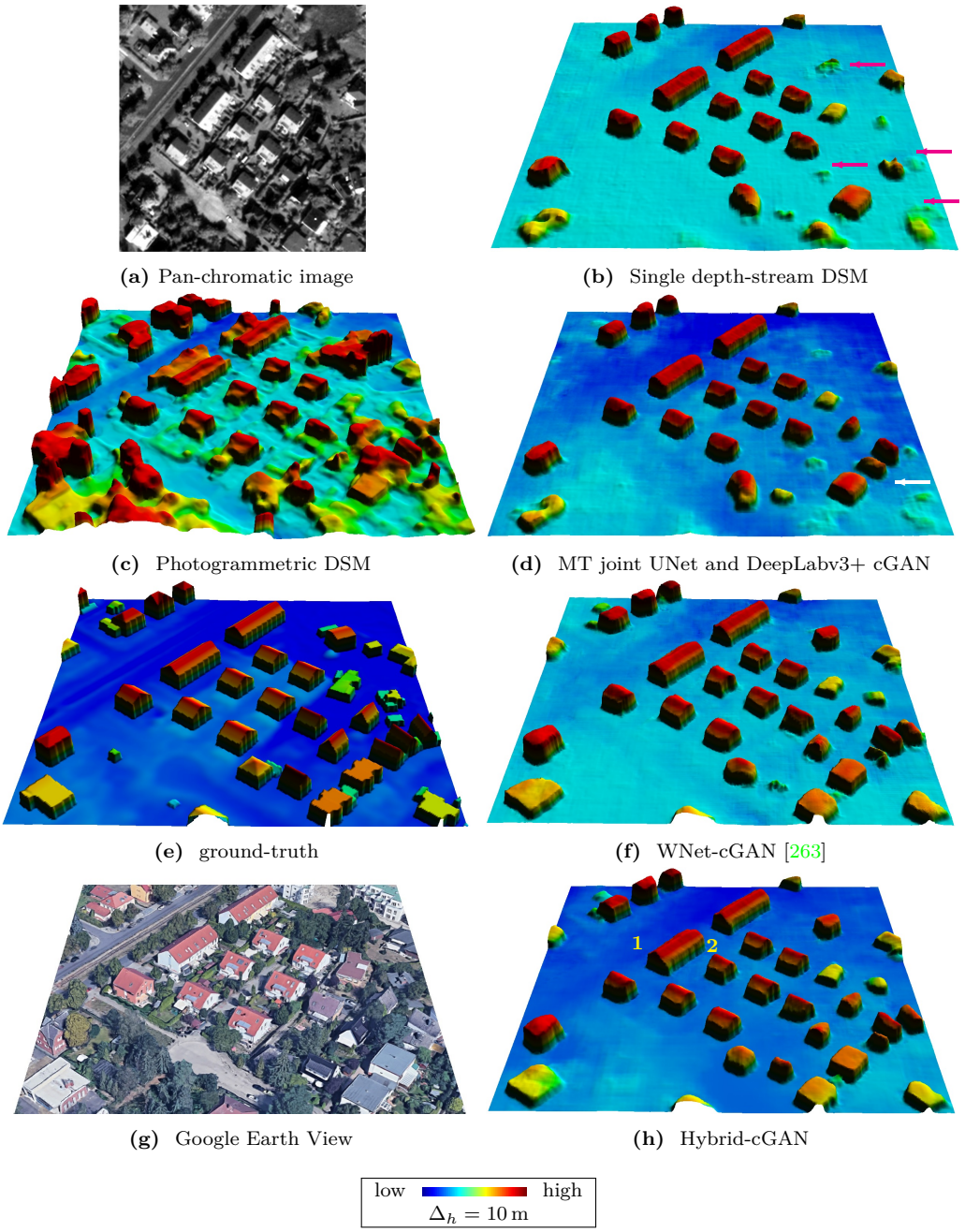
#### 5.4.3.1 Digital Surface Model Refinement

Selected test samples of DSMs generated by the single-stream cGAN model presented in Section 5.2, the two-stream WNet-cGAN model [263], and Hybrid-cGAN model are illustrated in Figures 5.26, 5.27 and 5.28. We also include the results of multi-task joint UNet and DeepLabv3+ cGAN model from Section 5.3 to make the comparison between all proposed network architectures. We point out that small buildings in all generated DSMs show more rectilinear borders and are not merged with adjacent trees, as present

in the input DSMs (*cf.* Figures 5.26c, 5.27c and 5.28c). The DSM generated from multi-task learning shows better results compared to the DSM from the single-stream cGAN model. More buildings are reconstructed (see the building highlighted with a white arrow in Figure 5.26d) with improved quality and form. This proves the statement made in Section 5.3 that the joint training of multiple tasks positively contributes to the solution of each problem. However, the integration of spectral information into the model clearly benefits the building reconstruction process even more. First of all, the number of reconstructed buildings is increased. For instance, the magenta arrows in Figure 5.26b highlight the areas in the DSM generated by the single-stream cGAN model, where the model is not able to reconstruct individual buildings, as opposed to the WNet-cGAN network [263] and Hybrid-cGAN network. Second, the roof ridge lines are more distinguishable and rectilinear. This statement can be also confirmed by exemplary investigating the profiles of the two buildings highlighted in Figure 5.26h. From Figure 5.29 we notice that the Hybrid-cGAN network is able to reconstruct much finer building shapes more similar to the ground-truth.

Moreover, the surfaces of roof planes are smoother or even more flat in many cases, affirming the influence of the normal vector loss. The profiles in Figure 5.29 also demonstrate the strength of all networks to separate the buildings from adjacent vegetation. From the demonstrated results we further conclude that most of the generated building shapes followed the correct pattern and feature improved roof forms. However, this statement is true for many residential but not large industrial buildings. How does the network behave in case of large and complicated building structures? The single depth-stream cGAN model only partially extracts such buildings (see Figures 5.27b and 5.28b). Providing the network with additional clues about building existence through its semantic information on the roof building mask via joint learning helps the model to better reconstruct the buildings' height representations depicted in Figures 5.27d and 5.28d. It can be seen that the inside of the construction as well as its border profile are more complete compared to the single depth-stream cGAN model. In the case of spectral information integration, it helps to improve the silhouette of the buildings as the detailed constructions on the rooftops, although at the late fusion (*cf.* Figures 5.27f and 5.28f), but still misses or has incomplete inside parts of structures. Moreover, because the input photogrammetric DSMs contain noise and many outliers, they propagate along the height image reconstruction and influence the results, as indicated by the dark blue areas in Figures 5.27b and 5.27f, and Figures 5.28b and 5.28f. However, the proposed Hybrid-cGAN network is able to eliminate those artifacts and also reconstruct the complete building structures on any single detail. Additionally, although the building rooftops seem to be entirely flat in the ground-truth data (*cf.* Figures 5.27e and 5.28e)—which is not the case in reality—, such cases do not confuse the model during the training phase and make it possible to preserve detailed 3D information from input photogrammetric DSM. These observations prove that the introduced Hybrid-cGAN architecture may successfully blend the spectral and height information together. The earlier combination of both modalities forces the network to accommodate the information even better.

In order to evaluate the generated elevation models quantitatively, the $\varepsilon_{\mathrm{MAE}}$, $\varepsilon_{\mathrm{RMSE}}$



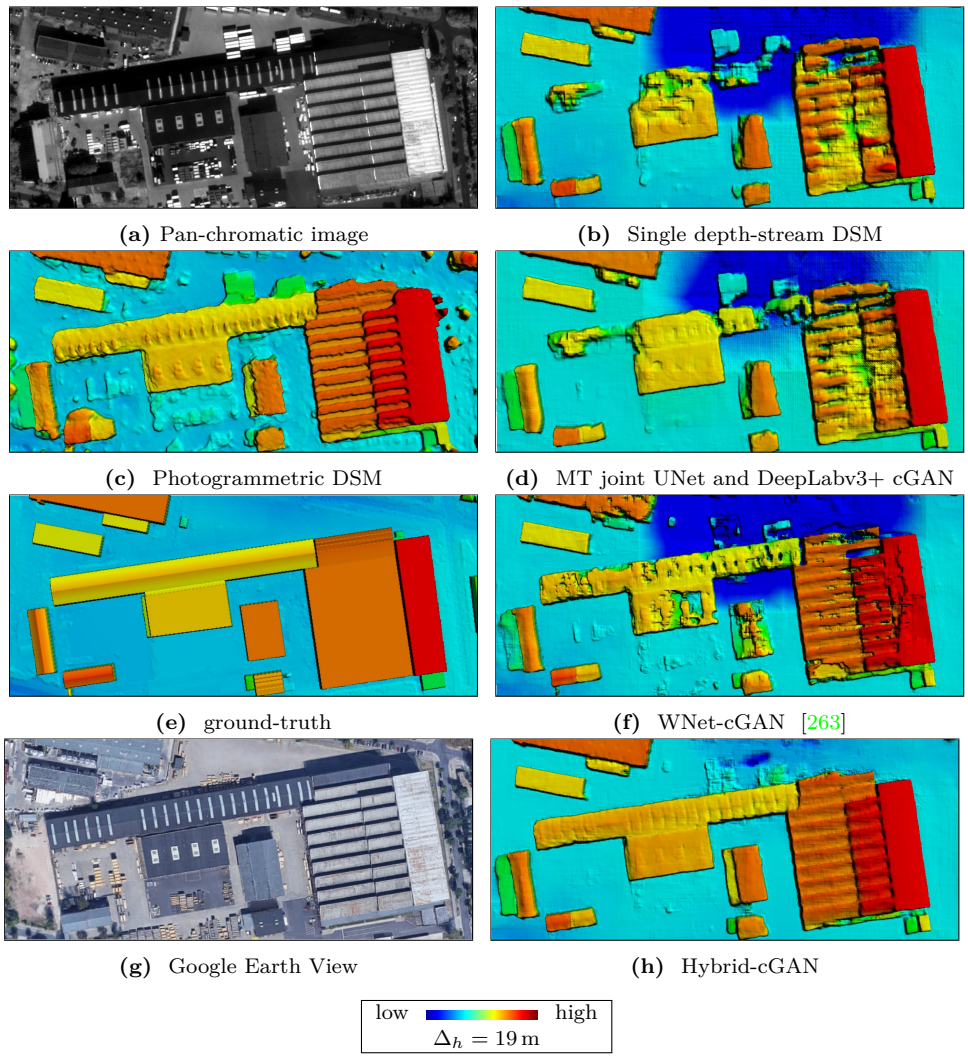

**(a)** Pan-chromatic image **(b)** Single depth-stream DSM

**(c)** Photogrammetric DSM **(d)** MT joint UNet and DeepLabv3+ cGAN

**(e)** ground-truth **(f)** WNet-cGAN [263]

**(g)** Google Earth View **(h)** Hybrid-cGAN

**Figure 5.26:** Visual analysis of DSMs, generated by (c) a standard photogrammetric method, (b) the single-stream cGAN model from Section 5.2, (d) the *multi-task (MT)* joint UNet and DeepLabv3+ cGAN model from Section 5.3, (f) the two-stream WNet-cGAN [263], and (h) the earlier fusion Hybrid-cGAN architecture. (a) and (c) depict the PAN image and the photogrammetric DSM, respectively, which are the input data to fusion networks. (e) demonstrates the ground-truth data used for learning process and evaluation procedure, and (g) shows the actual ground appearance from Google Earth Engine. The DSM images are color-coded for better visualization.



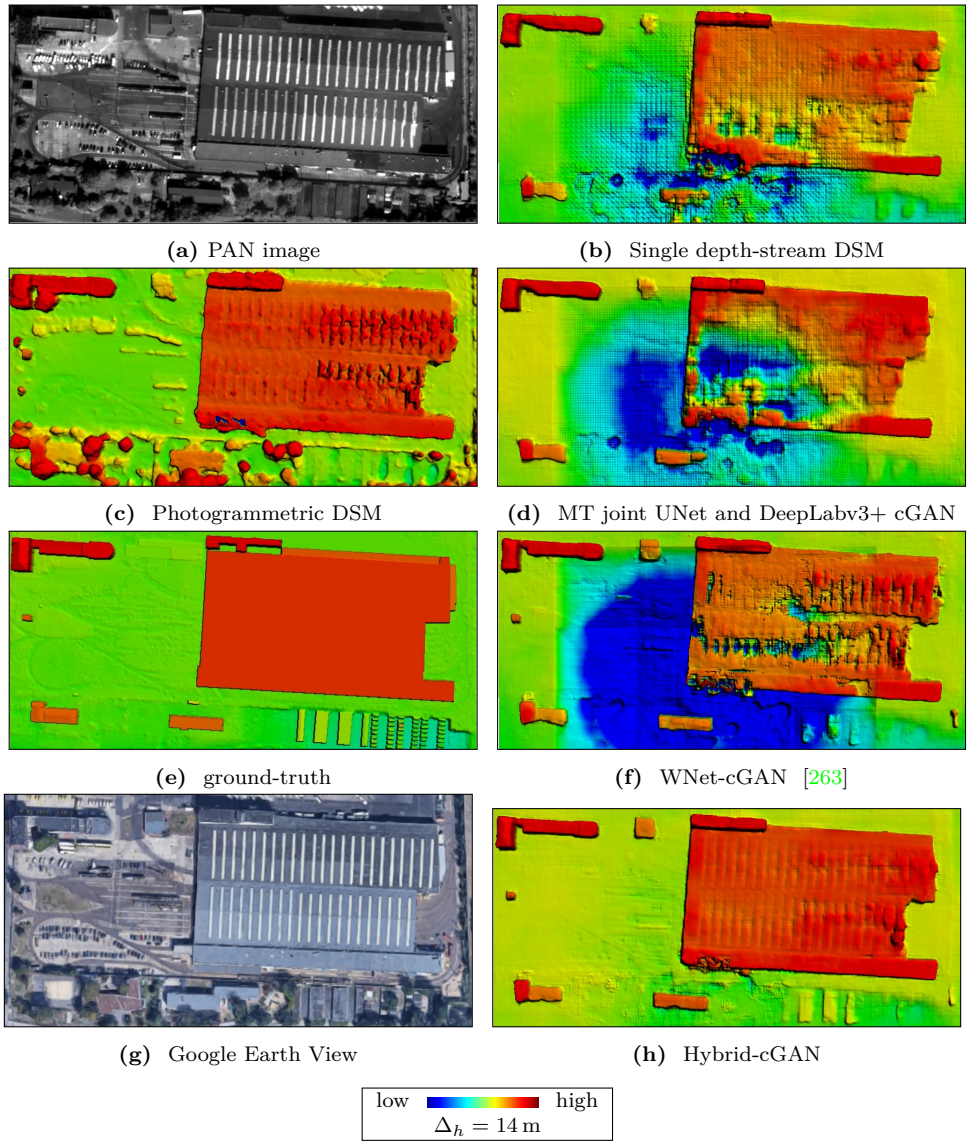

**(a)** Pan-chromatic image **(b)** Single depth-stream DSM

**(c)** Photogrammetric DSM **(d)** MT joint UNet and DeepLabv3+ cGAN

**(e)** ground-truth **(f)** WNet-cGAN [263]

**(g)** Google Earth View **(h)** Hybrid-cGAN

**Figure 5.27:** Visual analysis of DSMs, generated by (b) the single-stream cGAN model from Section 5.2, (d) MT joint UNet and DeepLabv3+ cGAN model from Section 5.3, (f) the two-stream WNet-cGAN [263], and (h) Hybrid-cGAN architecture. (g) shows the actual ground appearance from Google Earth Engine. The DSM images are color-shaded for better visualization.



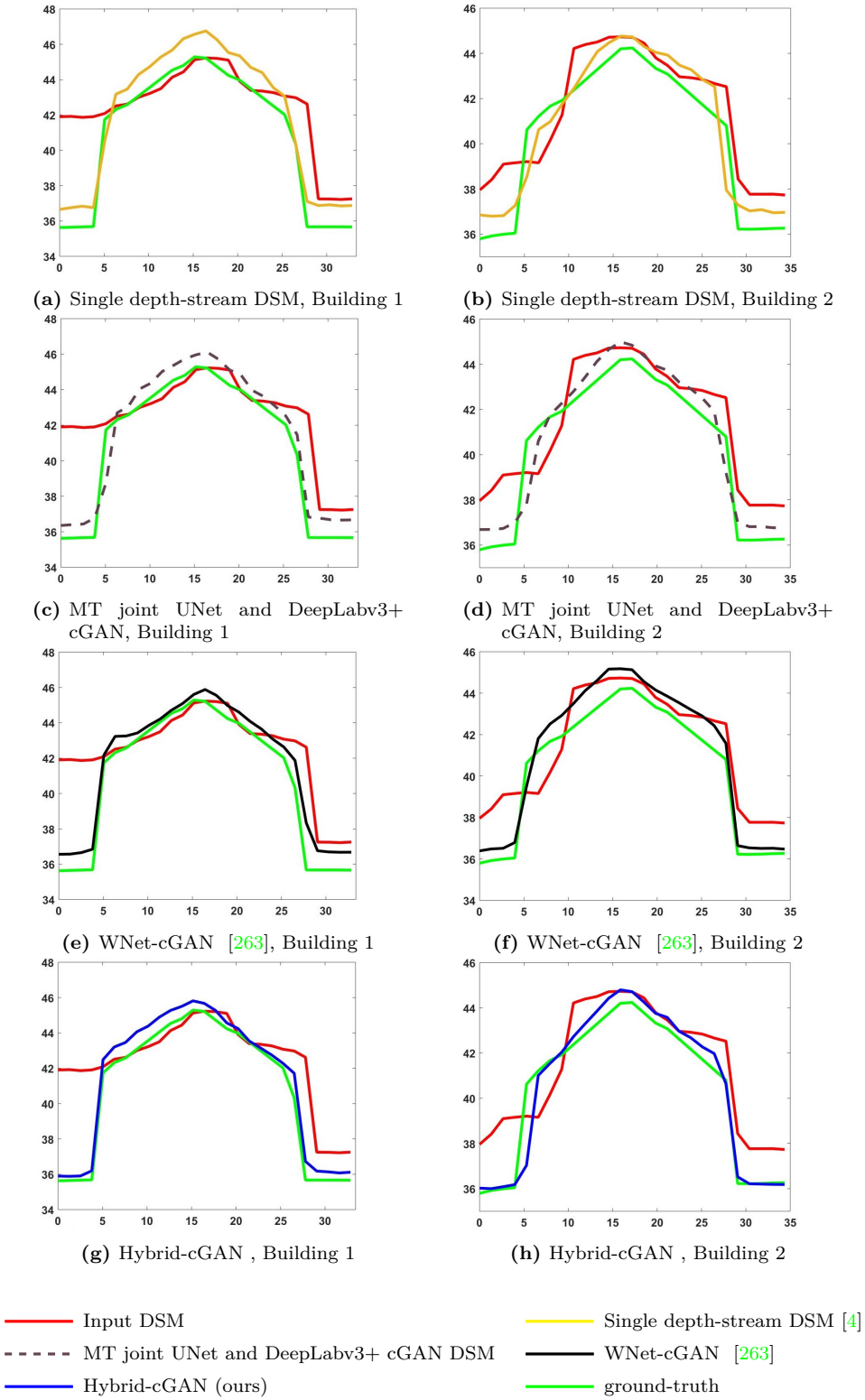

**(a)** PAN image **(b)** Single depth-stream DSM

**(c)** Photogrammetric DSM **(d)** MT joint UNet and DeepLabv3+ cGAN

**(e)** ground-truth **(f)** WNet-cGAN [263]

**(g)** Google Earth View **(h)** Hybrid-cGAN

**Figure 5.28:** Visual analysis of an DSM, generated by (b) the single-stream cGAN model from Section 5.2, (d) MT joint UNet and DeepLabv3+ cGAN model from Section 5.3, (f) the two-stream WNet-cGAN model [263], and (h) Hybrid-cGAN architecture. (g) shows the actual ground appearance from Google Earth Engine. The DSM images are color-coded for better visualization.




**(a)** Single depth-stream DSM, Building 1

**(b)** Single depth-stream DSM, Building 2

**(c)** MT joint UNet and DeepLabv3+ cGAN, Building 1

**(d)** MT joint UNet and DeepLabv3+ cGAN, Building 2

**(e)** WNet-cGAN [263], Building 1

**(f)** WNet-cGAN [263], Building 2

**(g)** Hybrid-cGAN , Building 1

**(h)** Hybrid-cGAN , Building 2

**Figure 5.29:** Visual investigation of profiles for two buildings selected from the DSMs generated by (a),(b) the single-stream cGAN model from Section 5.2 (c),(d) MT joint UNet and DeepLabv3+ cGAN model from Section 5.3, (e),(f) the two-stream WNet-cGAN model [263], and (g),(h) the proposed Hybrid-cGAN architecture. The first row shows profiles of "Building 1" and the second row depicts the profiles of "Building 2" (*cf.* Figure 5.26h).

and $\varepsilon_{\mathrm{NMAD}}$ error metrics previously defined by Equations (5.9), (5.10) and (5.11), respectively, are utilized. The same metrics that are used in Section 5.2 and Section 5.3 are used so as to be able to compare the generated LoD2-like DSMs from different models proposed in this chapter.

To exclude the influence of time acquisition difference between the photogrammetric DSMs and the CityGML data model—carrying the risk of absence or the appearance of new buildings— the evaluation regions are manually checked in this regard. The final evaluations are carried out on regions showing buildings in both the photogrammetric DSM as well as in the reference LoD2-DSM.

The obtained results for the three test area samples are presented in Tables 5.5, 5.6 and 5.7. In comparison to the other methods, the DSMs created by the single-stream model show inferior results in terms of $\varepsilon_{\mathrm{RMSE}}$ for all three areas. As has been mentioned before in Figure 5.26b, the model is not able to reconstruct some buildings, or only partially reconstructs them, which is highlighted in Figures 5.27b and 5.28b.

With the intensity information integrated into the learning process, the RMSE error $\varepsilon_{\mathrm{RMSE}}$ decreases and in the instance of the Hybrid-cGAN, achieves the lowest value smaller than for the input photogrammetric DSM. This observation provides evidence that the proposed model improves the noisy and inaccurate photogrammetric DSMs and changes them to highly detailed DSMs. Considering the other two metrics, Hybrid-cGAN also outperforms the competing models, except for the third test area. The single-stream DSM from Section 5.2 achieves the lowest NMAD error with $\varepsilon_{\mathrm{NMAD}} = 0.45$. This is most likely due to simplified roof representations on ground-truth data. More precisely, some buildings show flat roofs in the ground-truth without any complex structures, which is not the case in the reality, as the visual inspection of the input photogrammetric DSM and the DSMs generated by the models with spectral information integration demonstrates. It can also be even noticed that the DSM generated by Hybrid-cGAN model shows even better roof pattern reconstruction compared to the input photogrammetric DSM clearly influenced by intensity information from the PAN image, because the details in it are more clear. However, from statistics this refinement is not obvious. For example, because the NMAD metric is sensible to outliers, it shows higher values for the results with more detailed roofs demonstrated in Figures 5.27 and 5.28.

### 5.4.4 Practical Applications of Refined Digital Surface Models

Accurate information about the Earth topography can be used to understand hundreds of geoscience applications relevant to environmental monitoring, cartography, disaster management, and many other applications. Are the improved large-scale DSMs applicable for any of these? In this section, an example of remote sensing application obtained after applying improved DSMs (see Figure 5.30a) generated by the Hybrid-cGAN model is demonstrated.

The application related to DSM conversion into CityGML format is considered as a 3D city model representation. The result generated by applying the DLR software is depicted in Figure 5.30b. From the demonstrated result it can be seen that the method is able to generate a 3D city model directly from the provided DSM. As the

**Table 5.5:** Quantitative results for RMSE, NMAD, MAE metrics evaluated on 17 selected buildings existing on both the photogrammetric DSM and the ground-truth LoD2-DSM of the first area depicted in Figure 5.26.

| Method | Error | | |
|---|---|---|---|
| | RMSE (m) | NMAD (m) | MAE (m) |
| photogrammetric DSM | 1.66 | 1.01 | 1.23 |
| single-stream cGAN | 2.28 | 1.09 | 1.86 |
| multi-task cGAN | 2.18 | 0.78 | 1.64 |
| WNet-cGAN [263] | 1.63 | 0.72 | 1.22 |
| Hybrid-cGAN | **1.52** | **0.62** | **0.96** |

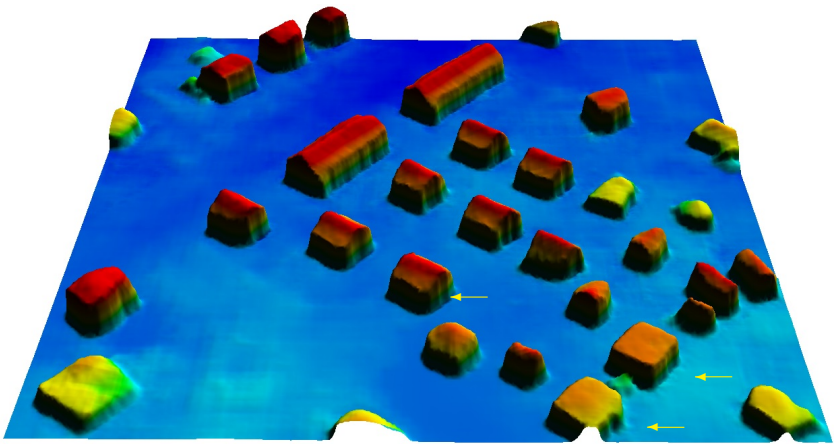
**Table 5.6:** Quantitative results for RMSE, NMAD, MAE metrics evaluated on 7 selected buildings existing on both the photogrammetric DSM and the ground-truth LoD2-DSM of the first area depicted in Figure 5.27.

| Method | Error | | |
|---|---|---|---|
| | RMSE (m) | NMAD (m) | MAE (m) |
| photogrammetric DSM | 2.72 | **1.09** | **1.57** |
| single-stream cGAN | 4.13 | 1.88 | 2.68 |
| multi-task cGAN | 3.97 | 1.79 | 2.55 |
| WNet-cGAN [263] | 3.89 | 2.03 | 2.64 |
| Hybrid-cGAN | **2.64** | 1.34 | 1.69 |

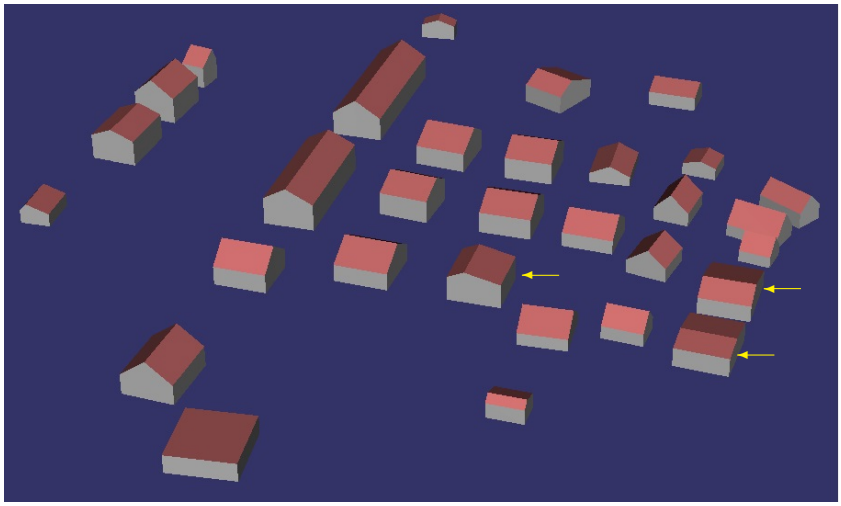
**Table 5.7:** Quantitative results for RMSE, NMAD, MAE metrics evaluated on 4 selected buildings existing on both the photogrammetric DSM and the ground-truth LoD2-DSM of the first area depicted in Figure 5.28.

| Method | Error | | |
|---|---|---|---|
| | RMSE (m) | NMAD (m) | MAE (m) |
| photogrammetric DSM | 2.19 | 0.62 | **1.12** |
| single-stream cGAN | 3.38 | **0.45** | 2.99 |
| multi-task cGAN | 4.84 | 1.88 | 3.41 |
| WNet-cGAN [263] | 3.16 | 1.15 | 2.02 |
| Hybrid-cGAN | **1.91** | 0.56 | 1.22 |

refined DSM does not contain any vegetation, no spectral images are needed for 3D city model extraction, which is normally the case when the protogrammetric DSM is used. However, we notice some inconsistency between the input DSM and the generated results highlighted by yellow arrows in Figure 5.30a and Figure 5.30b, respectively, although our input DSM is correct and coincides with the ground-truth (*cf.* Figures 5.26e and 5.26h). This refers to the DLR software for CityGML model generation failure itself.



**(a)** DSM refined by Hybrid-cGAN



**(b)** 3D city model form refined DSM

**Figure 5.30:** Illustration of 3D city model obtained from refined DSMs generated by applying the developed Hybrid-cGAN neural network (*cf.* Section 5.4.1). The example depicts the Berlin area.

## 5.5 Comparison and Discussion

The obtained results confirm that cGAN-based networks enables us to not only generate ground surface models containing improved and meaningful height information in the form of continuous values, it also refines the shapes of existing objects, e.g., buildings.

The proposed cGAN-based models are independent from the type and form of data. For example, in our first experiment, we tested the proposed cGAN and cLSGAN networks on two datasets consisting of photogrammetric DSMs as input data and either LiDAR-DSMs or LoD2-DSMs as ground-truth data. In both cases, the generated DSMs are close to the ground-truth in appearance. Depending on the desired output used for training, the model eliminated or created areas with a small amount of vegetation while generating the height images containing buildings (see Figures 5.6f, 5.6g, 5.6n and 5.6o). This is achieved through the neural network capability to model the desired level of output. Moreover, both the cGAN and the cLSGAN are resistant to generating fake building constructions on areas where there are no buildings present in the input photogrammetric DSM. The network attention is concentrating only on reconstruction existing buildings and simultaneous refinement of their structures, i.e., roofs' ridge lines, plane surfaces and outlines in general, despite the possible mismatch between the available input DSM generated from stereo satellite images and the given ground-truth data during the learning process. This mismatch happens in reality due to a data acquisition time difference.

Regarding the quality of the generated DSMs, we can say that the results from the investigated cLSGAN technique were very similar to those from cGAN in appearance. However, the detailed examination of specific building constructions and their profiles showed that cLSGAN outperformed cGAN model. The buildings reconstructed by the cGAN model miss some construction parts or feature artifacts on roof surfaces while the cLSGAN-based DSMs demonstrate more complete and realistic building shapes with smoother roof planes. The reason for this lies in the fact that the cGAN model uses the sigmoid cross entropy loss function for the discriminator [65] that leads to the vanishing gradient problem when updating the generator using the created samples that are on the correct side of the decision boundary but are still far from the real data [265]. As a result, the discriminator believes that the created images come from real samples and causes almost no error by updating the generator as the images are on the correct side of the decision boundary. Another explanation for the cLSGAN advantage compared to cGAN is the ability of the least squares loss function to penalize samples that lie far from the decision boundary, even if they were correctly classified, and move them towards the decision boundary [265]. As a result, this allows the generation of samples that are closer to the real data.

The generalization on an entirely new city unseen during the learning process showed that the network learned the general surface representation reconstruction. Specifically, independent from the topographic height of the input photogrammetric DSM, the proposed networks are able to generate the correct form of buildings with their actual heights as well as improve their geometries compared to the low-quality photogrammetric DSM.

This means that the network is capable of distinguishing the buildings from other objects by their specific features as well as their relative height.

In two follow-up experiments we investigated the influence of additional information inclusion into the neural network model for generating even better DSMs. In general, additional information can be presented in different forms. We have chosen a MT learning strategy to impact the correctness of building shape reconstruction by joint training of the depth regression task together with the pixel-wise classification problem, and a data fusion approach for integrating the intensity information into the DSM generation pipeline.

The obtained results prove that the incorporation of pixel-wise building roof type classification problems into the DSM generation task enables the improvement of each of them. This occurs due to the data interconnection, i.e., tasks relation. Through simultaneous learning of multiple problems, the model has to find a representation that is captured by all tasks. It makes the model more robust to the noise and prevents its overfitting as the model learns more generic representations. Moreover, it influences an integration of information through the sharing of relevant parameters of multiple tasks. Therefore, the proposed cGAN architecture has some common layers at the beginning, as earlier layers typically detect features like edges and corners, but separates the rest of the layers in order to avoid performance impairment due to the difference of physical meaning between the tasks.

The investigation of suitability of different network architectures for each task led to the decision of combining the U-form network for DSM generation together with the DeepLabv3+ architecture for pixel-wise classification within an end-to-end cGAN-based framework because they provide the best performance. The drawbacks of the examined ResNet34 and DeepLabv3+ architectures for depth image generation are the presence of incorrect height information due to failure propagation from the input photogrammetric DSM and oversmoothed building shape representations, respectively. The explanation of the UNet superior performance could be the presence of long skip connections compared to ResNet34, where only local residual connections were built. These long skip connections sent more detailed information from earlier layers to the top ones in the decoder, helping to compensate the incorrect reconstruction propagated from the failure region. The reason behind a smooth transition from roof to ground on DSMs produced by the DeepLabv3+ architecture is the last bilinear up-sampling layer with a factor of four in the decoder, which smoothed the results and was not able to learn such a complicated task such as elevation model generation. This was the main logic of why we did not consider the DeepLabv3+ model for the DSM generation task, although in terms of $\varepsilon_{\text{RMSE}}$ and $\varepsilon_{\text{MAE}}$ metrics it showed the superior performance. Overall, the joint learning of UNet and DeepLabv3+ also quantitatively proofed the positive influence of roof form knowledge on the depth image generation problem because evaluation metrics went down compared to other combinations. The newly introduced vector normal loss term into the final objective function also penalized irregularities on roof surfaces, leading to smoother roof shapes which are closer to the ground-truth.

For the pixel-wise classification task, the DeepLabv3+ network qualitatively and quan-

titatively demonstrated the best results. It was able to detect not only big urban constructions but also small residential houses. Moreover, the rate of correctly classified roof types is significantly higher compared to results generated by UNet- and ResNet34-based models. This observation confirms that the combination of short and long skip connections, together with the à trous convolution at multiple scales within the DeepLabv3+ architecture, positively influenced objects' extraction task.

With intensity knowledge integration into the DSM improvement procedure we were able to detect more buildings and also improve their geometry. These results were due to the increase of model confidence to make decisions, as additional observation provides supplementary information. Moreover, intensity images depict more accurate object appearance, e.g., boundaries and ridge lines, together with their texture information which is very useful for distinguishing the objects, for example from the ground or for reconstructing complicated patterns of their structures.

The obtained results demonstrated that data fusion models qualitatively and quantitatively outperforms even the multi-task model. This is reasonable, as demonstrated in the case of the data fusion model where the intensity information flows directly through the network adding profitable feature to feature extraction from photogrammetric DSMs. In the instance of multi-task learning, the pixel-wise classification masks implicitly influence the model via joint learning of multiple problems. Moreover, the study showed that the moment of merging intensity and height information also matters. For example, the WNet architecture [263] merge intermediate features from the PAN and the photogrammetric DSM images at the end, allowing the network to learn the height image generation task almost independently from each image. It still improves the reconstruction results (*cf.* $\varepsilon_{\text{RMSE}}$ metric Tables 5.5, 5.6 and 5.7) compared to other investigated models, because the network is still able to combine the relevant knowledge. However, it does not provide optimal representations in the instance of the presented industrial buildings. Earlier fusion in the Hybrid-cGAN architecture, on the other hand, blends intermediate features from the PAN image and the photogrammetric DSM in the middle of the proposed architecture, giving the network the possibility to benefit from complementary information. As a result, the refinement of building edge lines occurs together with its complete shape reconstruction, keeping meaningful roof patterns if applicable. Merging the images right at the beginning is not reasonable as the data have different physical meanings. Additionally, as demonstrated in Section 5.3, the normal vector loss added to the final objective function improved roof surfaces, which we considered in this experiment as well.

It should be mentioned that regarding the roof pattern reconstruction, all investigated networks are consistent (although in some cases the buildings were only partially reconstructed) even if the ground-truth used for comparison has a simplified form. This is also the reason why the $\varepsilon_{\text{RMSE}}$ metric shows differences over 1 m between the reconstructed elevation models and the ground-truth. Our results demonstrated that deep learning models visually produce fairly reasonable reconstructed DSMs. However, the used RMSE, NMAD, MAE metrics do not give good enough insight into the depth estimation quality, as they mainly consider the overall accuracy by reporting global statistics

of depth residuals. Moreover, the available ground-truth data with insufficient quality also influences the evaluation procedure.

To demonstrate the applicability and correctness of the generated DSMs for different remote sensing applications, an additional experiment was performed. A 3D city model was extracted. Regarding the 3D city model results, we can state that our DSM produced by the Hybrid-cGAN model provides full and valuable information for reliable building shape reconstruction. The presence of incorrect roof types on the generated 3D city model is related to the weakness of the city modeling methodology but not to our elevation model, as this can be confirmed comparing our elevation model with the ground-truth. From the obtained results we conclude that the cGAN models have a tremendous potential towards photogrammetric DSM refinement task and their use for different applications will be investigated in detail in the future.

## 5.6 Summary

Detailed and reliable *Digital Surface Models (DSMs)* are indispensable for applications related to building information extraction and reconstruction. Traditional stereo matching techniques are able to produce large-scale DSMs from multi-view satellite images, but in most cases the quality of building shapes in them is not reliable. In this chapter, three methodologies towards automatic DSMs enhancing from optical *Very High-Resolution (VHR)* satellite images were investigated. In the first method, the *conditional Generative Adversarial Network (cGAN)* with objective function based on least squares was developed. The approach demonstrated the superior performance of least squares objective function over negative log-likelihood objective function and the ability of deep neural network to deal with regression problems. More specifically, the cGAN is able to generate DSMs with improved building forms to a high level of accuracy from low-quality photogrammetric DSMs. The simultaneous training of deep neural network towards multiple tasks estimation, such as DSM enhancing and building roof type classification, further improved the photogrammetric DSMs. These findings were examined in the second method applying the multi-task learning *conditional Least Square Generative Adversarial Network (cLSGAN)* to the single-input photogrammetric DSM. Incorporation of scene information from different data, such as *pan-chromatic (PAN)* images and photogrammetric DSMs, added the complementary knowledge to the network, especially in cases of unclear object representations in one of the data sources. These observations were demonstrated by the third proposed approach based on cLSGAN deep neural network for data fusion. The possibility to reconstruct more correct building outlines and roof ridge lines integrating intensity and depth information, led to more realistic building representations in the resulting DSM. Additional experiments demonstrated the generalization capability of developed models on different urban scenes, unseen during the training, and the applicability of the resulting improved DSMs to other remote sensing applications.

Chapter 6

# Conclusion

## 6.1 Summary

Modern satellites collect digital data at high spatial resolution within specific wavelength ranges. They cover large areas of the Earth surface and are able to access even remote places rapidly. Moreover, the multi-view stereo image acquisition functionality of recent satellite platforms enables the generation of *Digital Surface Models (DSMs)*, valuable knowledge for analyzing the elevated objects. These characteristics make the satellite images ideal for information extraction about the Earth surface. One of the challenging, but critical tasks for urban environment analysis from satellite images is building information extraction and reconstruction, as structured components provide a fundamental knowledge for many remote sensing applications. This problem has been widely explored in the past, but still remains open due to the complexity of the urban regions, and the individual non-standardized shapes of building constructions. In this thesis, the possibilities of combining complementary knowledge from spectral, height and segmentation, or categorization (e.g., roof type) information obtained from satellite images is investigated, in order to achieve more comprehensive and reliable terrestrial scene understanding, particularly about buildings.

Two-dimensional building information extraction for complex urban areas was performed on the basis of *Fully Convolutional Network (FCN)* which is able to integrate different remote sensing data sources. The developed end-to-end Fused-FCN4s framework for pixel-wise classification enabled the fusion of the automatically learned relevant contextual features from height and spectral information, separately presented to a single architecture by *normalized Digital Surface Models (nDSMs)*, *red, green, and blue (RGB)*, and *pan-chromatic (PAN)* images, respectively. The final result was a unique binary building mask. Each of data modalities, e.g., spectral image and height image, has its advantages and limitations. Together, they provide vital, complementary information, i.e., elevation information of the objects in case of nDSMs and texture information and more detailed boundaries in case of spectral images, to compensate these limitations and increase the model confidence while making the decision. Experimental

results on two areas of Munich, Germany, and Istanbul, Turkey, unseen during training, have shown that the proposed Fused-FCN4s architecture can be successfully generalized over residential and industrial buildings of different urban scenes, without any difficulties due to their diversity. The applied deep neural network framework was able to merge most relevant information from satellite images with different modalities and extracted even small objects with miniscule details in the building outline. The boundaries of extracted building footprints are also more rectilinear due to spectral information integration. Therefore, the designed model does not require any additional post-processing steps. Some inaccuracies in the resulting building mask still can be observed, which can be a result of building complex construction or its coverage by trees. Moreover, the presence of noisy in nDSMs can additionally influence the results, as the distortion in building silhouette entails the failure in its footprint. Nevertheless, the proposed Fused-FCN4s architecture shows a great potential in providing a more robust solution for building footprint extraction problems from satellite images over a wide area.

High-quality DSMs with close to reality building shapes is a very valuable data source for many remote sensing applications. Although the existing photogrammetric methodologies are able to generate large-scale high-resolution DSMs from stereo imagery, they still exhibit irregularities and noise, due to problems like occlusion by trees, neighboring constructions, atmospheric effects, shadows, or matching errors. This leads to noisy or partly missing building structures and as a result, requires a refinement procedure. Many attempts were made to refine 3D objects on DSMs, like buildings, by fitting models from libraries or reconstructing 3D information from detected 2D building footprints using prior knowledge about the type of roofs. However, these algorithms were not robust to the huge variety of existing building forms.

Three *conditional Generative Adversarial Network (cGAN)* based methodologies were developed to achieve automatic DSMs reconstruction with accurate and close to reality building shapes. To accomplish the generation of surface models with buildings, which exhibit detailed shapes and roof forms, a high-quality DSM is necessary for formation of ground-truth data during the training process. Therefore, a methodology for transformation of *City Geography Markup Language (CityGML)* data to DSMs was proposed which delivers the necessary ground-truth for training. By applying the first proposed single-stream cGAN model, an improvement of photogrammetric DSMs has been achieved already. Building geometries in the generated DSMs mainly demonstrated enhanced roof ridge lines, giving buildings a more realistic appearance. Moreover, the detailed analysis reported that the model did not hallucinate new buildings, which at times is an issue using *Generative Adversarial Networks (GANs)*, but only reconstructed and improved the existing ones. Vegetation reconstruction was also not present and did not influence the generated results. Additionally, training on *Light Detection and Ranging (LiDAR)* DSM data was performed to demonstrate the generalization ability on different types of data. The trained system was tested on two unseen areas in the cities of Berlin and Munich and achieved positive results. The evaluation of the results showed the potential of the proposed methodology to generalize not only over diverse urban and industrial building shapes with complex constructions, but also in different cities without

major problems. However, some noise or unreconstructed parts of buildings were still present in the resulting DSMs. This can be explained as the consequence of the presence of very inaccurate parts in the input stereo DSM from which some buildings cannot be recognized, even with the human eye.

Aiming to improve the refinement, a deep learning approach based on a cGAN architecture was introduced consisting of two separated generators intended for multiple tasks. The proposed methodology was able to refine a large variety of building shapes in photogrammetric DSMs automatically and at the same time, produced improved roof classification maps. Although both generators were only connected at the beginning, they were able to contribute to each other for the reconstruction through joint learning and, as a result, produced more accurate results. The height information mainly helped to improve distinguishing buildings among various terrestrial targets, and roof classifications, in turn helped to refine building boundaries, making them more rectangular. Additionally, this is valuable information for roof forms' reconstruction as the network was able to determine this knowledge and use it for better learning. A *vector normal loss* term introduced to the objective function penalized irregularities on roof surfaces, leading to smoother roof shapes, which were closer to the ground-truth. The limitation of the proposed methodology is the separate streams in the generator architecture. Combining two generators $G_1$ and $G_2$ of the network by sharing even more hidden layers between all tasks while keeping only task-specific output layers, will overcome the problem of a large number of training parameters and should better influence the generation of both tasks, because complementary information will be jointly learned at the beginning.

To approach the problem differently and investigate the neural network potentials deeper, a further methodology for automatic building shape refinement from low-quality DSMs to *Level of Detail (LoD)* 2 from multiple space-borne remote sensing data was presented on the basis of cGANs. The Hybrid-cGAN network automatically combined the advantages of PAN imagery and photogrammetric DSM, while working with their individual drawbacks and from the obtained results, demonstrated the possibility of generating DSMs with completed residential and also industrial building structures. Moreover, the generated roof surfaces were smoother and more planar, giving evidence of the positive influence of the auxiliary normal vector loss function. A 3D visualization of the generated elevation models illustrates the realistic appearance of the buildings and their strong resemblance to the ground-truth. The comparison between three proposed architectures revealed that the Hybrid-cGAN produces the best results. However, the tremendous potential of the multi-task learning approach can be seen. It will therefore be worthwhile to investigate it in more detail in further work.

## 6.2 Future Work

Although the proposed methods demonstrate reliable results, further research could address additional improvements, for instance:

**Building footprint extraction**

- In light of the performed investigations, it was demonstrated that the earlier fusion of data from different modalities via deep neural networks benefits the resulting outputs more than the late fusion. Considering this discovery, the earlier fusion of spectral and elevation information towards building footprint extraction could be established. Additionally, this will make the network less complicated, because the number of parameters will be reduced.

- The presented method to binary building mask generation could be extended to utilizing original DSMs instead of nDSMs. The deep neural network will be able to learn the relevant building heights itself, which will exclude the influence of possible failures due to the pre-processing step (nDSM generation). For the same reason, the pan-sharpening step could be integrated into the network framework.

**DSM enhancement**

- The performance of the multi-task network can be further improved by avoiding a manual tuning of balancing hyper-parameters responsible for weighting the losses of individual tasks. A multi-task loss function that can learn to balance various regression and classification losses automatically could be established.

- The number of learning parameters in the multi-task network could be reduced by sharing more hidden layers between different tasks. This could also potentially influence the performance of individual tasks.

- The integration of both short skip connections as in the ResNet network and long skip connections as in the UNet network within one single architecture could further advance 3D reconstruction and roof classification map generation.

- The performance of multi-task as well as multi-model networks could be further improved by integration of color information in form of different spectral channels to the learning process instead of using only single intensity information from PAN image because it could introduce more rich features related to buildings.

- The geometry reconstruction and roof classification results could be further improved by combining multi-task and multi-modality networks into one framework to benefit from their individual advantages.

- The space-borne DSM enhancement methods could be extended to the refinement of low-resolution DSMs, like those generated from SPOT-6/-7 stereo imagery.

# Acronyms

| Notation | Description |
|---|---|
| ASPP | À trous Spatial Pyramid Pooling. |
| ASTER | Advanced Space-borne Thermal Emission and Reflection. |
| BN | Batch Normalization. |
| BPTT | Backpropagation Through Time. |
| BSP | Binary Space Partitioning. |
| CAVIS | Clouds, Aerosols, Water Vapor, Ice and Snow. |
| cGAN | conditional Generative Adversarial Network. |
| CIR | color infrared. |
| CIS | Channel-wise Inhibited Softmax. |
| CityGML | City Geography Markup Language. |
| cLSGAN | conditional Least Square Generative Adversarial Network. |
| CNN | Convolutional Neural Network. |
| CRF | Conditional Random Field. |
| CRFasRNN | Conditional Random Field as a Recurrent Neural Network. |
| CVF | Curvature Vector Flow. |
| DEM | Digital Elevation Model. |
| DSM | Digital Surface Model. |
| DTM | Digital Terrain Model. |
| ERS | European Remote Sensing Satellite. |

| Notation | Description |
|---|---|
| FC | Fully Connected. |
| FCN | Fully Convolutional Network. |
| FFT | Fast Fourier Transformation. |
| GAN | Generative Adversarial Network. |
| GCP | Ground Control Point. |
| GGVF | Generalized Gradient Vector Flow. |
| GIS | Geographic Information System. |
| GPS | Global Positioning System. |
| GPU | Graphics Processing Unit. |
| GSD | Ground Sampling Distance. |
| HF-FCN | Hierarchically Fused FCN. |
| IDW | Inverse Distance Weighting. |
| InSAR | Interferometric Synthetic Aperture Radar. |
| IoU | Intersection over Union. |
| IR | infrared. |
| JERS | Japan Earth Resources Satellite. |
| LAD | Least Absolute Deviation. |
| LAPGAN | Laplacian Pyramid GAN. |
| LiDAR | Light Detection and Ranging. |
| LoD | Level of Detail. |
| LReLU | Leaky Rectified Linear Unit. |
| LRN | Local Response Normalization. |
| LS | least squares. |
| LSM | Least Squares Matching. |
| MAE | Mean Absolute Error. |
| MLP | Multilayer Perceptron. |
| MPP | Marked Point Process. |
| MQ | multi-quadric. |
| MRF | Markov Random Field. |
| MT | multi-task. |

| **Notation** | **Description** |
|---|---|
| NAIP | National Agriculture Imagery Program. |
| nDSM | normalized Digital Surface Model. |
| NDVI | Normalized Difference Vegetation Index. |
| NIR | near-infrared. |
| NMAD | Normalized Median Absolute Deviation. |
| OSM | OpenStreetMap. |
| PAN | pan-chromatic. |
| PCA | Principal Component Analysis. |
| PCA1 | Principal Component Analysis 1. |
| PolSAR | Polarimetric Synthetic Aperture Radar. |
| R-CNN | Region-based CNN. |
| ReLU | Rectified Linear Unit. |
| ResNet | residual network. |
| RGB | red, green, and blue. |
| RMSE | Root Mean Square Error. |
| RNN | Recursive Neural Network. |
| RPC | Rational Polynomial Coefficient. |
| RPN | Region Proposal Network. |
| SAR | Synthetic Aperture Radar. |
| SGD | Stochastic Gradient Descent. |
| SGM | Semi-Global Matching. |
| SIFT | Scale Invariant Feature Transform. |
| SRGAN | Super-Resolution GAN. |
| SRTM | Shuttle Radar Topographic Mission. |
| SVM | Support Vector Machine. |
| SWIR | short wave infrared. |
| VHR | Very High-Resolution. |

# List of Figures

# List of Tables

# Bibliography


[1] F. Xu, N. Woodhouse, Z. Xu, D. Marr, X. Yang, and Y. Wang, "Blunder elimination techniques in adaptive automatic terrain extraction", *ISPRS J*, vol. 29, no. 3, p. 21, 2008.

[2] B. Sirmacek, P. d'Angelo, T. Krauss, and P. Reinartz, "Enhancing urban digital elevation models using automated computer vision techniques", in *ISPRS Commission VII Symposium*, 2010.

[3] K. Bittner, F. Adam, S. Cui, M. Körner, and P. Reinartz, "Building footprint extraction from vhr remote sensing images combined with normalized dsms using fused fully convolutional networks", *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 8, pp. 2615–2629, 2018.

[4] K. Bittner, P. d'Angelo, M. Körner, and P. Reinartz, "Dsm-to-lod2: Spaceborne stereo digital surface model refinement", *Remote Sensing*, vol. 10, no. 12, p. 1926, 2018. DOI: 10.3390/rs10121926.

[5] K. Bittner, M. Körner, F. Fraundorfer, and P. Reinartz, "Multi-task cgan for simultaneous spaceborne dsm refinement and roof-type classification", *Remote Sensing*, vol. 11, no. 11, p. 1262, 2019.

[6] K. Bittner, P. Reinartz, and M. Korner, "Late or earlier information fusion from depth and spectral data? large-scale digital surface model refinement by hybrid-cgan", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[7] T. Krauß, B. Sirmacek, H. Arefi, and P. Reinartz, "Fusing stereo and multispectral data from worldview-2 for urban modeling", in *Proc. of SPIE Vol*, vol. 8390, 2012, pp. 83901X–1.

[8] S. Gopi *et al.*, *Advanced surveying: total station, GIS and remote sensing*. Pearson Education India, 2007.

[9] R. P. Gupta, *Remote sensing geology*. Springer, 2017.

[10] P. L. Basgall, F. A. Kruse, and R. C. Olsen, "Comparison of lidar and stereo photogrammetric point clouds for change detection", in *Laser Radar Technology and Applications XIX; and Atmospheric Propagation XI*, International Society for Optics and Photonics, vol. 9080, 2014, 90800R.

[11] T. Toutin and L. Gray, "State-of-the-art of elevation extraction from satellite sar data", *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 55, no. 1, pp. 13–33, 2000.

[12] M Crosetto, "Calibration and validation of sar interferometry for dem generation", *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 57, no. 3, pp. 213–227, 2002.

[13] J. H. Yu, X. Li, L. Ge, and H.-C. Chang, "Radargrammetry and interferometry sar for dem generation", in *15th Australasian Remote Sensing & Photogrammetry Conf. Alice Springs, Australia*, Citeseer, 2010, pp. 1212–1223.

[14] U Soergel, U Thoennessen, and U Stilla, "Visibility analysis of man-made objects in sar images", in *2003 2nd GRSS/ISPRS Joint Workshop on Remote Sensing and Data Fusion over Urban Areas*, IEEE, 2003, pp. 120–124.

[15] J Everaerts, "Pegasus—bridging the gap between airborne and spaceborne remote sensing", *New Strategies For European Remote Sensing*, pp. 395–401, 2005.

[16] L. Zhang, *Automatic digital surface model (DSM) generation from linear array images*. ETH Zurich, 2005.

[17] P. d'Angelo and P. Reinartz, "Semiglobal matching results on the isprs stereo matching benchmark", *ISPRS Hannover Workshop*, vol. 38, no. 4/W19, pp. 79–84, 2011.

[18] W. Turner, S. Spector, N. Gardiner, M. Fladeland, E. Sterling, and M. Steininger, "Remote sensing for biodiversity science and conservation", *Trends in ecology & evolution*, vol. 18, no. 6, pp. 306–314, 2003.

[19] S. Goetz, *Crisis in earth observation*, 2007.

[20] S. R. Loarie, L. N. Joppa, and S. L. Pimm, "Satellites miss environmental priorities", *Trends in Ecology & Evolution*, vol. 22, no. 12, pp. 630–632, 2007.

[21] *Satellite imaging corporation*, https://www.satimagingcorp.com/.

[22] *European space imaging*, https://www.euspaceimaging.com/.

[23] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information", *IEEE Transactions on pattern analysis and machine intelligence*, vol. 30, no. 2, pp. 328–341, 2008.

[24] P. d'Angelo and G. Kuschk, "Dense multi-view stereo from satellite imagery", in *2012 IEEE International Geoscience and Remote Sensing Symposium*, IEEE, 2012, pp. 6944–6947.

[25] Z. M. Moratto, M. J. Broxton, R. A. Beyer, M. Lundy, and K. Husmann, "Ames stereo pipeline, nasa's open source automated stereogrammetry software", in *Lunar and Planetary Science Conference*, vol. 41, 2010, p. 2364.

[26] O. C. Ozcanli, Y. Dong, J. L. Mundy, H. Webb, R. Hammoud, and V. Tom, "A comparison of stereo and multiview 3-d reconstruction using cross-sensor satellite imagery", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 17–25.

[27] K Gong and D Fritsch, "A detailed study about digital surface model generation using high resolution satellite stereo imagery.", *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, vol. 3, no. 1, 2016.

[28] M. Lehner, P. Reinartz, *et al.*, "Stereo evaluation of cartosat-1 data summary of dlr results during cartosat-1 scientific assessment program", *International Society for Photogrammetry and Remote Sensing (ISPRS)*, pp. 1207–1212, 2008.

[29] R. Müller, T. Krauß, M. Lehner, and P. Reinartz, "Automatic production of a european orthoimage coverage within the gmes land fast track service using spot 4/5 and irs-p6 liss iii data", in *ISPRS conference proceedings*, vol. 46, 2007, p. 6.

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks", in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[31] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[32] R. Girshick, "Fast r-cnn", in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[33] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn", in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[34] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation", in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 801–818.

[35] M. Lin, Q. Chen, and S. Yan, "Network in network", *arXiv preprint arXiv:1312.4400*, 2013.

[36] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition", *arXiv preprint arXiv:1409.1556*, 2014.

[37] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[38] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[39] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks", *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 10, pp. 6232–6251, 2016.

[40] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors", *nature*, vol. 323, no. 6088, p. 533, 1986.

[41] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Convolutional neural networks for large-scale remote-sensing image classification", *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 2, pp. 645–657, 2017.

[42] N. Qian, "On the momentum term in gradient descent learning algorithms", *Neural networks*, vol. 12, no. 1, pp. 145–151, 1999.

[43] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization", *Journal of Machine Learning Research*, vol. 12, no. Jul, pp. 2121–2159, 2011.

[44] Y. Nesterov, "A method for unconstrained convex minimization problem with the rate of convergence o $(1/\hat{k2})$", in *Doklady AN USSR*, vol. 269, 1983, pp. 543–547.

[45] M. D. Zeiler, "Adadelta: An adaptive learning rate method", *arXiv preprint arXiv:1212.5701*, 2012.

[46] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization", *arXiv preprint arXiv:1412.6980*, 2014.

[47] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting", *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[48] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3156–3164.

[49] A. Graves and J. Schmidhuber, "Offline handwriting recognition with multidimensional recurrent neural networks", in *Advances in neural information processing systems*, 2009, pp. 545–552.

[50] M. S. Al-Radhi, T. G. Csapó, and G. Németh, "Deep recurrent neural networks in speech synthesis using a continuous vocoder", in *International Conference on Speech and Computer*, Springer, 2017, pp. 282–291.

[51] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks", in *2013 IEEE international conference on acoustics, speech and signal processing*, IEEE, 2013, pp. 6645–6649.

[52] J. Zhang and K. Man, "Time series prediction using rnn in multi-dimension embedding phase space", in *SMC'98 Conference Proceedings. 1998 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No. 98CH36218)*, IEEE, vol. 2, 1998, pp. 1868–1873.

[53] S. Ebrahimi Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal, "Recurrent neural networks for emotion recognition in video", in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ACM, 2015, pp. 467–474.

[54] S. Sigtia, E. Benetos, S. Cherla, T. Weyde, A Garcez, and S. Dixon, "Rnn-based music language models for improving automatic music transcription", 2014.

[55] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks", in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[56] M. Auli, M. Galley, C. Quirk, and G. Zweig, "Joint language and translation modeling with recurrent neural networks", 2013.

[57] I. V. Serban, T. Klinger, G. Tesauro, K. Talamadupula, B. Zhou, Y. Bengio, and A. Courville, "Multiresolution recurrent neural networks: An application to dialogue response generation", in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[58] S. Yildirim, "Design of adaptive robot control system using recurrent neural network", *Journal of Intelligent and Robotic Systems*, vol. 44, no. 3, pp. 247–261, 2005.

[59] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

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

[61] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[62] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition", *International journal of computer vision*, vol. 104, no. 2, pp. 154–171, 2013.

[63] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation", *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 1, pp. 142–158, 2015.

[64] 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*, 2015, pp. 91–99.

[65] 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*, 2014, pp. 2672–2680.

[66] M. Mirza and S. Osindero, "Conditional generative adversarial nets", *arXiv preprint arXiv:1411.1784*, 2014.

[67] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks", *arXiv preprint arXiv:1611.07004*, 2016.

[68] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8798–8807.

[69] R. Villegas, J. Yang, S. Hong, X. Lin, and H. Lee, "Decomposing motion and content for natural video sequence prediction", *arXiv preprint arXiv:1706.08033*, 2017.

[70] P. Burt and E. Adelson, "The laplacian pyramid as a compact image code", *IEEE Transactions on communications*, vol. 31, no. 4, pp. 532–540, 1983.

[71] E. L. Denton, S. Chintala, R. Fergus, *et al.*, "Deep generative image models using a laplacian pyramid of adversarial networks", in *Advances in neural information processing systems*, 2015, pp. 1486–1494.

[72] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network.", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, 2017, p. 4.

[73] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from rgb-d images for object detection and segmentation", in *European conference on computer vision*, Springer, 2014, pp. 345–360.

[74] M. Schwarz, H. Schulz, and S. Behnke, "Rgb-d object recognition and pose estimation based on pre-trained convolutional neural network features", in *2015 IEEE international conference on robotics and automation (ICRA)*, IEEE, 2015, pp. 1329–1335.

[75] K. J. Åström and B. Wittenmark, *Computer-controlled systems: theory and design*. Courier Corporation, 2013.

[76] Y. Li, J. Zhang, Y. Cheng, K. Huang, and T. Tan, "Semantics-guided multi-level RGB-D feature fusion for indoor semantic segmentation", in *2017 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2017, pp. 1262–1266.

[77] J. Kim, J. Koh, Y. Kim, J. Choi, Y. Hwang, and J. W. Choi, "Robust deep multi-modal learning based on gated information fusion network", *arXiv preprint arXiv:1807.06233*, 2018.

[78] R. Socher, B. Huval, B. Bath, C. D. Manning, and A. Y. Ng, "Convolutional-recursive deep learning for 3d object classification", in *Advances in neural information processing systems*, 2012, pp. 656–664.

[79] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard, "Multimodal deep learning for robust rgb-d object recognition", in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2015, pp. 681–687.

[80] A. Wang, J. Lu, G. Wang, J. Cai, and T.-J. Cham, "Multi-modal unsupervised feature learning for rgb-d scene labeling", in *European Conference on Computer Vision*, Springer, 2014, pp. 453–467.

[81] H Jhuang, H Garrote, E Poggio, T Serre, and T Hmdb, "A large video database for human motion recognition", in *Proc. of IEEE International Conference on Computer Vision*, vol. 4, 2011, p. 6.

[82] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition", *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2013.

[83] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos", in *Advances in neural information processing systems*, 2014, pp. 568–576.

[84] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: A large video database for human motion recognition", in *2011 International Conference on Computer Vision*, IEEE, 2011, pp. 2556–2563.

[85] D. McLaughlin, "An integrated approach to hydrologic data assimilation: Interpolation, smoothing, and filtering", *Advances in Water Resources*, vol. 25, no. 8-12, pp. 1275–1286, 2002.

[86] Y Liberman, R Samuels, P Alpert, and H Messer, "New algorithm for integration between wireless microwave sensor network and radar for improved rainfall measurement and mapping", *Atmospheric Measurement Techniques*, vol. 7, no. 10, pp. 3549–3563, 2014.

[87] H. Seyyedi, *Comparing satellite derived rainfall with ground based radar for Northwestern Europe*. University of Twente Faculty of Geo-Information and Earth Observation (ITC), 2010.

[88] S. T. Shivappa, M. M. Trivedi, and B. D. Rao, "Audiovisual information fusion in human–computer interfaces and intelligent environments: A survey", *Proceedings of the IEEE*, vol. 98, no. 10, pp. 1692–1715, 2010.

[89] M. Turk, "Multimodal interaction: A review", *Pattern Recognition Letters*, vol. 36, pp. 189–195, 2014.

[90] M. Liggins II, D. Hall, and J. Llinas, *Handbook of multisensor data fusion: theory and practice*. CRC press, 2017.

[91] H. C. Lai, R. Yang, and G. W. Ng, "Enhanced self-organizing map for passive sonar tracking to improve situation awareness", in *Information Fusion, 2007 10th International Conference on*, IEEE, 2007, pp. 1–7.

[92] L. Matthies, Y. Xiong, R Hogg, D. Zhu, A Rankin, B. Kennedy, M. Hebert, R Maclachlan, C. Won, T. Frost, *et al.*, "A portable, autonomous, urban reconnaissance robot", *Robotics and Autonomous Systems*, vol. 40, no. 2, pp. 163–172, 2002.

[93] D. L. Hall and J. Llinas, "An introduction to multisensor data fusion", *Proceedings of the IEEE*, vol. 85, no. 1, pp. 6–23, 1997.

[94] G. L. Foresti and C. S. Regazzoni, "Multisensor data fusion for autonomous vehicle navigation in risky environments", *IEEE Transactions on Vehicular Technology*, vol. 51, no. 5, pp. 1165–1185, 2002.

[95] R. Caruana, "Multitask learning", *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[96] R. K. Ando and T. Zhang, "A framework for learning predictive structures from multiple tasks and unlabeled data", *Journal of Machine Learning Research*, vol. 6, no. Nov, pp. 1817–1853, 2005.

[97] Y. Zhang and Q. Yang, "An overview of multi-task learning", *National Science Review*, vol. 5, no. 1, pp. 30–43, 2017.

[98] S. Ruder, "An overview of multi-task learning in deep neural networks", *arXiv preprint arXiv:1706.05098*, 2017.

[99] S. J. Pan and Q. Yang, "A survey on transfer learning", *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[100] Q. Xia, P. Jiang, F. Sun, Y. Zhang, X. Wang, and Z. Sui, "Modeling consumer buying decision for recommendation based on multi-task deep learning", in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, ACM, 2018, pp. 1703–1706.

[101] G. Obozinski, B. Taskar, and M. I. Jordan, "Joint covariate selection and joint subspace selection for multiple classification problems", *Statistics and Computing*, vol. 20, no. 2, pp. 231–252, 2010.

[102] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning", in *Proceedings of the 25th international conference on Machine learning*, ACM, 2008, pp. 160–167.

[103] L. Deng, G. Hinton, and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: An overview", in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, 2013, pp. 8599–8603.

[104] A. Torralba, K. P. Murphy, and W. T. Freeman, "Sharing visual features for multiclass and multiview object detection", *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 5, pp. 854–869, 2007.

[105] Z. Kang, K. Grauman, and F. Sha, "Learning with whom to share in multi-task feature learning.", in *ICML*, vol. 2, 2011, p. 4.

[106] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition", in *International conference on machine learning*, 2014, pp. 647–655.

[107] J. Yim, H. Jung, B. Yoo, C. Choi, D. Park, and J. Kim, "Rotating your face using multi-task deep neural network", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 676–684.

[108] G. Gkioxari, B. Hariharan, R. Girshick, and J. Malik, "R-cnns for pose estimation and action detection", *arXiv preprint arXiv:1406.5212*, 2014.

[109] Y. Liao, S. Kodagoda, Y. Wang, L. Shi, and Y. Liu, "Understand scene categories by objects: A semantic regularized scene classifier using convolutional neural networks", in *2016 IEEE international conference on robotics and automation (ICRA)*, IEEE, 2016, pp. 2318–2325.

[110] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks", *arXiv preprint arXiv:1312.6229*, 2013.

[111] I. Misra, A. Shrivastava, A. Gupta, and M. Hebert, "Cross-stitch networks for multi-task learning", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3994–4003.

[112] M. Teichmann, M. Weber, M. Zoellner, R. Cipolla, and R. Urtasun, "Multinet: Real-time joint semantic reasoning for autonomous driving", in *2018 IEEE Intelligent Vehicles Symposium (IV)*, IEEE, 2018, pp. 1013–1020.

[113] J. Uhrig, M. Cordts, U. Franke, and T. Brox, "Pixel-level encoding and depth layering for instance-level semantic labeling", in *German Conference on Pattern Recognition*, Springer, 2016, pp. 14–25.

[114] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7482–7491.

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

[116] I. Kokkinos, "Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6129–6138.

[117] L. Liebel and M. Körner, "Auxiliary tasks in multi-task learning", *arXiv preprint arXiv:1805.06334*, 2018.

[118] X. Lin, D. Sánchez-Escobedo, J. R. Casas, and M. Pardàs, "Depth estimation and semantic segmentation from a single rgb image using a hybrid convolutional neural network", *Sensors*, vol. 19, no. 8, p. 1795, 2019.

[119] S. Saito and Y. Aoki, "Building and road detection from large aerial imagery", in *Image Processing: Machine Vision Applications VIII*, International Society for Optics and Photonics, vol. 9405, 2015, 94050K.

[120] A. Suliman, Y. Zhang, and R. Al-Tahir, "Registration-based mapping of above-ground disparities (rmad) for building detection in off-nadir vhr stereo satellite imagery", *Photogrammetric Engineering & Remote Sensing*, vol. 82, no. 7, pp. 535–546, 2016.

[121] D. Konstantinidis, T. Stathaki, V. Argyriou, and N. Grammalidis, "Building detection using enhanced hog–lbp features and region refinement processes", *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 3, pp. 888–905, 2016.

[122] D. H. Douglas and T. K. Peucker, "Algorithms for the reduction of the number of points required to represent a digitized line or its caricature", *Cartographica: The International Journal for Geographic Information and Geovisualization*, vol. 10, no. 2, pp. 112–122, 1973.

[123] A. Huertas and R. Nevatia, "Detecting buildings in aerial images", *Computer Vision, Graphics, and Image Processing*, vol. 41, no. 2, pp. 131–152, 1988.

[124] R. B. Irvin and D. M. McKeown, "Methods for exploiting the relationship between buildings and their shadows in aerial imagery", *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 19, no. 6, pp. 1564–1575, 1989.

[125] R. Guercke and M. Sester, "Building footprint simplification based on hough transform and least squares adjustment", in *Proceedings of the 14th workshop of the ICA commission on generalisation and multiple representation, Paris*, 2011.

[126] R. Mohan and R. Nevatia, "Using perceptual organization to extract 3d structures", *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 11, pp. 1121–1139, 1989.

[127] S. Krishnamachari and R. Chellappa, "Delineating buildings by grouping lines with mrfs", *IEEE Transactions on image processing*, vol. 5, no. 1, pp. 164–168, 1996.

[128] W. Liu and V. Prinet, "Building detection from high-resolution satellite image using probability model", in *Proceedings. 2005 IEEE International Geoscience and Remote Sensing Symposium, 2005. IGARSS'05.*, Citeseer, vol. 6, 2005, pp. 3888–3891.

[129] P. Saeedi and H. Zwick, "Automatic building detection in aerial and satellite images", in *2008 10th International Conference on Control, Automation, Robotics and Vision*, IEEE, 2008, pp. 623–629.

[130] J. Wang, X. Yang, X. Qin, X. Ye, and Q. Qin, "An efficient approach for automatic rectangular building extraction from very high resolution optical satellite imagery", *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 3, pp. 487–491, 2014.

[131] T. Kim and J.-P. Muller, "Development of a graph-based approach for building detection", *Image and Vision Computing*, vol. 17, no. 1, pp. 3–14, 1999.

[132] K. Segl and H. Kaufmann, "Detection of small objects from high-resolution panchromatic satellite imagery based on supervised image segmentation", *IEEE Transactions on geoscience and remote sensing*, vol. 39, no. 9, pp. 2080–2083, 2001.

[133] M. Molinier, J. Laaksonen, and T. Hame, "Detecting man-made structures and changes in satellite imagery with a content-based information retrieval system built on self-organizing maps", *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 4, pp. 861–874, 2007.

[134] B. Sirmacek and C. Unsalan, "Urban-area and building detection using sift keypoints and graph theory", *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 4, pp. 1156–1167, 2009.

[135] A. O. Ok, "Automated detection of buildings from single vhr multispectral images using shadow information and graph cuts", *ISPRS journal of photogrammetry and remote sensing*, vol. 86, pp. 21–40, 2013.

[136] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models", *International journal of computer vision*, vol. 1, no. 4, pp. 321–331, 1988.

[137] L. D. Cohen, "On active contour models and balloons", *CVGIP: Image understanding*, vol. 53, no. 2, pp. 211–218, 1991.

[138] C. Xu and J. L. Prince, "Generalized gradient vector flow external forces for active contours", *Signal processing*, vol. 71, no. 2, pp. 131–139, 1998.

[139] T. F. Chan and L. A. Vese, "Active contours without edges", *IEEE Transactions on image processing*, vol. 10, no. 2, pp. 266–277, 2001.

[140] D. Gil and P. Radeva, "Curvature vector flow to assure convergent deformable models for shape modelling", in *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, Springer, 2003, pp. 357–372.

[141] S. Ahmady, H. Ebadi, M. V. Zouj, and H. A. Moghaddam, "Automatic building extraction from high resolution aerial images using active contour model", *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 37, pp. 453–456, 2008.

[142] D. Mumford and J. Shah, "Optimal approximations by piecewise smooth functions and associated variational problems", *Communications on pure and applied mathematics*, vol. 42, no. 5, pp. 577–685, 1989.

[143] A. J. Fazan, A. Porfírio, A. J. F. Dal Poz, and D. Poz, "Building roof contours extraction from aerial imagery based on snakes and dynamic programming", 2010.

[144] D. Chen, S. Shang, and C. Wu, "Shadow-based building detection and segmentation in high-resolution remote sensing image.", *journal of multimedia*, vol. 9, no. 1, pp. 181–188, 2014.

[145] X. Huang and L. Zhang, "Morphological building/shadow index for building extraction from high-resolution imagery over urban areas", *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, no. 1, pp. 161–172, 2012.

[146] Y.-T. Liow and T. Pavlidis, "Use of shadows for extracting buildings in aerial images", *Computer Vision, Graphics, and Image Processing*, vol. 49, no. 2, pp. 242–277, 1990.

[147] J. C. McGlone and J. A. Shufelt, "Projective and object space geometry for monocular building extraction", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1994, 54—61.

[148] F. Dornaika, A. Moujahid, A. Bosaghzadeh, Y. El Merabet, and Y. Ruichek, "Object classification using hybrid holistic descriptors: Application to building detection in aerial orthophotos", *Polibits*, no. 51, pp. 11–17, 2015.

[149] H. Baluyan, B. Joshi, A. Al Hinai, and W. L. Woon, "Novel approach for rooftop detection using support vector machine", *ISRN Machine Vision*, vol. 2013, 2013.

[150] T.-T. Ngo, C. Collet, and V. Mazet, "Automatic rectangular building detection from vhr aerial imagery using shadow and image segmentation", in *Image Processing (ICIP), 2015 IEEE International Conference on*, IEEE, 2015, pp. 1483–1487.

[151] H. Mayer, "Automatic object extraction from aerial imagery—a survey focusing on buildings", *Computer vision and image understanding*, vol. 74, no. 2, pp. 138–149, 1999.

[152] C. Ünsalan and K. L. Boyer, "A system to detect houses and residential street networks in multispectral satellite images", *Computer Vision and Image Understanding*, vol. 98, no. 3, pp. 423–461, 2005.

[153] B. Sirmacek, H. Taubenbock, P. Reinartz, and M. Ehlers, "Performance evaluation for 3-d city model generation of six different dsms from air-and spaceborne sensors", *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, no. 1, pp. 59–70, 2012.

[154] M. Brédif, O. Tournaire, B. Vallet, and N. Champion, "Extracting polygonal building footprints from digital surface models: A fully-automatic global optimization framework", *ISPRS journal of photogrammetry and remote sensing*, vol. 77, pp. 57–65, 2013.

[155] F. Rottensteiner, J. Trinder, S. Clode, and K. Kubik, "Using the dempster–shafer method for the fusion of lidar data and multi-spectral images for building detection", *Information fusion*, vol. 6, no. 4, pp. 283–300, 2005.

[156] G. Shafer *et al.*, *A mathematical theory of evidence*. Princeton university press Princeton, 1976, vol. 1.

[157] W. Förstner, "A framework for low level feature extraction", in *European Conference on Computer Vision*, Springer, 1994, pp. 383–394.

[158] G. Sohn and I. Dowman, "Data fusion of high-resolution satellite imagery and lidar data for automatic building extraction", *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 62, no. 1, pp. 43–63, 2007.

[159] S. Zabuawala, H. Nguyen, H. Wei, and J. Yadegar, "Fusion of lidar and aerial imagery for accurate building footprint extraction", *Image Processing. Machine Vision Applications II*, vol. 7251, 72510Z–1, 2009.

[160] A. Turlapaty, B. Gokaraju, Q. Du, N. H. Younan, and J. V. Aanstoos, "A hybrid approach for building extraction from spaceborne multi-angular optical imagery", *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, no. 1, pp. 89–100, 2012.

[161] T. Partovi, R. Bahmanyar, T. Krauß, and P. Reinartz, "Building outline extraction using a heuristic approach based on generalization of line segments", *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 3, pp. 933–947, 2016.

[162] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps", *arXiv preprint arXiv:1312.6034*, 2013.

[163] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[164] M. Vakalopoulou, K. Karantzalos, N. Komodakis, and N. Paragios, "Building detection in very high resolution multispectral data with deep learning features", in *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, IEEE, 2015, pp. 1873–1876.

[165] K. Nogueira, O. A. Penatti, and J. A. dos Santos, "Towards better exploiting convolutional neural networks for remote sensing scene classification", *Pattern Recognition*, vol. 61, pp. 539–556, 2017.

[166] S. Kluckner, T. Mauthner, P. M. Roth, and H. Bischof, "Semantic classification in aerial imagery by integrating appearance and height information", in *Asian Conference on Computer Vision*, Springer, 2009, pp. 477–488.

[167] S. Kluckner and H. Bischof, "Semantic classification by covariance descriptors within a randomized forest", in *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, IEEE, 2009, pp. 665–672.

[168] V. Mnih and G. E. Hinton, "Learning to detect roads in high-resolution aerial images", in *European Conference on Computer Vision*, Springer, 2010, pp. 210–223.

[169] P. Dollar, Z. Tu, and S. Belongie, "Supervised learning of edges and object boundaries", in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, IEEE, vol. 2, 2006, pp. 1964–1971.

[170] V. Mnih, "Machine learning for aerial image labeling", PhD thesis, University of Toronto (Canada), 2013.

[171] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling", *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1915–1929, 2012.

[172] S. Saito, T. Yamashita, and Y. Aoki, "Multiple object extraction from aerial imagery with convolutional neural networks", *Electronic Imaging*, vol. 2016, no. 10, pp. 1–9, 2016.

[173] J. Yuan, "Automatic building extraction in aerial scenes using convolutional networks", *arXiv preprint arXiv:1602.06564*, 2016.

[174] T. Zuo, J. Feng, and X. Chen, "Hf-fcn: Hierarchically fused fully convolutional network for robust building extraction", in *Asian Conference on Computer Vision*, Springer, 2016, pp. 291–302.

[175] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark", in *IEEE International Symposium on Geoscience and Remote Sensing*, 2017.

[176] H. L. Yang, J. Yuan, D. Lunga, M. Laverdiere, A. Rose, and B. Bhaduri, "Building extraction at scale using convolutional neural network: Mapping of the united states", *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 8, pp. 2600–2614, 2018.

[177] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation", *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[178] S. J. M. Drozdzal, D. Vazquez, and A. R. Y. Bengio, "The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation",

[179] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation", in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2015, pp. 234–241.

[180] G. Wu, X. Shao, Z. Guo, Q. Chen, W. Yuan, X. Shi, Y. Xu, and R. Shibasaki, "Automatic building segmentation of aerial imagery using multi-constraint fully convolutional networks", *Remote Sensing*, vol. 10, no. 3, p. 407, 2018.

[181] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks", *arXiv preprint*, 2017.

[182] N. Merkle, P. Fischer, S. Auer, and R. Müller, "On the possibility of conditional adversarial networks for multi-sensor image matching", in *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, IEEE, 2017, pp. 2633–2636.

[183] D. Marmanis, W. Yao, F. Adam, M. Datcu, P. Reinartz, K. Schindler, J. D. Wegner, and U. Stilla, "Artificial generation of big data for improving image classification: A generative adversarial network approach on sar data", *arXiv preprint arXiv:1711.02010*, 2017.

[184] K. Davydova, S. Cui, and P. Reinartz, "Building footprint extraction from digital surface models using neural networks", in *Proceedings of SPIE*, vol. 10004, 2016, pp. 1–10.

[185] K. Bittner, S. Cui, and P. Reinartz, "Building extraction from remote sensing data using fully convolutional networks", *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences-ISPRS Archives*, vol. 42, no. W1, pp. 481–486, 2017.

[186] A. Lagrange, B. Le Saux, A. Beaupere, A. Boulch, A. Chan-Hon-Tong, S. Herbin, H. Randrianarivo, and M. Ferecatu, "Benchmarking classification of earth-observation data: From learning explicit features to convolutional networks", in *Geoscience and Remote Sensing Symposium (IGARSS), 2015 IEEE International*, IEEE, 2015, pp. 4173–4176.

[187] J. Sherrah, "Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery", *arXiv preprint arXiv:1606.02585*, 2016.

[188] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs", *arXiv preprint arXiv:1606.00915*, 2016.

[189] N. Audebert, B. Le Saux, and S. Lefèvre, "Semantic segmentation of earth observation data using multimodal and multi-scale deep networks", in *Asian Conference on Computer Vision*, Springer, 2016, pp. 180–196.

[190] Y. Xu, L. Wu, Z. Xie, and Z. Chen, "Building extraction in very high resolution remote sensing imagery using deep learning and guided filters", *Remote Sensing*, vol. 10, no. 1, p. 144, 2018.

[191] D. Marmanis, K. Schindler, J. D. Wegner, S. Galliani, M. Datcu, and U. Stilla, "Classification with an edge: Improving semantic image segmentation with boundary detection", *ISPRS J. Photogramm. Remote Sens.*, vol. 135, pp. 158–172, 2018.

[192] J. Hu, L. Mou, A. Schmitt, and X. X. Zhu, "Fusionet: A two-stream convolutional neural network for urban scene classification using polsar and hyperspectral data", in *Urban Remote Sensing Event (JURSE), 2017 Joint*, IEEE, 2017, pp. 1–4.

[193] L. Mou and X. X. Zhu, "Spatiotemporal scene interpretation of space videos via deep neural network and tracklet analysis", in *Geoscience and Remote Sensing Symposium (IGARSS), 2016 IEEE International*, IEEE, 2016, pp. 1823–1826.

[194] P. Schuegraf and K. Bittner, "Automatic building footprint extraction from multi-resolution remote sensing images using a hybrid fcn", *ISPRS International Journal of Geo-Information*, vol. 8, no. 4, p. 191, 2019.

[195] N. R. Chrisman, "The error component in spatial data", *Geographical information systems*, vol. 1, no. 12, pp. 165–174, 1991.

[196] A. M. Felicísimo, "Parametric statistical method for error detection in digital elevation models", *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 49, no. 4, pp. 29–33, 1994.

[197] C. López, "Improving the elevation accuracy of digital elevation models: A comparison of some error detection procedures", *Transactions in GIS*, vol. 4, no. 1, pp. 43–64, 2000.

[198] P. Wang, "Applying two dimensional kalman filtering for digital terrain modelling", *Proceedings of International Archives of Photogrammetry, Remote Sensing, and Spatial Information Sciences*, pp. 649–656, 1998.

[199] J. P. Walker and G. R. Willgoose, "A comparative study of australian cartometric and photogrammetric digital elevation model accuracy", *Photogrammetric Engineering & Remote Sensing*, vol. 72, no. 7, pp. 771–779, 2006.

[200] K. Arrell, S. Wise, J. Wood, and D. Donoghue, "Spectral filtering as a method of visualising and removing striped artefacts in digital elevation data", *Earth Surface Processes and Landforms: The Journal of the British Geomorphological Research Group*, vol. 33, no. 6, pp. 943–961, 2008.

[201] P. Goovaerts *et al.*, *Geostatistics for natural resources evaluation*. Oxford University Press on Demand, 1997.

[202] E. Anderson, J. Thompson, and R. Austin, "Lidar density and linear interpolator effects on elevation estimates", *International Journal of Remote Sensing*, vol. 26, no. 18, pp. 3889–3900, 2005.

[203] S. Smith, D. Holland, and P. Longley, "Quantifying interpolation errors in urban airborne laser scanning models", *Geographical Analysis*, vol. 37, no. 2, pp. 200–224, 2005.

[204] W. Shi and Y Tian, "A hybrid interpolation method for the refinement of a regular grid digital elevation model", *International Journal of Geographical Information Science*, vol. 20, no. 1, pp. 53–67, 2006.

[205] R. L. Hardy, "Multiquadric equations of topography and other irregular surfaces", *Journal of geophysical research*, vol. 76, no. 8, pp. 1905–1915, 1971.

[206] R. Franke, "Scattered data interpolation: Tests of some methods", *Mathematics of computation*, vol. 38, no. 157, pp. 181–200, 1982.

[207] C. Chen and Y. Li, "A robust multiquadric method for digital elevation model construction", *Mathematical Geosciences*, vol. 45, no. 3, pp. 297–319, 2013.

[208] D. Milledge, S. N. Lane, and J Warburton, "Optimization of stereo-matching algorithms using existing dem data", *Photogrammetric Engineering & Remote Sensing*, vol. 75, no. 3, pp. 323–333, 2009.

[209] M. Karkee, B. L. Steward, and S. A. Aziz, "Improving quality of public domain digital elevation models through data fusion", *Biosystems Engineering*, vol. 101, no. 3, pp. 293–305, 2008.

[210] T. Krauß and P. Reinartz, "Enhancement of dense urban digital surface models from vhr optical satellite stereo data by pre-segmentation and object detection", *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 38, no. 1, p. 6, 2010.

[211] D. Poli and P. Soille, "Refinement of digital surface models through constrained connectivity partitioning of optical imagery", *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 38, no. 4/ W19, 2011.

[212] P. Soille, "Constrained connectivity for hierarchical image partitioning and simplification", *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 7, pp. 1132–1145, 2008.

[213] D. Poli and I. Caravaggi, "Digital surface modelling and 3d information extraction from spaceborne very high resolution stereo pairs", *JRC Scientific and Technical Reports, Ispra*, pp. 1–31, 2012.

[214] D. Canu, J.-P. Gambotto, J. A. Sirat, and N. Ayache, "Reconstruction of buildings from multiple high resolution images", in *Proceedings of 3rd IEEE International Conference on Image Processing*, IEEE, vol. 2, 1996, pp. 621–624.

[215] S. Vinson, L. D. Cohen, and F. Perlant, "Extraction of rectangular buildings in aaerial images", in *Proceedings of the Scandinavian Conference on Image Analysis*, 2001, pp. 431–438.

[216] L. D. Cohen and S. Vinson, "Segmentation of complex buildings from aerial images and 3d surface reconstruction", in *Sixth IEEE Workshop on Applications of Computer Vision, 2002.(WACV 2002). Proceedings.*, IEEE, 2002, pp. 215–219.

[217] B Sirmacek, P dAngelo, and P Reinartz, "Detecting complex building shapes in panchromatic satellite images for digital elevation model enhancement", in *ISPRS Workshop on Modeling of Optical Airborne and Space Borne Sensors*, Citeseer, 2010.

[218] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network", in *Advances in neural information processing systems*, 2014, pp. 2366–2374.

[219] H. Tian, B. Zhuang, Y. Hua, and A. Cai, "Depth inference with convolutional neural network", in *2014 IEEE Visual Communications and Image Processing Conference*, IEEE, 2014, pp. 169–172.

[220] B. Li, C. Shen, Y. Dai, A. Van Den Hengel, and M. He, "Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1119–1127.

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

[222] J. Zhu and R. Ma, *Real-time depth estimation from 2d images*, 2016.

[223] J. Jeon and S. Lee, "Reconstruction-based pairwise depth dataset for depth image enhancement using cnn", 2018, pp. 422–438.

[224] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese, "3d-r2n2: A unified approach for single and multi-view 3d object reconstruction", in *European conference on computer vision*, Springer, 2016, pp. 628–644.

[225] A. Dai, C. R. Qi, and M. Nießner, "Shape completion using 3d-encoder-predictor cnns and shape synthesis", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 3, 2017.

[226] J. Wu, C. Zhang, T. Xue, B. Freeman, and J. Tenenbaum, "Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling", in *Advances in Neural Information Processing Systems*, 2016, pp. 82–90.

[227] D. J. Rezende, S. A. Eslami, S. Mohamed, P. Battaglia, M. Jaderberg, and N. Heess, "Unsupervised learning of 3d structure from images", in *Advances in Neural Information Processing Systems*, 2016, pp. 4996–5004.

[228] B. Yang, H. Wen, S. Wang, R. Clark, A. Markham, and N. Trigoni, "3d object reconstruction from a single depth view with adversarial learning", *arXiv preprint arXiv:1708.07969*, 2017.

[229] Y.-X. Guo and X. Tong, "View-volume network for semantic scene completion from a single depth image", *arXiv preprint arXiv:1806.05361*, 2018.

[230] L. Mou and X. X. Zhu, "Im2height: Height estimation from single monocular imagery via fully residual convolutional-deconvolutional network", *Arxiv Prepr. Arxiv:1802.10249*, 2018.

[231] P. Ghamisi and N. Yokoya, "Img2dsm: Height simulation from single imagery using conditional generative adversarial net", *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 794–798, 2018.

[232] G. Costante, T. A. Ciarfuglia, and F. Biondi, "Towards monocular digital elevation model (dem) estimation by convolutional neural networks-application on synthetic aperture radar images", in *EUSAR 2018; 12th European Conference on Synthetic Aperture Radar*, VDE, 2018, pp. 1–6.

[233] D. Paschalidou, O. Ulusoy, C. Schmitt, L. Van Gool, and A. Geiger, "Raynet: Learning volumetric 3d reconstruction with ray potentials", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3897–3906.

[234] S. Srivastava, M. Volpi, and D. Tuia, "Joint height estimation and semantic labeling of monocular aerial images with cnns", in *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, IEEE, 2017, pp. 5173–5176.

[235] S. Ahmadi, M. V. Zoej, H. Ebadi, H. A. Moghaddam, and A. Mohammadzadeh, "Automatic urban building boundary extraction from high resolution aerial images using an innovative model of active contours", *International Journal of Applied Earth Observation and Geoinformation*, vol. 12, no. 3, pp. 150–157, 2010.

[236] G Sohn and I. Dowman, "Extraction of buildings from high resolution satellite data", *Automated Extraction of Man-Made Objects from Aerial and Space Images (III). Balkema Publishers, Lisse*, pp. 345–355, 2001.

[237] C. Lin and R. Nevatia, "Building detection and description from a single intensity image", *Computer vision and image understanding*, vol. 72, no. 2, pp. 101–121, 1998.

[238] Z. Kim and R. Nevatia, "Uncertain reasoning and learning for feature grouping", *Computer Vision and Image Understanding*, vol. 76, no. 3, pp. 278–288, 1999.

[239] D. H. Lee, K. M. Lee, and S. U. Lee, "Fusion of lidar and imagery for reliable building extraction", *Photogrammetric Engineering & Remote Sensing*, vol. 74, no. 2, pp. 215–225, 2008.

[240] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition", *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[241] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[242] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks", in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 2011, pp. 315–323.

[243] 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*, vol. 4, no. 2, pp. 26–31, 2012.

[244] S. Ruder, "An overview of gradient descent optimization algorithms", *arXiv preprint arXiv:1609.04747*, 2016.

[245] H. Le and A. Borji, "What are the Receptive, Effective Receptive, and Projective Fields of Neurons in Convolutional Neural Networks?", *arXiv preprint arXiv:1705.07049*, 2017.

[246] R. Qin, J. Tian, and P. Reinartz, "Spatiotemporal inferences for use in building detection using series of very-high-resolution space-borne stereo images", *International Journal of Remote Sensing*, vol. 37, no. 15, pp. 3455–3476, 2016.

[247] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding", *arXiv preprint arXiv:1408.5093*, 2014.

[248] K. Jacobsen, "Dem generation from satellite data", *European Association of Remote Sensing Laboratories Ghent*, vol. 273276, no. 4, 2003.

[249] U. G. Sefercik, "Productivity of terrasar-x 3d data in urban areas: A case study in trento", *European Journal of Remote Sensing*, vol. 46, no. 1, pp. 597–612, 2013.

[250] K. Bittner, P d'Angelo, M Körner, and P Reinartz, "Automatic large-scale 3d building shape refinement using conditional generative adversarial networks", *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 422, pp. 103–108, 2018.

[251] K. Bittner and M. Körner, "Automatic large-scale 3d building shape refinement using conditional generative adversarial networks", in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2018, pp. 2000–2002.

[252] G Gröger, T. Kolbe, C Nagel, and K. Häfele, "Ogc city geography markup language (citygml) encoding standard, version 2.0, ogc doc no. 12-019", *Open Geospatial Consortium*, 2012.

[253] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks", in *Computer Vision (ICCV), 2017 IEEE International Conference on*, IEEE, 2017, pp. 2813–2821.

[254] J. R. Shewchuk, "Triangle: Engineering a 2d quality mesh generator and delaunay triangulator", in *Applied computational geometry towards geometric engineering*, Springer, 1996, pp. 203–222.

[255] B. N. Delaunay, "Sur la Sphère Vide", *Bulletin of Academy of Sciences of the USSR*, no. 6, pp. 793–800, 1934.

[256] J. Höhle and M. Höhle, "Accuracy assessment of digital elevation models by means of robust statistical methods", *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 64, no. 4, pp. 398–406, 2009.

[257] O. Sener and V. Koltun, "Multi-task learning as multi-objective optimization", in *Advances in Neural Information Processing Systems*, 2018, pp. 527–538.

[258] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[259] J. Hu, M. Ozay, Y. Zhang, and T. Okatani, "Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries", *Arxiv Prepr. Arxiv:1803.08673*, 2018.

[260] J. Zhang, T. Zhu, Y. Tang, and W. Zhang, "Geostatistical approaches to refinement of digital elevation data", *Geo-spatial Information Science*, vol. 17, no. 4, pp. 181–189, 2014.

[261] A. F. Elaksher and J. Bethel, "Refinement of digital elevation models in urban areas using breaklines via a multi-photo least squares matching algorithm", *Journal of Terrestrial Observation*, vol. 2, no. 2, p. 7, 2010.

[262] M. L. Hobi and C. Ginzler, "Accuracy assessment of digital surface models based on worldview-2 and ads80 stereo remote sensing data", *Sensors*, vol. 12, no. 5, pp. 6347–6368, 2012.

[263] K. Bittner, M. Körner, and P. Reinartz, "DSM Building Shape Refinement from Combined Remote Sensing Images based on WNet-cGANs", in *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, 2019, pp. 783–786.

[264] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier Nonlinearities improve Neural Network Acoustic Models", in *ICML Workshop on Deep Learning for Audio, Speech and Language Processing*, 2013.

[265] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. P. Smolley, "On the Effectiveness of Least Squares Generative Adversarial Networks", *arXiv preprint arXiv:1712.06391*, 2017.

# Appendix A

**K. Bittner, F. Adam, S. Cui, M. Körner, and P. Reinartz. Building Footprint Extraction from VHR Remote Sensing Images Combined with Normalized DSMs using Fused Fully Convolutional Networks. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 11, no. 8, 2018**

https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8447548

# Appendix B

**K. Bittner, P. d'Angelo, M. Körner, and P. Reinartz. DSM-to-LoD2: Spaceborne Stereo Digital Surface Model Refinement. Remote Sensing, vol. 10, no. 12, 2018**

https://www.mdpi.com/2072-4292/10/12/1926

# Appendix C

**K. Bittner, M. Körner, F. Fraundorfer, and P. Reinartz, Multi-Task cGAN for Simultaneous Spaceborne DSM Refinement and Roof-Type Classification, Remote Sensing, vol. 11, no. 11, p. 1262, 2019.**

https://www.mdpi.com/2072-4292/11/11/1262

# Appendix D

**K. Bittner, M. Körner, and P. Reinartz, Late or Earlier Information Fusion from Depth and Spectral Data? Large-Scale Digital Surface Model Refinement by Hybrid-cGAN, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2019.**

http://openaccess.thecvf.com/content_CVPRW_2019/html/EarthVision/Bittner_Late_or_Earlier_Information_Fusion_From_Depth_and_Spectral_Data_CVPRW_2019_paper

# Appendix E

# Related Publications

## E.1 Journals

- **K. Bittner**, M. Körner, F. Fraundorfer, and P. Reinartz, "Multi-Task cGAN for Simultaneous Spaceborne DSM Refinement and Roof-Type Classification", *Remote Sensing*, vol. 11, no. 11, p. 1262, 2019.
- P. Schuegraf, and **K. Bittner**, "Automatic Building Footprint Extraction from Multi-Resolution Remote Sensing Images using a Hybrid-FCN", *ISPRS International Journal of Geo-Information*, vol. 8, no. 4, p. 191, 2019.
- **K. Bittner**, P. d'Angelo, M. Körner, and P. Reinartz, "DSM-to-LoD2: Spaceborne stereo digital surface model refinement", *Remote Sensing*, 2018, 10, 1926.
- **K. Bittner**, F. Adam, S. Cui, M. Körner, and P. Reinartz, "Building Footprint Extraction from VHR Remote Sensing Images Combined with Normalized DSMs using Fused Fully Convolutional Networks", *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 8, pp. 2615-2629, 2018.

## E.2 Conferences

- **K. Bittner**, M. Körner, and P. Reinartz, "Late or Earlier Information Fusion from Depth and Spectral Data? Large-Scale Digital Surface Model Refinement by Hybrid-cGAN", *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019.
- **K. Bittner**, M. Körner, and P. Reinartz, "DSM Building Shape Refinement from Combined Remote Sensing Images based on WNet-cGAN", in *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, pp. 783-786, 2019.

- P. d'Angelo, D. Cerra, S. Azimi, N. Merkle, J. Tian, S. Auer, M. Pato, R. Reyes, X. Zhuo, **K. Bittner**, T. Krauss, and P. Reinartz "3D Semantic Segmentation from Multi-View Optical Images", in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, to appear, 2019.

- D. Cerra, M. Pato, E. Carmona, S. Azimi, J. Tian, R. Bahmanyar, F. Kurz, E. Vig, **K. Bittner**, and C. Henry, "Combining Deep and Shallow Neural Networks with Ad HOC Detectors for the Classification of Complex Multi-Modal Urban Scenes", In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 3856-3859, 2018.

- **K. Bittner**, P. d'Angelo, M. Körner, and P. Reinartz, "Automatic Large-Scale 3D Building Shape Refinement using Conditional Generative Adversarial Networks", *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42, 2, 2018.

- **K. Bittner**, and M. Körner, "Automatic Large-Scale 3D Building Shape Refinement using Conditional Generative Adversarial Networks", In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, Salt Lake City, UT, USA, 18-22 June 2018; pp. 1887-1889.

- **K. Bittner**, S. Cui, and P. Reinartz, "Building Extraction from Remote Sensing Data using Fully Convolutional Networks", *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences-ISPRS Archives*, vol. 42, no. W1, pp. 481-486, 2017.

- **K. Davydova**, S. Cui, and P. Reinartz, "Building Footprint Extraction from Digital Surface Models using Neural Networks", In *SPIE Remote Sensing, International Society for Optics and Photonics*, pp. 100040J-100040J, 2016.