TU WIEN Informatics

Alpine Terrain Relighting

Deep-Learning basierende Einzelbild Schattenentfernung mit Digitalen Höhenmodellen.

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Media and Human-Centered Computing

eingereicht von

Maximilian Staats, Bsc
Matrikelnummer 01624279

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Assistant Prof. Dr.in techn. Manuela Waldner
Mitwirkung: Assistant Prof. Doctor Pedro Hermosilla Casajus

Wien, 23. Jänner 2024

Maximilian Staats	Manuela Waldner

TU WIEN Informatics

Alpine Terrain Relighting

Deep-Learning Based Single Image Shadow-Removal with Digital Elevation Models

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Media and Human-Centered Computing

by

Maximilian Staats, Bsc
Registration Number 01624279

to the Faculty of Informatics

at the TU Wien

Advisor: Assistant Prof. Dr.in techn. Manuela Waldner
Assistance: Assistant Prof. Doctor Pedro Hermosilla Casajus

Vienna, 23rd January, 2024

Maximilian Staats	Manuela Waldner

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Erklärung zur Verfassung der Arbeit

Maximilian Staats, Bsc

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 23. Jänner 2024

Maximilian Staats

Danksagung

In erster Linie möchte ich meiner Betreuerin, Manuela Waldner, für ihre Unterstützung, ihr Engagement und die Menge an Zeit danken, die sie in regelmäßige Treffen mit mir investiert hat. Ihr strukturiertes, konstruktives Feedback und ihr Interesse an den Ergebnissen während der vielen Iterationen haben maßgeblich dazu beigetragen, den Verlauf dieser Forschung zu gestalten.

Darüber hinaus möchte ich Pedro Hermosilla Casajus meinen Dank für seine Unterstützung und seine Verfügbarkeit für alle Arten von Fragen zum Thema Computer Vision aussprechen.

Besonderer Dank gilt Adam Celarek und Johannes Eschner für ihre nützlichen Beiträge zur Verbesserung der Qualität des Schattenwurf-Algorithmus und Adam dafür, dass er sich die Zeit genommen hat, die großen Mengen an Höhendaten zu übertragen.

Ein großes Dankeschön an Angeliki Grammatikaki für ihre großzügige Zeit und Mühe bei der Erläuterung der Feinheiten von QGIS und DEMs, die mir beim Start des Projekts geholfen haben.

Schließlich möchte ich meiner Familie und meinen Freunden meine Anerkennung für ihre unermüdliche emotionale und finanzielle Unterstützung während der vielen Jahre, in denen ich mein Studium verfolgt habe, aussprechen. Ihre Ermutigungen und Unterstützungen war eine ständige Quelle der Kraft, und dafür bin ich zutiefst dankbar.

Acknowledgements

First and foremost, I would like to thank my advisor, Manuela Waldner, for her support, commitment, and the significant amount of time invested in conducting regular meetings with me. Her structured, constructive feedback and her interest in the results throughout the many iterations have been instrumental in shaping the trajectory of this research.

Furthermore, I would like to express my gratitude to Pedro Hermosilla Casajus for his assistance and availability for all kinds of questions regarding computer vision topics.

Special thanks to Adam Celarek and Johannes Eschner for their useful inputs to improve the quality of the shadowing algorithm, and to Adam for taking the time to transfer the huge amounts of elevation data.

A huge thanks to Angeliki Grammatikaki for her generous time and effort in explaining the intricacies QGIS and DEMs which really helped me to get started with the project.

Finally, I express my deepest appreciation to my family and friends for their unwavering emotional and financial support throughout the many years I pursued my studies. Their encouragement has been a constant source of strength and for that, I am profoundly grateful.

Kurzfassung

Luftbilder zusammen mit digitalen Höhenmodellen (DHM) ermöglichen die Darstellung von 3D-Repräsentationen der Erde, einschließlich alpiner Gebiete. Diese virtuellen Landschaften bieten die Möglichkeit, Lichtverhältnisse zu verschiedenen Tageszeiten zu simulieren, was die Planung von Bergtouren vereinfachen kann. Allerdings enthalten die als Textur verwendeten Orthofotos oft große Schatten von Bergen und Felsen, die die visuelle Qualität der künstlich beleuchteten Texturen erheblich beeinträchtigen. Der notwendige Prozess, um Schatten aus Einzelbildern zu entfernen, stellt ein entscheidendes Problem für das Gebiet der Computer Vision dar und dient auch als Voraussetzung für viele andere Aufgaben wie Segmentierung und Klassifizierung. Einige vielversprechende Ansätze wurden bereits entwickelt, aber im Gegensatz zu früheren Methoden versucht diese Arbeit mithilfe von den Verfügbaren DHMs den Schattenentfernungsprozess zu verbessern. Schatten in Orthofotos sind inhärent mit der zugrunde liegenden raumbezogenen Topologie verbunden und DHMs bieten eine wertvolle Informationsquelle, um Schatteneffekte zu verringern. Daher beschäftigt sich diese Arbeit mit der Integration von DHMs in eine moderne Deep-Learning Pipeline. DHMs werden auf ihre Rolle bei der Erzeugung von Trainingsdatensätzen und als zusätzlicher Input für ein multimodales Netzwerk untersucht. Insbesondere wird die aus DHMs abgeleitete 3D-Geometrie, komplementiert durch Raytracing, verwendet, um künstliche Schatten mit realistischen Formen zu erzeugen. Anschließend wird ein Experiment mit dem erstellten Datensatz durchgeführt, um empirisch und qualitativ zu prüfen, ob zusätzliche Höhendaten die Leistung der Modelle verbessern können. Darüber hinaus wurde die Fähigkeit der Modelle, von künstlichen Schatten auf reale Schatten zu verallgemeinern, geprüft. Das Experiment mit virtuellen Schatten zeigte, dass die Bereitstellung zusätzlicher Höhendaten für das Schattenentfernungsnetzwerk signifikant bessere Ergebnisse mit einer mittleren bis großen Effektgröße liefert. Anfänglich konnte keiner der trainierten Modelle auf echte Schatten verallgemeinern. Das Verkleinern des Datensatzes auf eine niedrigere Detailstufe verringerte dieses Problem. Zusammen mit einer Analyse der Ausgaben jeder Netzwerkschicht wurde geschlussfolgert, dass der Grund für die unzureichende Leistung bei echten Schatten kleine im Trainingsset verbliebene echte Schatten sind. Die aus dieser Kenntnis gewonnenen verbesserten Modelle wurden einer visuellen Analyse unterzogen und zeigten, dass Höhendaten und die generierten realistischeren Schattenformen zu sichtbaren Verbesserungen bei der Verallgemeinerungsfähigkeit der Modelle beitragen.

Abstract

Aerial orthophotos together with digital elevation models (DEMs) allow the rendering of 3D representations of the earth, including alpine terrain. These virtual landscapes provide the opportunity to simulate light conditions at different times of the day, aiding in trip planning. However, orthophotos used as texture often contain large shadows stemming from cliffs and rocks, which significantly impact the visual quality of relighted textures. The necessary single-image shadow-removal process presents a crucial problem for the computer vision domain, which also functions as a prerequisite for many other tasks like segmentation and classification. Many promising approaches have already been developed, but unlike previous methods, this study tries to capitalize on the availability of DEMs to enhance the shadow removal process. Shadows in orthophotos are inherently linked to the underlying geospatial topology, and DEMs provide a valuable source of information for mitigating their impact. Therefore, this thesis explores the integration of DEMs into a state-of-the-art deep learning pipeline. DEMs are examined for their role in generating training sets and as supplementary input for a multi-modal network. Notably, 3D geometry derived from DEMs complemented by ray-tracing is used to generate artificial shadows with realistic shapes. Subsequently, an experiment is conducted with the created dataset to empirically test if additional elevation data is beneficial for the performance of the models. Additionally, the model's ability to generalize from artificial to real shadows was probed. The experiment on virtual shadows showed that providing additional elevation data to the shadow-removal network does yield significantly better results with a medium to large effect size. Initially, all trained models failed to generalize to real shadow data. Downsizing the dataset to a lower level of detail mitigated this problem. Together with an analysis of the output of each network layer, it was concluded that the reason for the subpar real data performance are remaining small-scale shadows in the train set. A visual analysis of the improved models showed noticeable improvements with the generated realistic shadow shapes compared to random ones. Moreover, the utility of additional elevation data as input for the models was demonstrated.

Contents

CHAPTER 1

Introduction

1.1 Motivation

Planning hiking and touring ski routes in alpine areas beyond known paths can be difficult due to limited information about the terrain. Conventional 2D maps help to overcome this knowledge gap, but they lack the ability to interactively navigate the terrain and rely on the user to create an accurate cognitive 3D model of the observed area. High-resolution terrain models and orthophotos captured by airplanes or satellites can help to create 3D visualizations of landscapes (see Google Earth [Goo], Alpine Maps [Alp], Flight Simulator [Fli]). These 3D landscape models can help hikers locate their position more accurately in comparison to conventional 2D maps [SP07] and make navigation interactive. This can enhance the overall understanding of the terrain and streamline the planning process for a trip.

Furthermore, these visualizations provide the opportunity to realistically simulate sun exposure in the given terrain. This can be achieved by using a 3D terrain model and established computer graphics methods like shadow maps [Wil78] or ray-tracing [App68]. Additionally, the amount of solar radiation a hillside receives and at what time of day can be displayed in real time. This information can be essential in assessing avalanche risks [HNL09], the physiological load of the trip due to heat exposure, or the ideal time to ascend a mountain. Services like *Shadow Map* [Sha] already offer sunlight and shadow simulations with 3D maps. However, only a simple texture is used to display the terrain and buildings. Orthophotos would offer more information about the vegetation and terrain conditions, but they also contain real shadows cast by hillsides, rocks, and cliffs, which can be confused with virtual shadows. This can be irritating for the user and hinder correct assessment (see Figure 1.1). Therefore, it is advantageous to remove large-scale shadows from the orthophotos to make the artificial shadows easier to interpret and more visually appealing.

(a) (b)

Figure 1.1: Examples from Google Earth [Goo] (a) and Alpine Maps [Alp] (b) demonstrate how virtual and real shadows combined can create confusing visualizations.

1.2 Shadow-Removal

Within the image domain, shadows correspond to regions characterized by a reduced photon flux reaching the sensors, thus resulting in a reduced capacity to extract information. This means that shadow-removal is more of a reconstruction process than a removal process. The goal is to restore the information not transmitted by the photons. However, shadow-removal is a widely used term in the community and in many papers; hence, shadow-removal will be used throughout the thesis.

Shadows depend on many factors, like the light source, how much diffuse or direct light is emitted, the material and shape of the occluding object, as well as the properties of the surface the shadow is cast on [VST+23]. In addition, shadow properties are also dependent on the image capturing process [AHO10]. All these factors combined make it very difficult to infer a shadow-free version of an image and reconstruct the missing information in the unlit areas.

Shadow removal is a fundamental problem in the computer vision domain. It offers use cases for many tasks like recognition, tracking, segmentation and, in our case, it can also be used to relight an image. There are already many approaches to address this problem [FHLD05, QTH+17, LS19, WLY18, AI22] that show success in their tested domains. However, shadow removal is a very diverse and complex problem, and a general solution is still to be developed. In contrast to these single-image methods, alternative approaches, such as those demonstrated by Rahman et al. [RMML19], adopt a multi-image strategy. In this method, multiple images captured on the same day are utilized to eliminate shadows from orthophotos. This is achieved by capitalizing on the dynamic nature of shadows, which shift throughout the day. The process involves seamlessly stitching together shadow-free areas to create a shadow-free output. Nevertheless, most of the available orthophotos are not captured at sufficient frequency, not to mention multiple

times a day. Thus, we focus on single-image shadow removal during this thesis.

In recent years, the shadow-removal approach has shifted towards supervised deep-learning methods [AI22, QTH+17, LS19, WLY18]. Here, one of the biggest challenges is the creation of a representative training dataset containing shadow and shadow-free image pairs. There are some available datasets, like the Image Shadow Triplets Dataset (ISTD) [WLY18], the ISTD+ with corrected colors [LS19] or the Shadow-Removal Dataset (SRD) [QTH+17]. However, these datasets are relatively small (between 1800 and 3000 images) and have a limited variety of scenes due to the tedious manual creation process. This limited amount of available data also restricts the size of the trainable networks [QTH+17]. The images are created by taking two consecutive photos of a scene, one with an occluding object and one without. Figure 1.2 and Figure 1.3 show some of the scenes and shadows depicted in the ISTD and SRD respectively. Nevertheless, none of these datasets includes aerial orthophotos and creating a new orthophoto datasets with this method is not feasible.

Figure 1.2: Examples from the ISTD [WLY18]. The input images are at the top, with their respective ground truth images at the bottom.

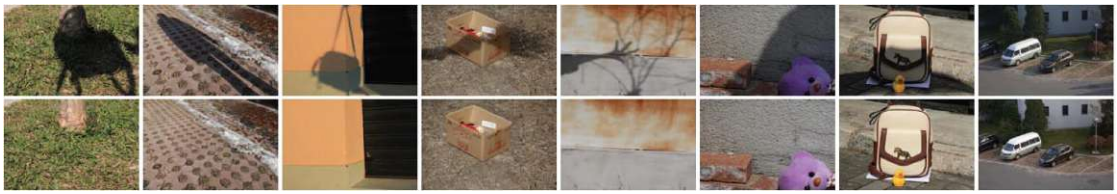


Figure 1.3: Examples from the SRD [QTH+17]. The input images are at the top, with their respective ground truth images at the bottom.

Another approach to producing training data is to take more or less “shadow-free” images and introduce synthetic shadows to the image. Morales et al. [MHT19] use this method to add procedural generated shadows to aerial satellite images. In their shadow creation process, they use Perlin noise and apply thresholds to produce random cloud-like shadow shapes. These artificial shadows enable their model to learn to remove similar real shadows cast by clouds. However, in our case, the orthophotos do not contain shadows cast by clouds but by the terrain, which produces widely different shapes depending on the underlying geospatial topology. In order to create artificial shadows cast by

the terrain, a DEM of the considered area will be used to render physically plausible shadows with a rendering engine and ray-tracing. This method allows the creation of a novel dataset depicting orthophotos of alpine terrain that will also be used in a later experiment.

1.3 Research Questions

This thesis focuses on the utility of DEMs in the context of shadow removal and tries to answer the following two research questions:

1) **RQ1**: How can terrain elevation data and orthophotos be encoded and incorporated into a deep-learning shadow-removal pipeline?

2) **RQ2**: To what extent does additional elevation data affect the performance of deep learning shadow-removal models?

1.4 Methods Overview

To properly answer the stated research questions, an investigation into state-of-the-art shadow-removal techniques was conducted, and a strategy to incorporate DEMs into a shadow-removal pipeline was developed. As a crucial first step, a dataset creation pipeline was implemented, employing DEMs and a 3D rendering engine to create physically plausible artificial shadows. This pipeline facilitated the generation of a novel dataset comprising orthophotos of the Austrian Alps, utilized for the training of two distinct models: one with and one without elevation data. Afterward, both models were evaluated with established quantitative metrics to discern possible performance differences. Hypothetically, the additional elevation data should help the model learn better features and increase its shadow-removal capacity. Furthermore, the models' ability to remove real shadows was investigated through a qualitative visual analysis. These two evaluations will collectively contribute insights into answering **RQ2**.

To further assess whether the pipeline designed to address **RQ1** yields improved generalizability to real shadows, a qualitative comparison with a baseline dataset will be conducted. This dataset is created in the exact same way as the primary dataset, but with random shadow shapes generated using Perlin noise.

1.5 Contribution

This thesis provides the following main contributions to the field of single image shadow-removal in the context of orthophotos:

- **Shadow-Removal Dataset Generation Pipeline:**
  We introduce a novel dataset generation pipeline designed for orthophotos, leveraging DEMs to create realistically shaped artificial shadows. This pipeline enhances the realism of shadow shapes compared to randomly generated ones and has shown improvements in the model's generalizability to real data.
- **Qualitative Comparison of Model Performance:**
  Through a qualitative comparison, we showcase the improved performance of models trained on the dataset generated by our proposed pipeline.
- **Incorporation of Elevation Data in a State-of-the-Art Models**
  We introduce an approach to integrating elevation data into a state-of-the-art shadow- removal model by encoding the data into an additional image channel. This enables the model to learn joint multi-modal features for terrain elevation and shadow shapes.
- **Quantitative Comparison of Model Variants:**
  A quantitative comparison of the model's performance—with and without elevation data—is conducted and shows a statistically significant difference with small improvements for the model trained with elevation data.
- **Qualitative Analysis of Real Shadow Performance:**
  The thesis extends to a qualitative examination of the models' performance in handling real shadows. Through visual analyses, we demonstrate noticeable improvements when elevation data is integrated.

CHAPTER 2

Related Work

Early shadow-removal methods are based on physical shadow formation models and try to estimate various parameters like direct and indirect light intensities and the reflectance of different colors and materials. The established shadow formation model introduced by Shor et al. [SL08] and used by Arbel et al. [AHO10] and Abiko et al. [AI22] is formulated as follows:

$$I^{shadow}(x,\lambda) = a(x)L^a(x,\lambda)R(x,\lambda). \tag{2.1}$$

In this equation, the illumination intensity of a shadow region I^{shadow} at position x and light wave frequency λ is depending on the indirect (ambient) illumination L^a and the reflectance (albedo) R. a denotes an additional attenuation factor of the ambient light due to the occluding object. It was assumed that this factor is constant at different light wave frequencies. With this formation model, we can formulate the shadow removal process as estimating the direct illumination L^d and the reflectance of the direct light R together with the inverse of the attenuation factor a to get the lit intensity I^{lit}.

$$I^{lit} = L^d(x,\lambda)R(x,\lambda) + \frac{1}{a(x)}I^{shadow} \tag{2.2}$$

Estimating these parameters can happen in image space or implicitly in gradient space and is quite difficult, especially with single images. Furthermore, even if parameters are estimated accurately, the used illumination models are not perfect and only represent an approximation of the illumination process [AI22].

Regarding shadow-removal, as with many image-related tasks, deep-learning has shown a lot of potential in recent years. Here, we can differentiate between paired (supervised) and unpaired (unsupervised) approaches. Unpaired approaches like the Mask-ShadowGAN [HJFH19] apply a feature transformation and do not need shadow-free ground truth

images, which are difficult to obtain. However, unsupervised approaches like this proposed method tend to create blurred output images [AI22].

Supervised methods (e.g., [WLY18, AI22, QTH+17]) are trained on a set of shadowed and shadow-free image tuples. Within supervised deep learning methods, Conditional Generative Adversarial Networks (CGANs) have proven to be a successful approach to performing shadow-to-shadow-free image translation [WLY18, AI22]. In the approach by Wang et al. [WLY18] two sequential (STacked) CGANs (ST-CGAN) are used for shadow detection and removal. This approach was further adapted by Abiko et al. [AI22] with a channel attention network, which focuses on the correlation between color channels.

Additionally, methods for estimating shadow parameters with deep learning and using them to adjust the image with physical lighting models are promising [LS19]. This approach by Le & Samaras also uses the shadow formation model introduced in Eq. 2.1 as basis for their mapping function to transform a shadow pixel to its non-shadow equivalent.

For the experimental part of the thesis, the ST-CGAN will be used because it is an already well-established model, and it uses the U-Net architecture. This architecture can infer higher-level global features of the terrain (e.g., to determine shadows cast from cliffs on the other side of the observed area) better than continuous condition continuation [CZZY17]. Furthermore, it is easy to add an elevation channel to the model, in contrast to the channel attention approach from Abiko et al. [AI22] where a correlation between channels is assumed, which is probably not the case with the elevation channel. Last but not least, code is available for this method, which provides all the necessary implementation details, mitigates the risk of errors, and reduces the time for implementation.

The next sections go through developments in deep-learning-based shadow-removal in more detail, starting by explaining generative adversarial networks as a general-purpose image-to-image translation tool and how this knowledge was used to create the sophisticated ST-CGAN shadow-removal network. Additionally, a few other approaches will be discussed in more detail and why they were not considered for this thesis.

2.1 Generative Adversarial Networks

Generative Adversarial Networks (GANs) are an established machine learning framework to train generative models. It consists of two main components: a generator and a discriminator, which play a two-player min-max game. The generator tries to produce fake samples to spoof the discriminator, and the discriminator guesses if a sample comes from the generator or the data. Worded differently, the goal of the discriminator D is to minimize its prediction loss, and the generator G tries to maximize the loss of the discriminator or minimize the inverse loss. If the input for the generator is random noise denoted as z and the considered data is described as p_{data}, the objective function $\mathcal{L}$ of this min-max game is defined as follows:

$$\min_D \max_G \mathcal{L}_{GAN}(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[logD(x)] + \mathbb{E}_{z\sim p_z(z)}[log(1-D(G(z))]. \quad (2.3)$$

If G and D have enough time and capacity, the generator learns the distribution of the given data, and a point of convergence is reached. At this optimum $D(x) = \frac{1}{2}$ and the discriminator guesses correctly half of the time. Goodfellow et al. [GPAM+14] show that this point is a global optimum. G now produces random non-deterministic examples based on the distribution of the training data and the input noise z.

However, if we want more control over the output of the generator, it is useful to condition the model with additional input y, e.g., class labels, text, or images [MO14]. In that case, Equation 2.3 is extended to:

$$\min_D \max_G \mathcal{L}_{CGAN}(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[logD(x|y)] + \mathbb{E}_{z\sim p_z(z)}[log(1-D(G(z|y))]. \quad (2.4)$$

These Conditioned GANs (CGANs) have proven to be a very powerful general-purpose tool for image-to-image translation tasks [IZZE17]. Introducing an additional $L1$ norm to the objective function was also found to be beneficial. This term is directly used on the output of G and brings it closer to the ground truth while also preserving more detail as using $L2$ [IZZE17]. However, $L1$ is averaging the distance of the whole image, therefore still focusing on low-frequency structures. To mitigate this effect, the image can be split up into $N \times N$ patches and let the discriminator predict the probability of each patch being fake. This so-called PatchGAN encourages the production of higher-frequency structures and models texture loss, whereas the L1 norm focuses more on the lower frequencies [IZZE17]. The final objective function for a general-purpose image-to-image CGAN was defined by Isola et al. [IZZE17] as:

$$L1(G) = \mathbb{E}_{x,y,z}(||y - G(x,z)||), \quad (2.5)$$

$$G^* = \min_D \max_G \mathcal{L}_{CGAN}(D,G) + \lambda L1(G). \quad (2.6)$$

The additional variable λ denotes a weighing hyperparameter for the introduced L1 metric. Experiments have shown that the additional noise z has no benefit to the model, and it just learns to ignore it [IZZE17] and makes the model non-deterministic [CZZY17]. Therefore, the noise input parameter can be left out.

Many image-to-image translation tasks require obtaining the underling structure of the image throughout the translation process. However, the widely used encoder-decoder network structure for translation tasks forces all the information through a bottleneck, which makes it harder for the network to transfer detailed information from input to output. Therefore, skip connections from the ith to the $n-i$th layer are introduced to form a "U-Net" type architecture [RFB15]. This allows high-frequency structures to

be passed through the model more easily, maintaining detail in the images. Another architecture beneficial for transferring very localized features, e.g., as with colorization tasks, is the continuous condition concatenation [CZZY17]. This architecture is also based on convolutional layers to obtain conditional features but refrains from downsampling. With this architecture, the network can keep all the spatial information throughout the layers, but it cannot collect global features as well as the U-Net can.

2.2 Deep-Learning Based Shadow-Removal

As previously discussed, CGANs are a powerful general-purpose tool for image-to-image translation tasks. Thus, they can also be used for shadowed to shadow-free image translations [WLY18, AI22].

Wang et al. [WLY18] use two stacked CGANs (ST-CGAN) to achieve a successful single image shadow-removal pipeline. The first CGAN detects the shadow areas and produces a shadow mask, which is then given to the second CGAN as additional input to reconstruct the shadow areas. Many approaches address the problem of shadow removal in two consecutive parts: detection and removal, respectively [WLY18, AI22, MHT19]. However, the approach from Wang et al. [WLY18] is interesting because both CGANs are trained together, which encourages joint learning of features. To establish this joint learning process as an objective function, Equation 2.6 was adapted. In the resulting adversarial objective function $\mathcal{L}_{CGAN}$, G_1 describes the shadow detection generator, G_2 the shadow-removal generator. D_1 and D_2 describe their respective discriminators. The parameter x denotes the input image containing shadows, y signifies the ground truth shadow mask expressed as a binary mask, and r represents the ground truth shadow-free image. With that, the adversarial losses for the shadow detection $\mathcal{L}_{CGAN_1}$ and shadow removal $\mathcal{L}_{CGAN_2}$ are defined as:

$$\begin{aligned}\mathcal{L}_{CGAN_1}(G_1, D_1) = \mathbb{E}_{x,y\sim p_{data}(x,y)}[logD_1(x,y)]+ \\ \mathbb{E}_{x\sim p_{data}(x)}[log(1 - D_1(G_1(x))],\end{aligned} \tag{2.7}$$

$$\begin{aligned}\mathcal{L}_{CGAN_2}(G_2, D_2|G_1) = \mathbb{E}_{x,y,r\sim p_{data}(x,y,r)}[logD_2(x,y,r)]+ \\ \mathbb{E}_{x\sim p_{data}(x)}[log(1 - D_2(x, G_1(x), G_2(G_1(x)))].\end{aligned} \tag{2.8}$$

The $L1$ term for detection and removal was defined as:

$$L1_{data_1}(G_1) = \mathbb{E}_{x,y\sim p_{data}(x,y)}(||y - G_1(x)||), \tag{2.9}$$

$$L1_{data_2}(G_2|G_1) = \mathbb{E}_{x,r\sim p_{data}(x,r)}(||r - G_2(G_1(x))||). \tag{2.10}$$

This results in the following linear combination of the partial functions to achieve a combined measure for the joined learning algorithm:

$$\begin{aligned} &\min_{D_1,D_2} \max_{G_1,G_2} L1_{data_1}(G_1) + \lambda_1 L1_{data_2}(G_2|G_1) + \\ &\lambda_2 \mathcal{L}_{CGAN_1}(G_1, D_1) + \lambda_3 \mathcal{L}_{CGAN_2}(G_2, D_2|G_1). \end{aligned} \tag{2.11}$$

In this formulation λ denotes an additional hyperparameter to determine the individual weighing of each loss term.

The darkening that occurs in a shadow area is not uniformly distributed over the spectrum of visible light. Abiko et al. [AI22] use this knowledge to extend the ST-CGAN with an additional channel attention layer that models the correlation between different colors and achieve quite good results with their approach. As mentioned earlier, in our context, this attention layer introduces an additional variable and could lead to unwanted side effects when used together with an elevation channel due to the assumed correlation between the channels. This probably does not apply to our setup with elevation as the fourth image channel.

Other authors have already investigated shadow-removal in orthophotos. Morales et al. [MHT19] compared CGANs with two different architectures to remove shadows from 4-band satellite images with red (0.63- 0.7µm), green (0.53-0.59µm), blue (0.45-0.50µm) and near-infrared (0.752-0.885µm). They juxtapose U-Net to continuous condition concatenation in removing shadows cast by clouds. The models were trained on a dataset consisting of handpicked, shadow-free satellite images from Peru altered with artificial shadows. It was concluded that continuous condition concatenation is superior in performance and that the trained models can generalize well from virtual shadows to real ones. However, the ability to generalize was evaluated by visually inspecting some examples, and it is not made clear how many or by which criteria this inspection was done. A "Real vs. Fake" Mechanical Turk study similar to Isola et al. [IZZE17] and Zhu et al. [ZPIE17] could provide more profound insights. Nevertheless, the approach by Morales et al. comes closest to our goal of removing shadows from aerial orthophotos. Their method of introducing artificial shadows to handpicked "shadow-free" images was used later in this thesis as a basis for the proposed data generation pipeline. Similar to the approach of Wang et al. two cGANs are used to detect and remove shadows sequentially, but the networks are not trained together, which prevents the joint learning of features. For this and the earlier stated reasons, the ST-CGAN by Wang et al. was selected over the deep-learning architecture of Morales et al.. Other differences between the approach of Morales et al. and our goal include different types of data (4-Band images vs. RGB images), scale of the images (2.8m/px vs. 16-29cm/px) and the type of shadow that should be removed (shadows from clouds vs. shadows cast by terrain).

In addition to GAN based methods, other deep-learning based methods have emerged as well. They try to estimate shadow parameters and adjust the images accordingly. Le & Samaras [LS19] use a linear illumination transformation function with parameters estimated by a neural network to relight the image. This approach could also be applicable to our goal. The shadow parameter estimator network could be adapted to incorporate elevation data as a fourth channel or even with a separate elevation network. The two modalities, RGB and elevation (or depth), could be fused in a feature fusion layer similar to approaches for RGB-D object recognition [GJZ+19] before predicting the shadow parameters. Their shadow matte prediction network, which is used to combine the relit image with the shadow mask and shadowed image, would need to be adapted similar to the ST-CGAN to incorporate the elevation data. For the prediction of the shadow mask, they use the approach of Zhu et al. [ZDH+18]. This shadow-removal strategy was not selected because the ST-CGAN offers a more holistic approach to the problem and the method lacks implementation details.

Qu et al. [QTH+17] use three different networks: a global localization network (G-Net), an appearance modeling network (A-Net), and a semantic modeling network (S-Net). Introducing elevation data into this complex collaboration of networks was deemed too delicate and could easily lead to failure.

CHAPTER 3

Dataset Generation

The goal is to create shadow-free orthophotos maintaining as much detail as possible. As briefly discussed in Chapter 2, unsupervised training methods lead to blurry results [AI22] and therefore the supervised ST-CGAN approach was chosen. For supervised learning, a representative training dataset is essential to achieving robust results. However, as mentioned in the introduction, creating shadowed and shadow-free image pairs is a very tedious process, and natural scenes are never really shadow-free. There are always some small occluding objects, like leaves, pebbles, or the side wall of a crevice, that cast shadows. It is important to note that even the ISTD [WLY18] and the SRD [QTH+17] datasets are only shadow-free within a certain scale factor. Furthermore, taking consecutive photos with and without an occluding object is not always an option, for example, with orthophotos. The occluding objects necessary to cast shadows of sufficient size would be too big to manage, and producing a dataset with a similar technique as ISTD at that scale is simply not feasible.

One could obtain shadow-free versions of orthophotos by taking images before and after noon. However, light conditions change over time, and a ground truth image taken in the afternoon does not accurately represent a shadow-free image taken before noon. Additionally, natural scenes can contain surfaces that are in shadow all day (e.g., in trenches) or objects that move (e.g., trucks, planes), which alter the scene over time. Apart from that, the shading of objects is entirely different due to varying illumination. Moreover, capturing images before and after noon doubles the time needed and is accompanied by additional costs.

Therefore, we utilize elevation data together with ray-tracing to create a new *Alpine Shadow Dataset* (ASD) with physically plausible shadows. These can be cast on more or less shadow-free areas of the orthophotos and help to circumvent some of the problems discussed. Additionally, the virtual shadows correlate directly with the underlying terrain encoded in the DEM. It was assumed that this correlation would help the model generalize

to real shadows better, due to the fact that the terrain topology is directly responsible for the shadows.

3.1 Digital Elevation Models

Elevation data is often encoded as a height map consisting of a planimetric grid structure and altitude values. Apart from a few exceptions, like overhanging cliffs, every point (x, y) on Earth's surface can be represented with one altitude value z as a bivariate function $z = f(x, y)$. Thus, the Earth's relief can be considered 2.5D [PEH20]. This simplified representation of height data is typically called a Digital Elevation Model (DEM) and is one of the most commonly used forms of geographic information. Two types of surfaces are frequently represented: the ground surface called Digital Terrain Model (DTM) or the surface including all natural or man-made objects, called the Digital Surface Model (DSM) (see Figure 3.1). The altitude values can be captured with photogrammetry, Light Detection and Ranging (LIDAR) or Radio Detection and Ranging (RADAR) whereas the last two are more appropriate for DTM [PEH20].

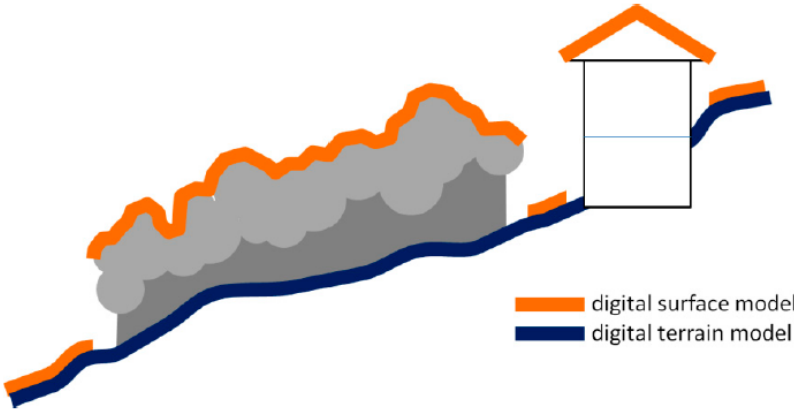


Figure 3.1: Visualization of the difference between a Digital Terrain Model (DTM) and a Digital Surface Model (DSM) [PEH20].

The input data for the dataset generation consists of a digital terrain model (DTM) and orthophotos of Austria. The data was obtained from Basemap [Bas] the official geodata source of the Austrian government. The DTM has a resolution of 1m/pixel and is encoded as a 32-bit float GeoTIFF. The GeoTIFF format is a compliant version of the Tag Image File Format (TIFF) containing georeference information as additional metadata and is commonly used to store DEMs [MR03]. The orthophotos have a resolution of 15-29cm/pixel and were captured by aerial surveying flights. There is also a digital surface model (DSM) available that also contains the height information of objects on the terrain. However, the DTM was chosen over the DSM because the DTM only includes the elevation data of the terrain and ignores the elevation of the tree tops, bushes, and

scrubs. The additional height information of vegetation would result in the generation of artificial shadows produced by this vegetation, which we want to ignore anyway.

3.2 Data Selection

As a basis for the dataset generation, a set of "shadow-free" areas was handpicked from the aerial images. Two main criteria were considered during the selection of these areas:

1. **No large-scale shadows.** Included shadows have to be smaller than 5 meters and shadows of vegetation can be ignored.
2. **Mountainous area.** Flat terrain can not be used for the shadow generation process and the corresponding DEM does not contain useful information for model.

These areas were sampled in a more or less random fashion, starting with an overview of the Austrian alps and spotting potential candidate areas. If an interesting area was found, it was magnified and checked for the criteria. The location of these areas can be seen in Figure 3.2. Hillsides directly facing the sun yielded the most shadow-free areas. It is important that these areas have a certain minimum size that is significantly larger than a single training image to allow shadow casting from surrounding areas onto the training image. The selected areas were between 18 and 300 times larger than the 256×256 pixel output images.

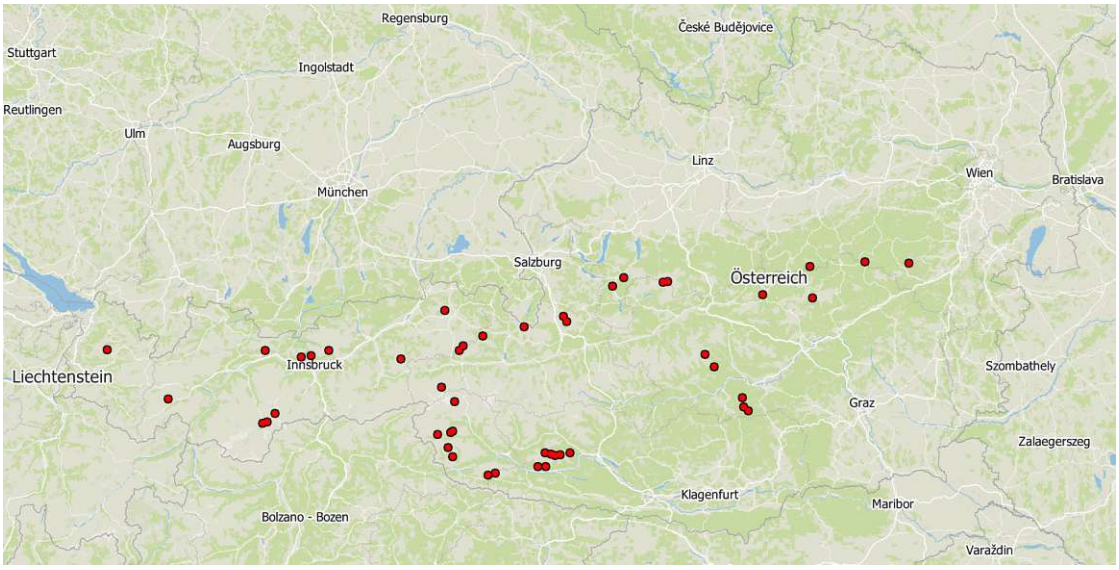


Figure 3.2: Sampling locations of the handpicked shadow-free areas.

The resulting set of 50 areas includes common terrain types from the Austrian alps like forest, fields, bushes, scree, snow patches, and rock, as well as some ponds, streets,

and cottages. Examples of the different terrain types can be seen in Figure 3.3. From each area, an image was extracted and encoded as a Portable Network Graphic (PNG) together with the corresponding DTM to enable the subsequent shadow rendering.

Figure 3.3: Examples of different types of terrain included in the Alpine Shadow Dataset.

3.3 Virtual Shadows

3.3.1 Shadow Mask Generation

As a first step, the DEM was converted to a 3D mesh to facilitate its integration into a rendering engine. For this process, the altitude values z at the raster image coordinates (x, y) can simply be interpreted as 3D vertices (x, y, z) and sequentially connected to form triangles. The included georeference information in the metadata of the GeoTIFF permits seamless scaling of the resultant mesh to real-world dimensions.

In order to create realistic shadow masks, the ray-tracing based Cycles Renderer by Blender[1] was chosen. The possibility of automating the rendering pipeline with Python scripts poses an additional advantage for Blender. For the conversion of the DEM into a 3D mesh, the BlenderGIS[2] plugin was used. This plugin already implements functionalities such as GeoTIFF file parsing, mesh generation, and the correct scaling of the mesh.

[1] https://www.blender.org
[2] https://github.com/domlysz/BlenderGIS

The general-purpose Principled BSDF shader[3] by Blender was used to render the material of the terrain. The final shadow masks should encode the shadow intensity, with 0 = completely shadowed and 1 = no shadow. Therefore, the color of the shader was set to pure white (Hex: #FFF). The other parameters of the shader were left with their default values.

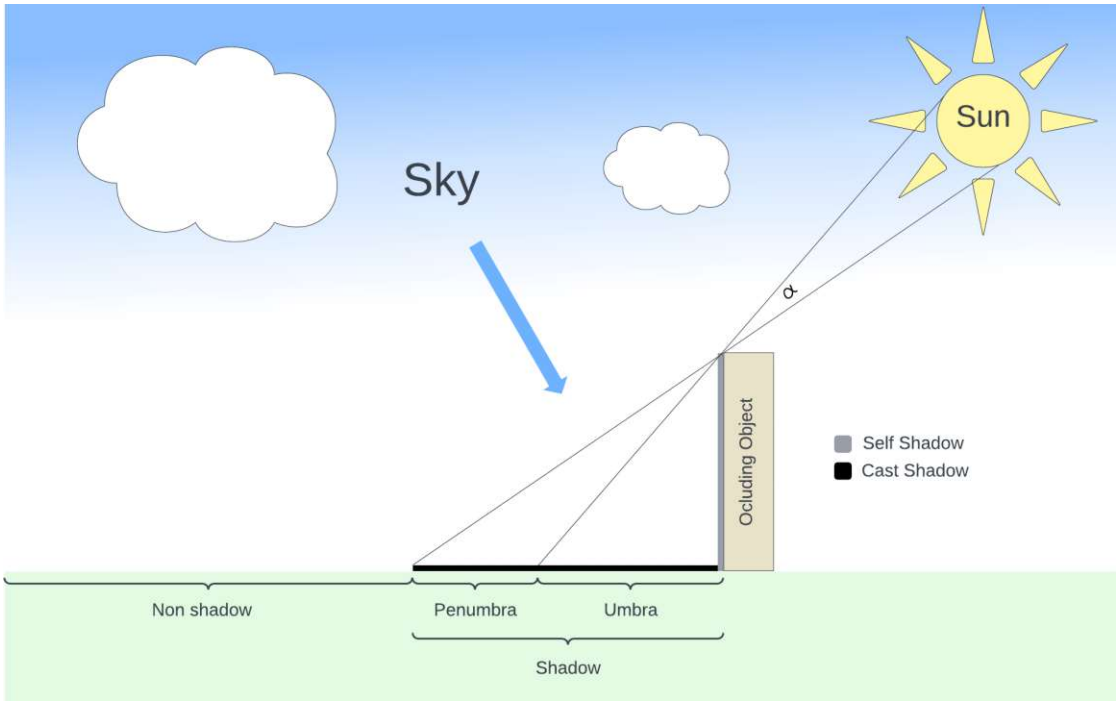
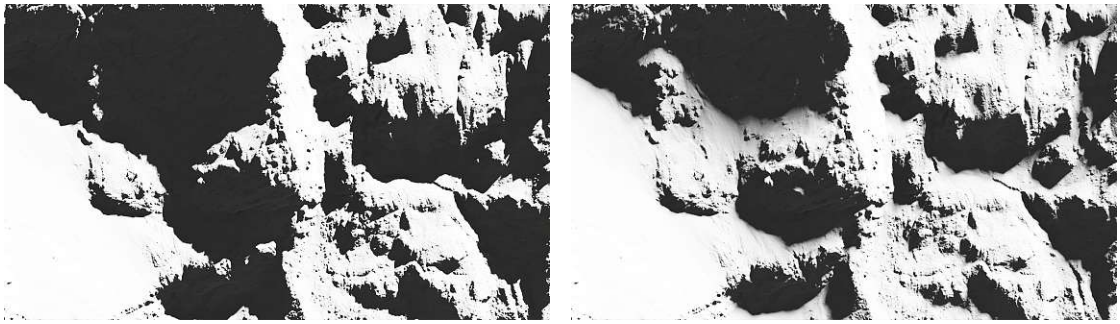


Figure 3.4: This diagram shows different types of shadows and how the angle α corresponds to the size of the penumbra region.

For the lighting of the scene, a single directional light source (Sunlight in Blender) and no ambient light were employed to simulate the sun. Depending on the clarity of the atmosphere, sun rays get refracted at varying strengths, causing softer or harsher shadows (larger or smaller penumbra regions). To replicate this effect, the angle parameter α of the sunlight was randomly sampled from a uniform distribution $\alpha \sim U(0.5, 10)$. This parameter describes the angular diameter, which is the size of a sphere or circle from a given view point. $\alpha = 0$ would denote perfectly parallel light rays and a non-existent transition (penumbra) region. Figure 3.4 visualizes the relation between angle α and the transition (penumbra) region between the shadow and non-shadow areas and depicts different types of shadows. The blue arrow denotes light reflected from the sky onto the shadow area, which leads to the often blue tint of shadows.
The white texture used for the terrain causes a lot of light reflections, which brighten areas inside the shadows unnaturally; thus, the number of ray bounces was reduced to

[3]https://docs.blender.org/manual/en/latest/render/shader_nodes/shader/principled.html

zero. This means that the shadow masks only represent areas with directly occluded sunlight.



(a) Hard shadow $\alpha = 0.5$. (b) Soft shadow $\alpha = 10$.

Figure 3.5: Rendered shadow masks using the DTM with different angle parameters (angular diameter) α determining the size of the penumbra region.

To increase data diversity, we rendered three randomized shadows for each image. Varying sun rotations and angles were sampled from uniform distributions, resulting in shadows of differing sizes and orientations. The rotation angles, denoted as x_{rot} and y_{rot}, were randomly sampled in degrees from the intervals $U(-45, -25)$ and $U(-45, -25) \cup U(25, 45)$, respectively. These rotation ranges reflect sun positions closer to the horizon, producing larger shadows on the terrain. The rendered shadow masks were stored as 16-bit gray-scale PNGs.

3.3.2 Shadowing Algorithm

For the actual shadowing of the image, the shadow mask $S = \{x \in \mathbb{R} : [0, 1]\}$ was taken together with the shadow-free input image $I = \{x \in \mathbb{R}^3 : [0, 1]\}$ to produce the shadowed version $I_S = \{x \in \mathbb{R}^3 : [0, 1]\}$. I will be multiplied by S to darken the shadow areas. Therefore, the shadow mask S is clipped with function c to prevent completely black areas from appearing in the image. Completely black areas are not realistic because there is always some light hitting the camera sensor. Furthermore, multiplying by zero would completely erase all information from this area, making it impossible for the model to reconstruct the original signal. The clipping threshold of 0.2 was determined by trying different values and visually comparing real shadows to generated ones. The shadow mask was inverted to represent higher shadow intensity with a higher value. This allows adaptive noise to be added in the next step.

$$c(x) = \begin{cases} x, & \text{if } x \geq 0.2, \\ 0.2, & \text{otherwise}, \end{cases} \tag{3.1}$$

$$S^{-1} = 1 - c(S). \tag{3.2}$$

The truncated normal distribution $\mathcal{N}_{trunc}(\mu, \sigma, a, b)$ is used to sample Gaussian noise, which is the most common noise occurring in images and other signals [Bon09]. As usual, μ denotes the mean and σ the standard deviation of the Gaussian distribution, whereas a and b describe the lower and upper truncation points. The noise is added to the mask depending on the shadow intensity S^{-1} at this point. This models the transition in penumbra regions more smoothly.

$$S^{-1}_{noise} = S^{-1} + S^{-1} * \nu \qquad \nu \sim \mathcal{N}_{trunc}(0, 0.05, -0.2, 0.2) \tag{3.3}$$

The noisy shadow mask is then filtered with a 3×3 median filter $median_{3x3}$ to lower the high frequency of the noise. The filtering and the following channel shifting steps were adopted from the shadow generation approach of Morales et al. [MHT19].

$$S^{-1}_f = median_{3\times3}(S^{-1}_{noise}) \tag{3.4}$$

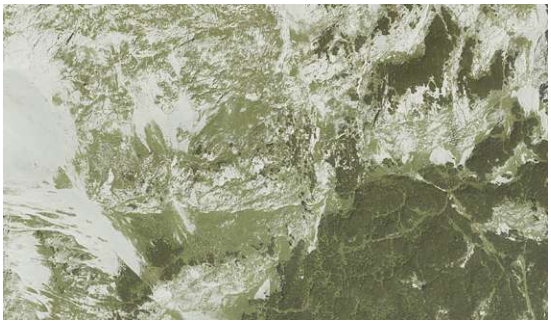
The inverted shadow mask S^{-1}_f will now be randomly shifted three times for each channel of the image and inverted back. This is done to simulate varying shadow intensities for different colors [MHT19]:

$$S_R = 1 - S^{-1}_f * u, \qquad u \sim U(-0.02, 0.02), \tag{3.5}$$
$$S_G = 1 - S^{-1}_f * v, \qquad v \sim U(-0.02, 0.02), \tag{3.6}$$
$$S_B = 1 - S^{-1}_f * w. \qquad w \sim U(-0.02, 0.02). \tag{3.7}$$

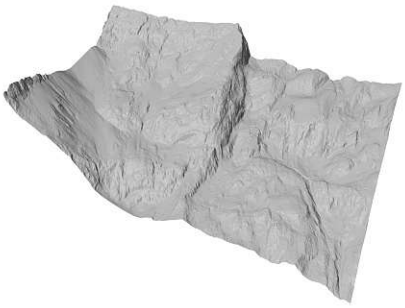
The three shadow masks will now be multiplied with their corresponding channels from the input image I to obtain the final shadowed image.

$$\begin{aligned} I'_R &= I_R \circ S_R \\ I'_G &= I_G \circ S_G \\ I'_B &= I_B \circ S_B \end{aligned} \tag{3.8}$$

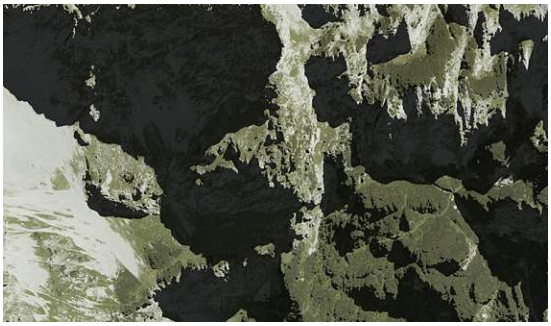
The resulting shadowed images are depicted in Figure 3.6.



(a) Original image



(b) Perspective 3D representation of the terrain



(c) Image with hard virtual shadows $\alpha = 0.5$

(d) Image with soft virtual shadows $\alpha = 10$

Figure 3.6: Original and artificially shadowed images with varying transition region (penumbra) sizes between shadow and non-shadow area.

3.4 Preprocessing

To facilitate feature learning for the elevation data, the DTM has to be added as input for the model. However, there are a few preprocessing steps needed to achieve this. First and foremost, the resolution of the DTM is lower than the orthophotos. Hence, the missing values will be bilinearly interpolated. Furthermore, the very fine-grained 32-bit floating point encoding of the DTM can be min-max scaled over the whole available dataset of Austria and down-sampled into a 16-bit unsigned integer encoding. This encoding allows the elevation data to be added as alpha channel to an PNG image, which makes it much easier to read and write data. A precision of $\sim$ 6cm is retained due to the min-max scaling.

The generated shadowed images have a resolution of at least 1552×898 pixels, but the neural network of the ST-CGAN requires images with a size of 256×256. Therefore, image tiles with 256×256 pixels are cropped out of the larger image. The border region of the image was excluded, because there is no terrain outside the image to cast shadows. Figure 3.7 shows the extracted image tiles from a larger image. The same cropping is also done to the original (ground truth) image and the DTM to produce the final image triplet, which can be seen in Figure 3.8. The elevation data from the DTM is added to the shadowed input image as the alpha channel. The generated *Alpine Shadow Dataset*

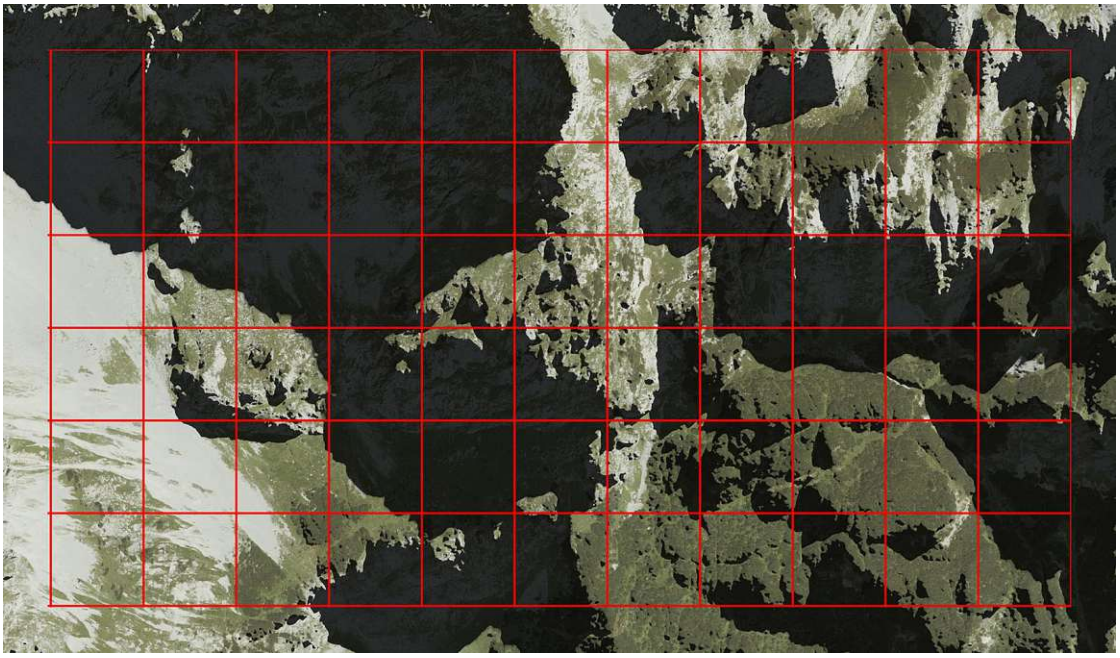


Figure 3.7: Cropped image tiles

consists of 50 images of 50 selected regions. Each image was altered with three different shadow masks with randomly chosen parameters, resulting in 150 images with varying sizes between 1552×898 and 6212×3593 pixels, depending on the size and level of detail of the selected area. Finally, the smaller 256×256 pixels images were cropped out of these 150 images, resulting in 16,740 image triplets for training and evaluation.

(a) Input (b) Ground truth (c) Shadow mask

Figure 3.8: Generated image triplet example of the *Alpine Shadow Dataset* consisting of an artificially shadowed input image (a) with an alpha channel containing elevation data (not visible), an original ground truth image (b) and a shadow mask (c).

3.5 Baseline Dataset with Random Shadow Shapes

To evaluate if the newly generated datasets has benefits for the models' generalizability over a dataset with random shadow shapes, a baseline dataset is also created. This baseline dataset is built on the exact same data as the ASD, but instead of using a rendering engine to create realistic shadow shapes, Perlin noise is used. The sampled noise together with a threshold produces blob-like shadow masks, similar to the process of Morales et al. [MHT19]. These shadow masks are then processed with the same proposed shadowing algorithm and preprocessing steps. Figure 3.9 depicts what these randomly darkened images look like.

(a) Input (b) Ground truth (c) Shadow mask

Figure 3.9: Generated image triplet with random shadow shape consisting of an artificially shadowed input image (a), an original ground truth image (b) and a shadow mask (c).

CHAPTER 4

Shadow-Removal Network

4.1 Network Architecture

The used shadow-removal network for the experiment ST-CGAN consists of generators G_1, G_2 and discriminators D_1 and D_2 which are responsible for shadow detection and removal. Both generators are implemented using an encoder-decoder-style network architecture, and both discriminators use the same convolutional network architecture.

The generators G_1 and G_2 consist of 8 encoding layers, 8 decoding layers, and a bottleneck layer in a U-Net architecture with skip connections from the encoding layer to the corresponding decoding layer (see Figure 4.1).

Every encoding layer, except for L0, performs a leaky ReLU operation with a negative slope of 0.2 before the convolution and a batch normalization afterwards. The applied 2D convolution has a kernel size of $4 times 4$ and a stride of 2 to downsample the image by half in each layer. The dimension of the input is reduced from $256 times 256$ at layer L0 to $1 times 1$ in the bottleneck layer. The decoder side of the U-net upscales the signal again with a transposed convolution analog to the encoder. In contrast to the encoder, the decoder applies a ReLU. The number of feature channels doubles every layer going downwards until L3, where a maximum of 512 feature channels is reached.

During training of the ST-CGAN, the discriminators D_1 and D_2 try to distinguish between real and fake images produced by G_1 and G_2 respectively. D_1 receives either the ground truth shadow mask or the generated shadow mask, concatenated with the input image. Similarly, D_2 receives either the ground truth image or the generated shadow-free image concatenated with the input image. Figure 4.2 shows a detailed overview of the training framework and how the data is passed through the networks during training.

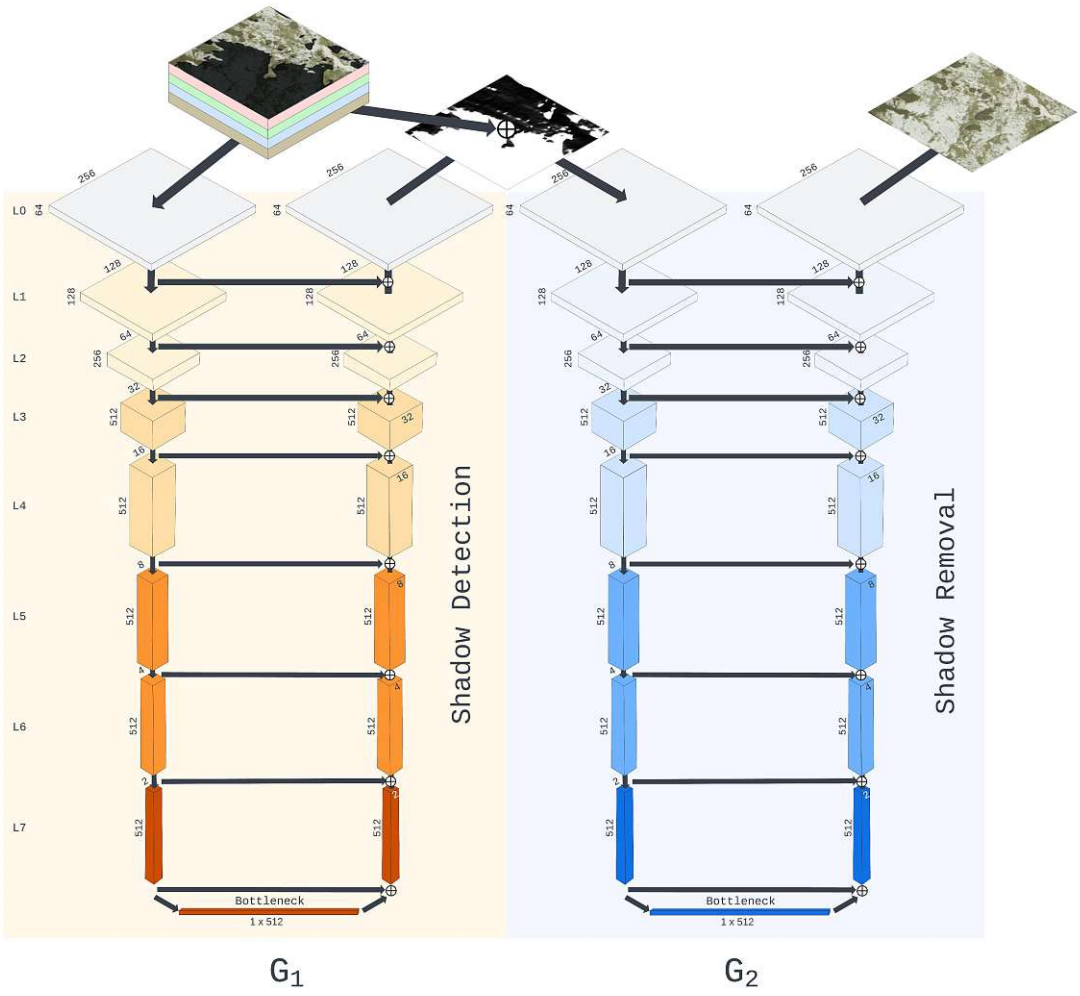


Figure 4.1: Layer visualization of the generators G_1 and G_2 in the ST-CGAN. The shadowed input image gets passed to G_1 and G_2. G_2 receives the input image together with the predicted shadow mask from G_1 to produce the shadow-free output image. The input image consists of the RGB channels and the elevation data as a fourth channel.

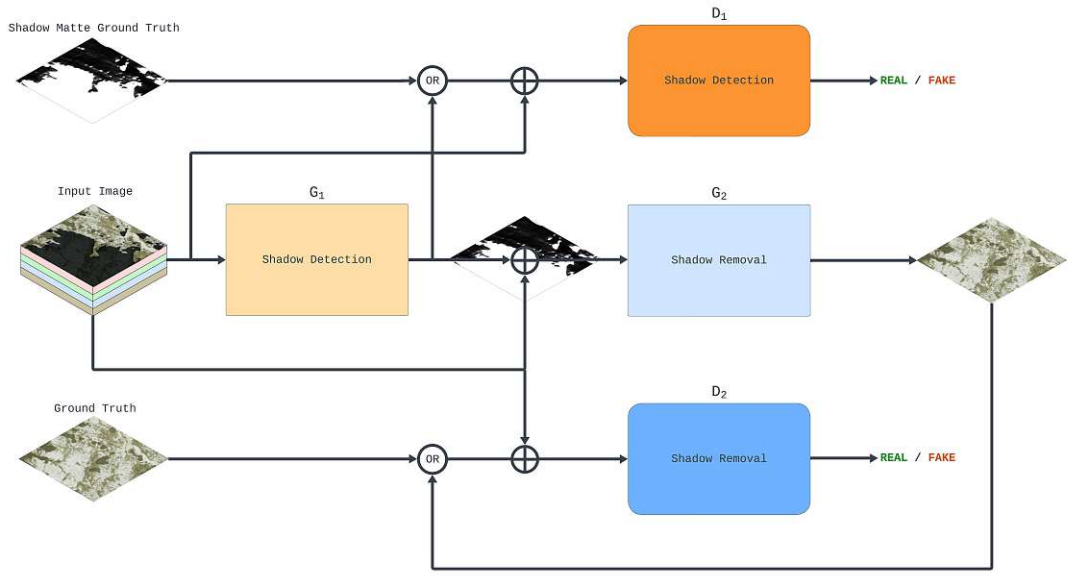


Figure 4.2: GAN style training framework for the ST-CGAN.

4.2 Implementation

The ST-CGAN was implemented using Python 3.9 and the PyTorch framework. An in-official implementation from Github[1] was used as the basis to mitigate implementation workload and errors. The available version had to be adapted in order to read and process images with 16-bit depth per channel instead of the usual 8-bit depth and to read the elevation data from the fourth channel. The used Python framework Pillow[2] used to read images does not allow the parsing of 16-bit channels and defaults to 8-bit implicitly. Therefore, the reading operations were replaced with the ones from the OpenCV[3] library. Here, it is important to pass the `IMREAD_UNCHANGED` flag to the `imread()` function, otherwise it also parses the image in 8-bit. The employed augmentation methods like rotations and flips were also unfit to handle the 16-bit data and have been completely replaced with analog methods from the Albumentations[4] library. Before the image and elevation data are sent to the models, the data is normalized and cast to 32-bit float values. The number of network input and output channels also had to be increased to accompany the additional elevation information.

[1] https://github.com/IsHYuhi/ST-CGAN_Stacked_Conditional_Generative_Adversarial_Networks
[2] https://python-pillow.org
[3] https://opencv.org
[4] https://albumentations.ai

CHAPTER 5

Experiment

For the experiment, two ST-CGANs were trained with the generated *Alpine Shadow Dataset*. The first network received the images together with the elevation data encoded in the alpha channel of the RGBA image. The other network only received the RGB data. This was done to measure performance differences between the two models and answer **RQ2**. It was assumed that the additional elevation data could help the model to remove shadows due to the direct correlation between terrain topology and shadows. The second goal of the experiment is to determine if the models are able to generalize to real shadows and if there are observable differences between the models. Additionally, a comparison with a baseline dataset comprising randomly shaped shadows was conducted to assess differences in generalizability. This elucidates the validity of the proposed data-generation method. The hyperparameters for training runs are held constant throughout all experiments to maintain consistency.

5.1 Training

The *Alpine Shadow Dataset* was split into train, validation, and test sets in a ratio of $(0.8, 0.1, 0.1)$, which results in 13392 training, 1674 validation, and 1674 test images. The dataset contains three different shadow versions of the same image depending on the sampled orientation of the sun and the angular diameter α. To make sure the provided image tiles of the test and validation set are really unseen, the data was split between images and not between different shadow versions.

The models were trained for 150 epochs with a batch size of 32 and a learning rate $lr = 0.0002$. The Adam solver was used as an algorithm for the gradient descent with the provided betas of $(0.5, 0.999)$. The λ-parameters defined in loss Equation 2.11 were taken from the original approach of Wang et al. [WLY18] and set to $\lambda_1 = 5$, $\lambda_2 = 0.1$, $\lambda_3 = 0.1$. The training took 16 hours for the model without elevation data and 19 hours for the

	lr	λ_1	λ_2	λ_3	beta1	beta2
Value	0.0002	5	0.1	0.1	0.5	0.999

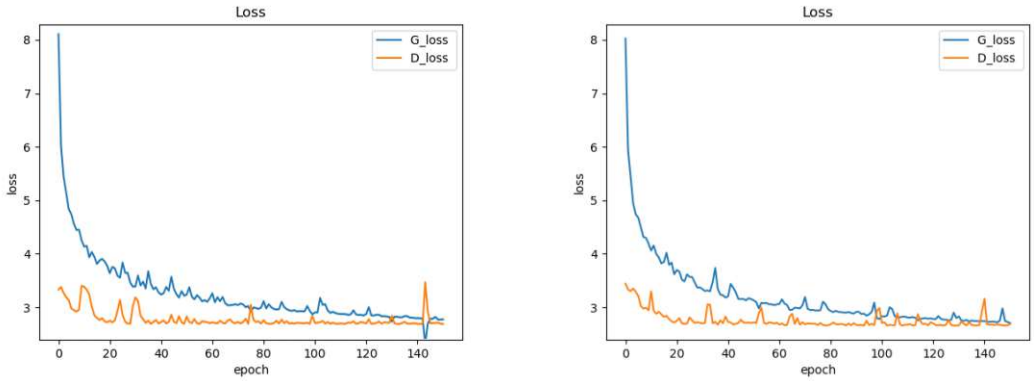
Table 5.1: Overview of the training hyperparameters of the ST-CGAN. The learning rate (lr) is used together with beta1 and beta2 as parameters for the Adam solver. The λ values describe the weighing of the individual parts of the loss function defined in Equation 2.11 The weights and learning rate were taken from the approach of Wang et al. [WLY18].

Model	G_1	G_2	D_1	D_2
Without Elevation	29.239M	29.245M	2.766M	2.769M
With Elevation	29.240M	29.246M	2.767M	2.770M

Table 5.2: Number of trainable parameters for each network.

model with elevation data on a desktop PC with an Nvidia RTX-3060 graphics card and an Intel Xeon-E3v3 processor. The number of trainable parameters for each network can be seen in Table 5.2. Table 5.1 shows an overview of the used hyperparameters.

During training, the loss from Equation 2.11 (*G_loss*) and the loss only comprised of the discriminator output (D_loss) were computed. As Figure 5.1 shows, these metrics for the generator and discriminator decreased well over time, suggesting a successful training algorithm (see Figure 5.1). After approximately 100 epochs, the loss curves flatten out, exhibiting the convergence of the models.



(a) without elevation data (b) with elevation data

Figure 5.1: Learning curves of the two trained models, without and with elevation data, over the 150 training epochs. The D_loss (orange) describes the discriminator loss and was calculated using binary cross entropy. The G_loss denotes the generator loss (blue), which was calculated using Equation 2.11.

5.2 Performance Measures

To make the performances quantitatively comparable, the root-mean-square error (RMSE) and the peak signal-to-noise ratio (PSNR) were used to calculate differences between the model output and the ground truth. The predicted output denoted as $\hat{y}$ and the ground truth denoted as y were used to calculate the RMSE and the PSNR with the following equations:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=0}^{n}(\hat{y}_i - y_i)^2}, \tag{5.1}$$

$$PSNR = 20 \cdot \log_{10}\left(\frac{2^b - 1}{\frac{1}{n}\sum_{i=0}^{n}(\hat{y}_i - y_i)^2}\right). \tag{5.2}$$

The parameter b of PSNR denotes the bit depth of the observed images and is used to calculate the maximal possible value with $2^b - 1$. For this experiment, the evaluated images were stored with 8 bits, resulting in a maximum value of 255. Both metrics are widely used to benchmark shadow-removal performance. In the subsequent analysis, RMSE was primarily used for comparison because RMSE is used in more referenced papers [LS19, AI22, QTH+17, WLY18] and has the advantage that it is defined for identical outcomes. In this case, the denominator of the inner fraction of the PSNR would equal zero. On the other hand, the PSNR metric is normalized, and comparison between differently scaled data is possible. We included the PSNR for this reason. However, similar to the mentioned papers, which use the RMSE, we transform the images to LAB space, so comparison between these other approaches is possible anyway. The two metrics were calculated over the test dataset within shadow areas and over the whole image. The metrics within shadow areas facilitate a more focused look at the performance, regardless of the size of the shadow region. The metrics over the full image additionally measure if the models alter regions outside the shadow area and the performance overall. The performance of the shadow mask prediction is measured with the Intersection over Union (IOU) or Jaccard index. If we denote the binarized ground truth shadow mask with A and the binarized predicted shadow mask with B, the IOU is calculated with this equation:

$$IOU = \frac{|A \cap B|}{|A \cup B|}. \tag{5.3}$$

The scores were calculated for each image individually. Afterward, the average of the RMSE, PSNR and IOU was calculated over all 1674 test images to produce the final aggregated scores. To measure the statistical significance of the metrics, the Wilcoxon signed-rank test was used, analogous to the comparative GAN study by Lv et al. [LZY21].

Performance evaluation on real shadow data can only be done by visual inspection of the predicted output, because there is no ground truth data as the basis for calculating the

metrics. A “real vs. fake” Mechanical Turk study could be conducted on the deshadowed images, similar to the studies done by Zhu et al. [ZPIE17] and Isola et al. [IZZE17]. However, this only makes sense if the predicted images are very close to real ones, which is not the case for our models.

CHAPTER 6

Results

This chapter presents the findings of the conducted experiments, commencing with an evaluation of model performance on artificial shadow data from the *Alpine Shadow Dataset*. Subsequently, the model's ability to generalize to real shadow data is assessed. This assessment is followed by a t-SNE analysis of the layers to elucidate the subpar real shadow performance. The t-SNE analysis reveals a deliberate discriminatory behavior between real and artificial shadows, which might be due to the remaining real shadows in the training data and the chosen GAN training framework. This suspicion was further underlined by training models with a downsized version of the *Alpine Shadow Dataset* to a lower level of detail, which mitigates the effect of remaining small-scale shadows. A comparison against the baseline dataset with random shadow shapes exhibited improved generalizability using the ASD. This comparison was conducted with models trained on downsized images due to the insufficient performance of the models trained on full-scale images.

6.1 Virtual Shadow Data

Metric	Mean	Variance
RMSE	3.308	28.957
RMSE Shadow	16.051	74.235
PSNR	40.841	38.193
PSNR Shadow	38.176	33.192

Table 6.1: RMSE and PSNR between the input and the ground truth averaged over the test dataset. Larger RMSE and lower PSNR scores signify larger differences.

Table 6.1 displays the baseline difference between the shadowed input images and the ground truth of the ASD computed with the RMSE and PSNR metrics. All metrics

observe very high variance, which is explained by the presence of image instances without shadows.

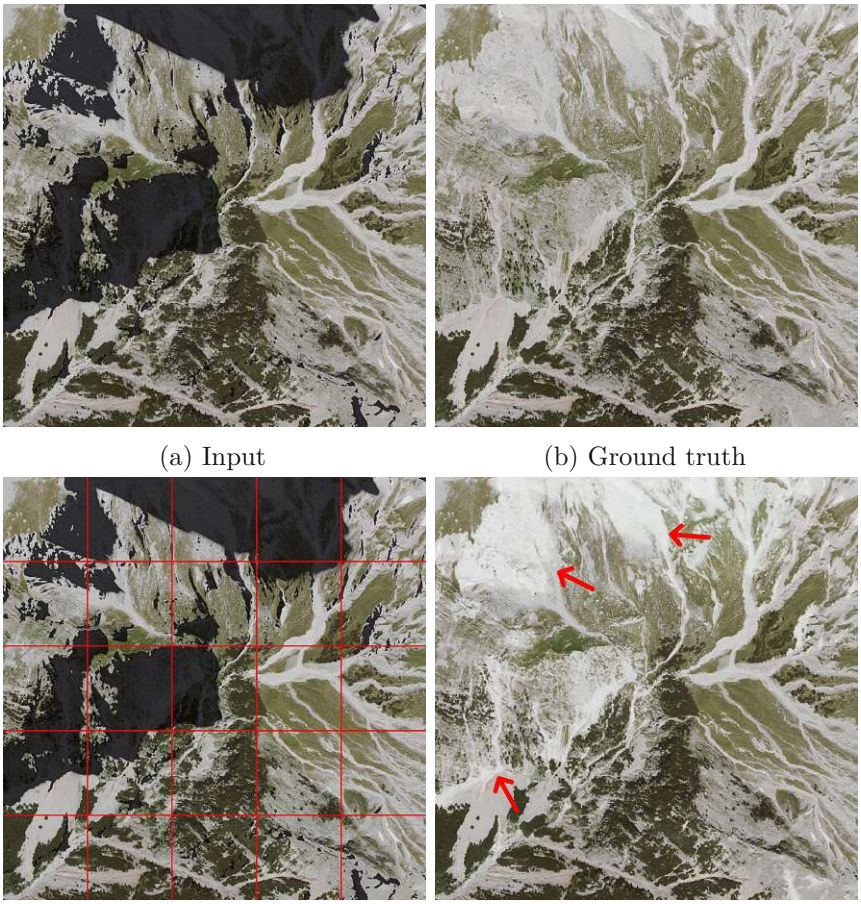
	No Elev.		Elev.		Difference		p-value	|r|
Metric	**mean**	**var**	**mean**	**var**	**mean**	**var**		
RMSE	0.723	0.199	0.693	0.229	-0.03	+0.03	0.0001	0.42
RMSE Sh.	3.387	3.398	3.414	3.721	+0.334	+0.323	0.0547	0.06
PSNR	41.928	24.739	42.771	29.645	+0.843	+4.906	0.0001	0.54
PSNR Sh.	30.192	2.751	30.256	2.711	+0.064	+0.04	0.0524	0.15
IOU	0.744	0.100	0.722	0.109	-0.022	+0.009	0.0487	0.04

Table 6.2: Overview of the computed metrics on the test set sampled from the ASD. Lower RMSE and higher PSNR and IOU scores signify better performance. In the difference column, green means that the model trained with elevation data improved over the one without, and red means the opposite. The p-values were calculated using the Wilcoxon signed-rank test and rounded to four decimal places. All other scores were rounded to three decimal places. Column "|r|" denotes the effect size of the observed differences.

Table 6.2 shows the performance of the model with elevation data $M_{elevation}$ and the model without elevation data M_{RGB} on the test dataset sampled from the ASD. The RMSE and PSNR scores suggest very good shadow-removal performance with artificial shadows. The lower the RMSE score, the better, in contrast to the PSNR, where a higher value is favorable. The RMSE score shows a statistically significant ($p < 0.05$) medium ($0.3 < |r| < 0.5$) sized improvement with elevation data. The same can be observed with the PSNR, where the effective difference is even large ($0.5 < |r|$).

The variance of the metrics is higher for $M_{elevation}$ most likely due to the higher number of trainable parameters (see Table 5.2). The PSNR Shadow and RMSE Shadow metrics show contrary results, even though they are closely related. This can be explained by the transformation to LAB color space only done for the RMSE scores. Both shadow metrics slightly exceed the $p < 0.05$ threshold for statistical significance, and their effective size is very small ($|r| < 0.1$) for RMSE and small ($0.1 < |r| < 0.3$) for PSNR. Considering the shadow mask prediction performance, the IOU score exhibits a very small ($|r| < 0.1$) decrease in performance for $M_{elevation}$ which is statistically significant by a small margin.

The success of artificial shadow-removal can also be seen in the example from Figure 6.1. The output was taken from M_{RGB} and is almost identical to $M_{elevation}$. Therefore, it was excluded from the figure to make comparisons easier. The example was extracted independently of the other test data because the smaller images tiles of ASD were split up randomly into the train, test, and validation sets. Thus, combining multiple tiles into a larger image was not possible with these images. Although this example was selected geographically far away from the ASD, the models were still able to reconstruct the shadow-free version quite well. However, some imperfections are visible, like the overly bright reconstruction of certain shadowed scree and rock terrain. Another positive aspect



(a) Input (b) Ground truth

(c) Image tiles (d) Output

Figure 6.1: This example shows how the models are able to successfully remove artificial shadows from the input image. The example taken from M_{RGB} comprises 5×5 image tiles with 256×256 pixels each seen in (c). The only visible differences are the overly bright reconstructed rock and scree areas indicated by the red arrows.

that can be observed is that although the images consist of 5×5 smaller image tiles, the transitions between those tiles are barely visible.

In Figures 6.2 and 6.3 the individual RMSE scores for each image are visualized as box plot diagrams. Here, a close resemblance between the two models can be examined. One visually observable distinction within the box plots lies in the presence of more pronounced shadow RMSE outliers of $M_{elevation}$ which is also reflected in the higher variance of the model.

For qualitative result inspection, Figure 6.4 shows the most extreme RMSE outliers from the test dataset. Here, we can see that even though these are the worst-performing images, they are visually not far away from the ground truth. The most common effect

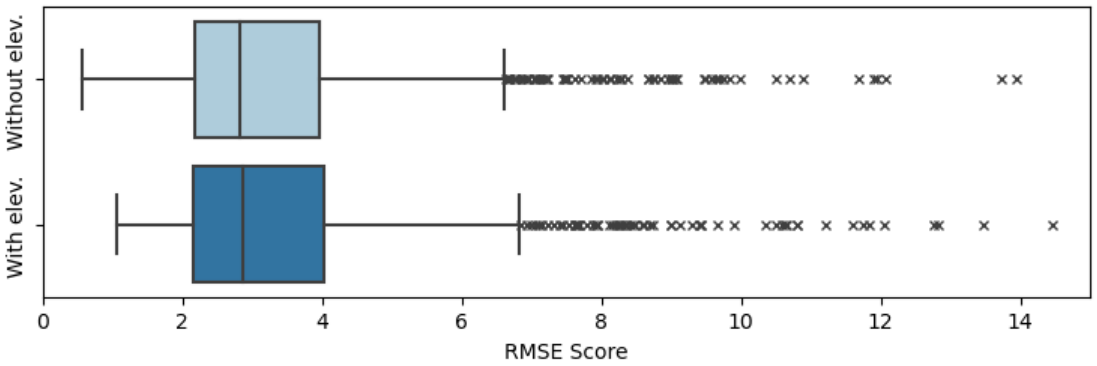


Figure 6.2: Juxtaposition of shadow RMSE scores between the two models visualized as box plots.

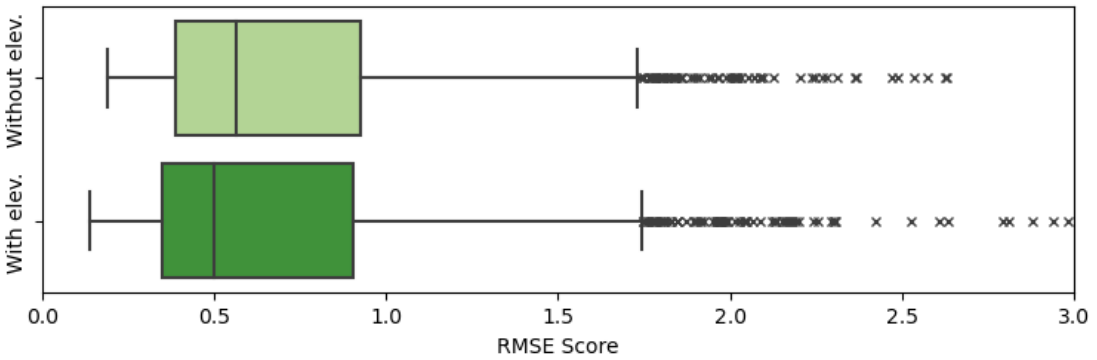


Figure 6.3: Juxtaposition of RMSE scores between the two models visualized as box plots.

is decoloring like in Figure 6.4 (a) or blurring, as seen in Figure 6.4 (c). Additionally, hallucination effects can be observed as well, like in Figure 6.4 (d), where the larger bush on the right side was replaced with smaller bushes with grass in between. Another interesting aspect of this observation is that both models seem to struggle with the same images (Figure 6.4 (a), (b), (d), and Figure 6.5 (a), (b), (f)). This indicates that these are especially difficult ones. The outliers from $M_{elevation}$ are also very similar, so much so that even the hallucination effect of Figure 6.4 (d) is almost identical to the effect in Figure 6.5 (b).

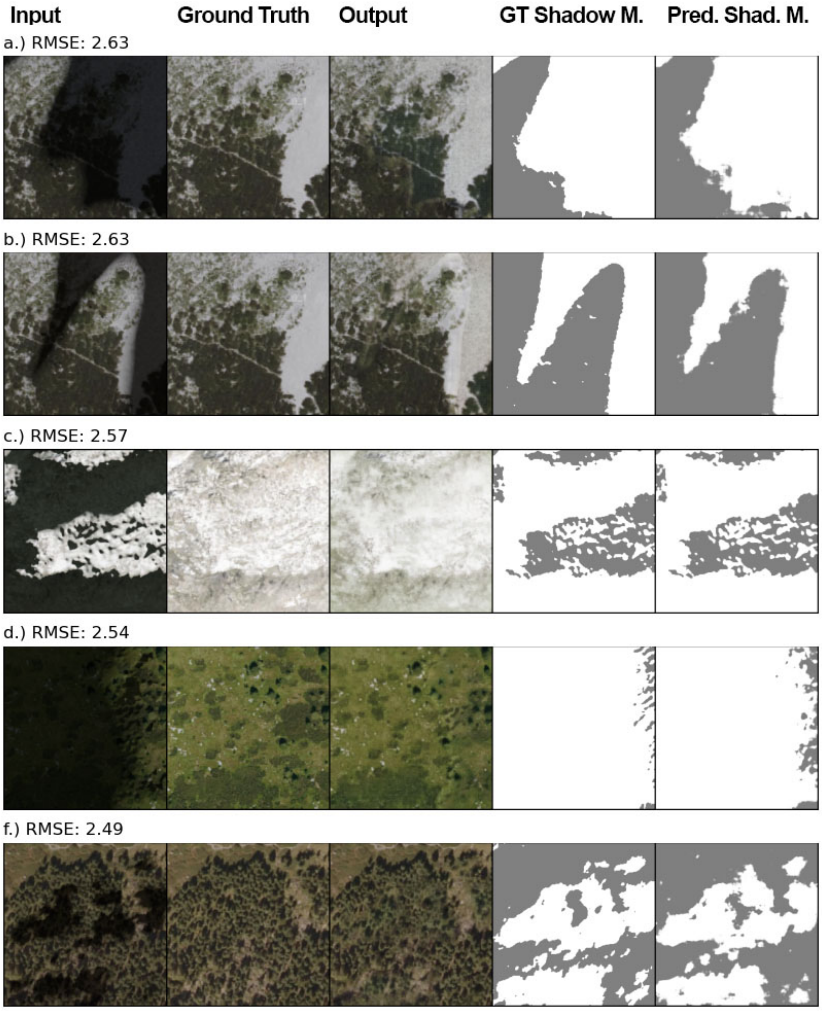


Figure 6.4: Top five outliers with the largest RMSE value from model M_{RGB} trained without elevation data. The images depict input, ground truth, output, ground truth shadow mask, and shadow mask prediction from left to right.

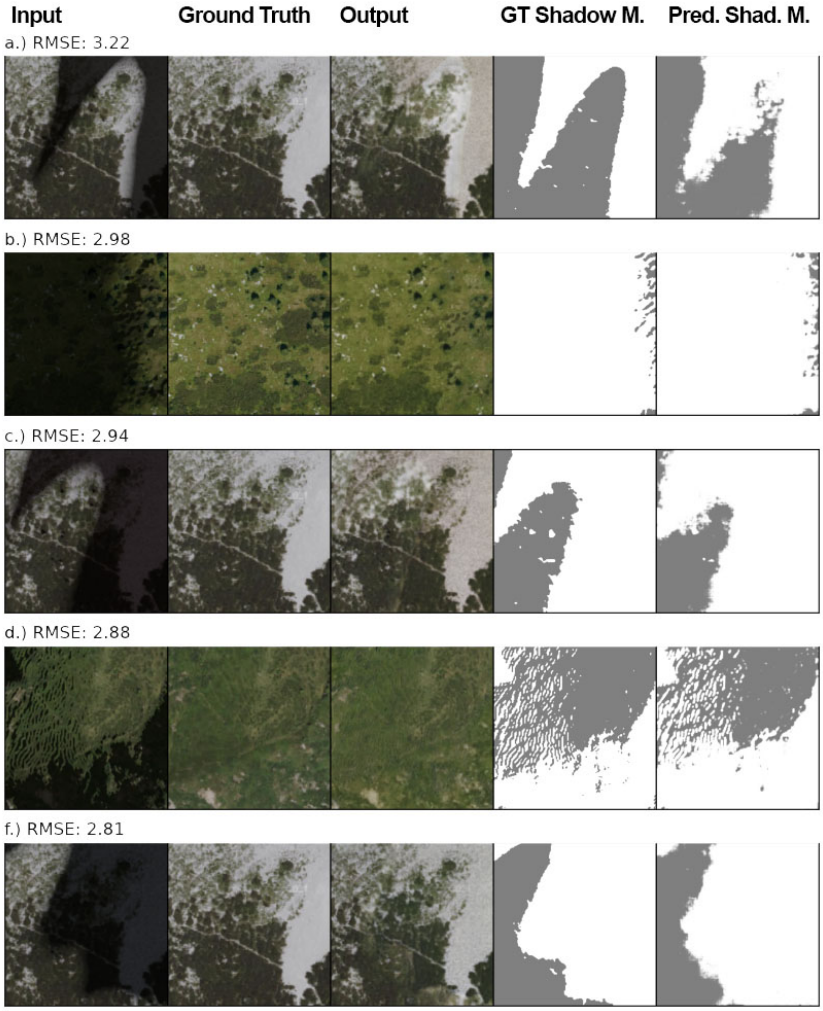
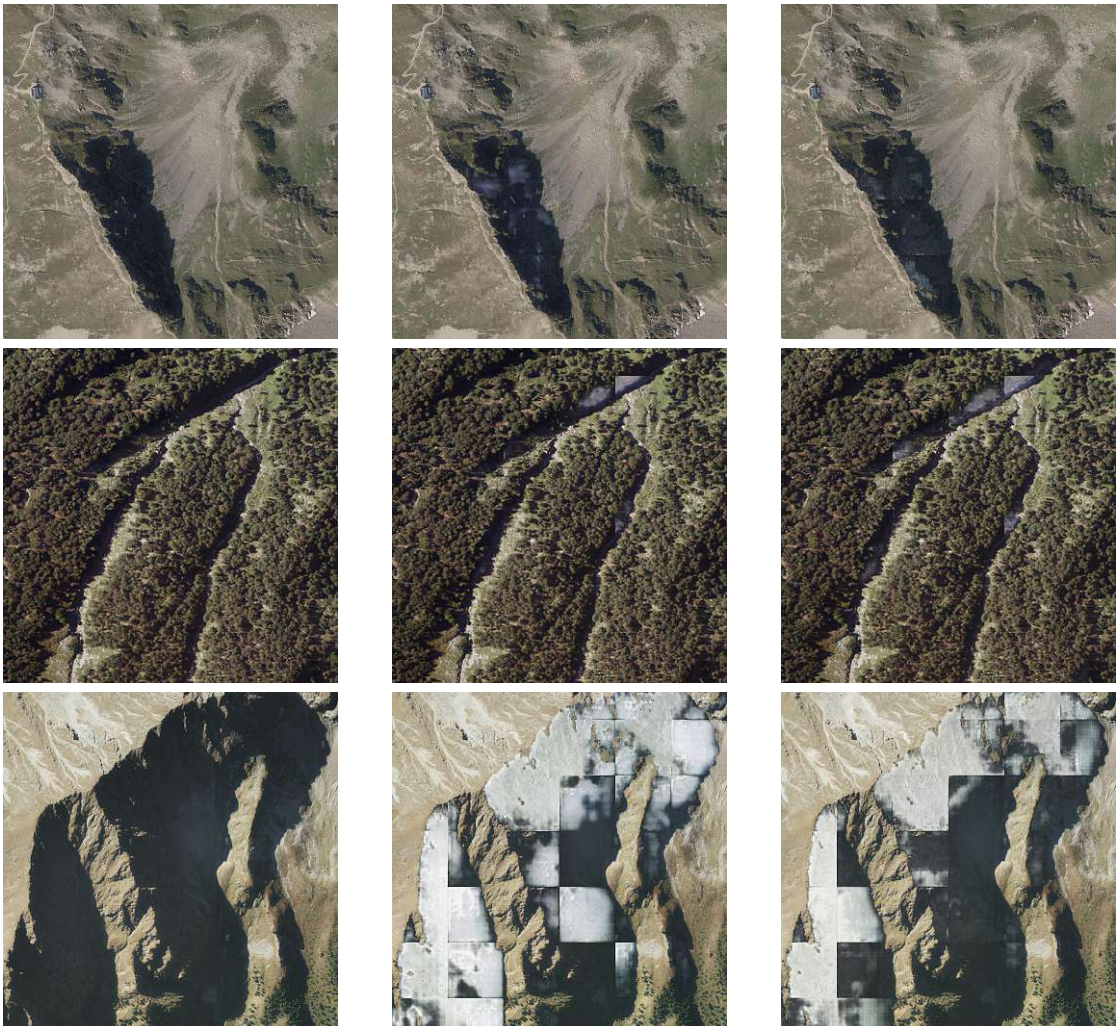


Figure 6.5: Top five outliers with the largest RMSE value from model $M_{elevation}$ trained with elevation data. The images depict input, ground truth, output, ground truth shadow mask, and shadow mask prediction from left to right.

6.2 Real Shadow Data

While the presented results on artificial shadows look promising, introducing real shadows to the models yields substantially different outputs.



(a) input (b) without elevation (c) with elevation

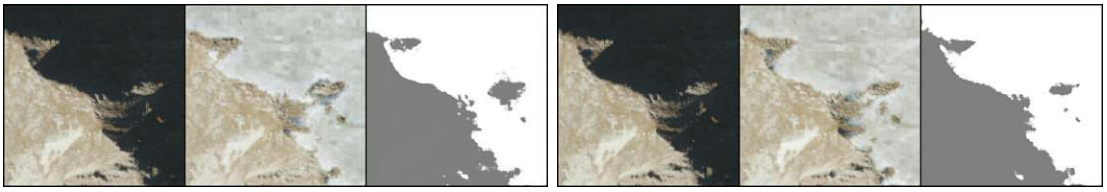
Figure 6.6: Shadow removal examples with real shadows. Each of the images comprises 36 smaller (256×256) image tiles sent to the network individually and stitched together afterwards.

Figure 6.6 illustrates three real shadow examples, each composed of 36 256×256 pixel image tiles. The evaluation was solely done by visual inspection, because there is no ground truth for real data. In the initial two instances, the output images exhibit minimal alterations, with only subtle brightening present in the larger shadow regions. The final

example shows that the detection of shadow areas works to some extent, but the removal process does not work properly. Notably, cloud-like structures consistently emerge across multiple examples, despite the absence of clouds in the training data. This intriguing behavior persisted consistently across all trained models, including preliminary tests.

(a) Without Elevation Data (b) With Elevation Data

Figure 6.7: Examples of both models failing to detect shadows.



(a) Without Elevation Data (b) With Elevation Data

Figure 6.8: Examples of both models successfully detecting shadows.

(a) Without Elevation Data (b) With Elevation Data

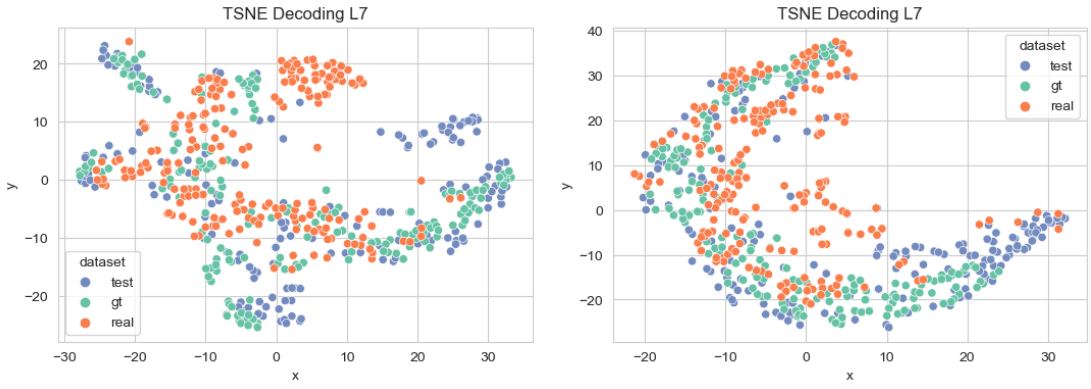
Figure 6.9: Examples depicting that the model with elevation data seems to be less prone to false positives.

The shadow detection performance of both models exhibits noticeable shortcomings when evaluated against real shadow data. This deficiency is illustrated in Figure 6.7, where both models completely disregard the shadow regions. While there are instances, as depicted in Figure 6.8, where the detection performs quite well, such cases are infrequent and appear to be confined to scenarios involving very dark and pronounced shadows. Notably, in certain instances, $M_{elevation}$ demonstrates a reduced tendency for false positives, exemplified in Figure 6.9, where M_{RGB} mistakenly identifies a bush as a shadow, while $M_{elevation}$ does not. Nevertheless, true positive predictions are still an infrequent occurrence in both models.

6.3 Layer Analysis with t-SNE

To elucidate if distribution shifts between virtual shadow, real shadow and ground truth images are contributing factors to the poor real shadow removal performance, an analysis of neural network layer outputs was conducted. In order to visualize the high-dimensional data, the t-distributed stochastic neighbor embedding (t-SNE) was used to reduce the data to two dimensions. Multiple data points are given to the model, stemming from three classes (real shadow, artificial shadow, and ground truth). The layer output of each data point is then projected on a plane and color coded depending on the class label. Clustering of these points would reveal differences in layer outputs across classes and helps to pinpoint if there is/are specific layer(s) where the models acquired discriminative capabilities towards ground truth, real and artificial shadows, suggesting distribution shifts between them.

To conduct this analysis, a subset of artificially shadowed and ground truth images was sampled from the test dataset, alongside a distinct set of real shadow images. The three sets contain 187 samples each, for a total of 561 data points. Subsequently, these datasets were given to the models, and the output of each layer of G_1 and G_2 was extracted and projected with t-SNE.



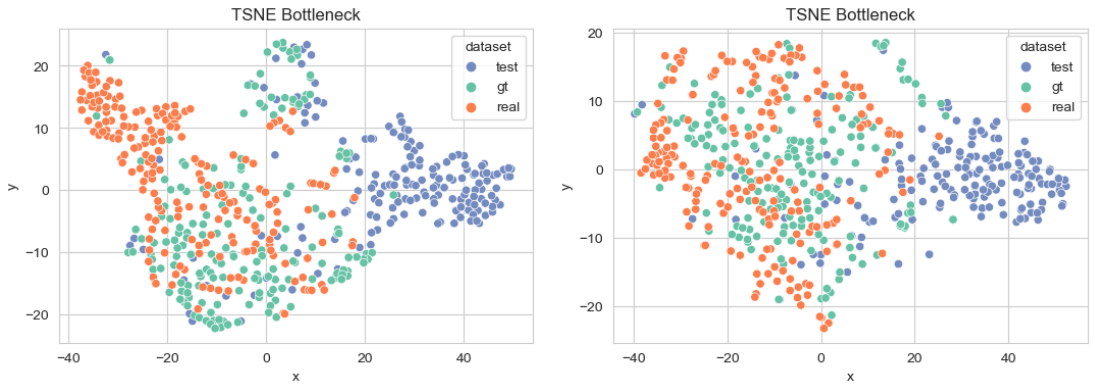
(a) without elevation (b) with elevation

Figure 6.10: T-SNE of the decoding layer L7 output of shadow detection network G_1 with artificial shadow (test, blue) and ground truth (gt, green) images from the test data set sampled from the ASD as well as real shadow images (real, red).

The t-SNE of the sampled sets shows that the shadow detection network G_1 does not distinguish between the three datasets visibly. Figure 6.10 depicts the output of the first decoding layer L7 of the shadow detection network G_1. In this graph, some minor clusters are visible, but the three classes are still more or less evenly distributed over the projected space.

This behavior can also be observed throughout encoding layers L0 to L7 of the shadow-removal network G_2. However, some clustering was visible in the bottleneck and decoding

layer L7 of G_2. Figure 6.11 shows the sample distribution with a considerable distinction between real and artificial shadows. The t-SNE also shows that the features of real and ground truth data are mixed. This indicates that the selected images for training are representative samples.



(a) without elevation (b) with elevation

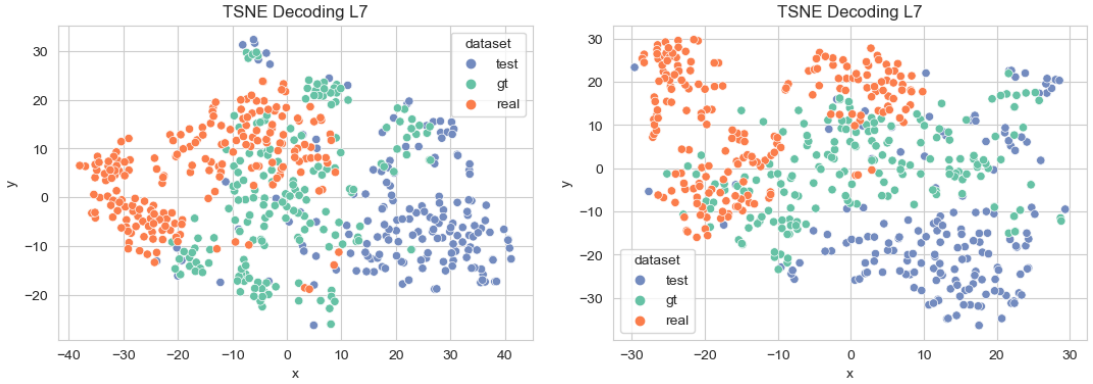
Figure 6.11: T-SNE of the shadow-removal network G_2 bottleneck layer output with artificial shadow (test, blue) and ground truth (gt, green) images from the test data set sampled from the ASD as well as real shadow images (real, red).

The distinction between the classes *real*, *ground truth*, and *test* (artificial shadows) is even more pronounced in decoding layer L7, where the ground truth also starts to separate from the other two classes (see Figure 6.12). The plot shows that the real and artificial shadows are on opposite sites, and the ground truth images are between. This suggests that the model tries to maximize the distance between artificial and real shadows, suggesting discriminative capabilities in these layers.

To determine if the images are discriminated by shadow type and not by another feature present, class centroids for *test, real* and *ground truth* were computed. The centroid of the real shadow class and the centroid of the artificial shadows class can then be used to determine the most distance examples to each other. These examples probably include the discriminating feature, according to the model. The centroids were computed with the average of all class samples, and the distances were determined using the Euclidean distance. Both calculations were done in feature space before the dimension reduction with t-SNE. In Figure 6.13, the furthest and closest examples are visualized.

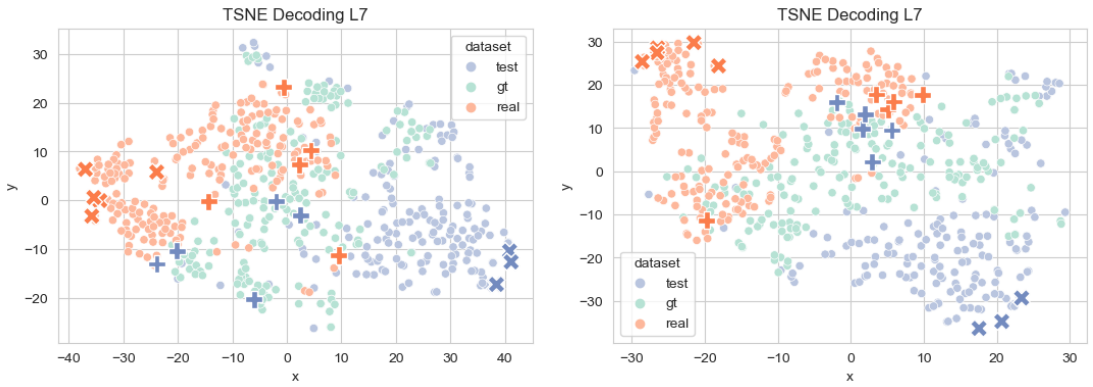
The top five most distant samples from each other's class centroids can be seen in Figure 6.14. Each of the depicted images contains a lot of artificial or real shadows, which underlines the assumption that shadows are indeed a discriminating factor for the models.

The five closest samples to the other class, as seen in Figure 6.15, depict forests and shadow-free areas. This makes sense, because these areas are also contained in the training and test sets. Interestingly, the closest real images to the test set, according to model M_{RGB}, contain rough, rocky terrain. A possible reason could be that rough, rocky



(a) without elevation (b) with elevation

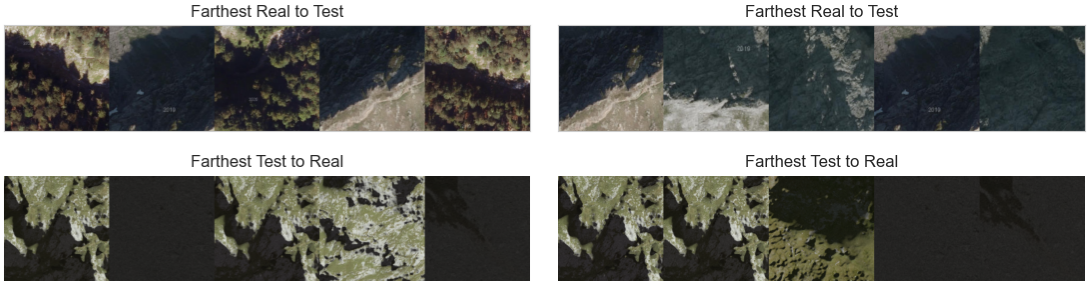
Figure 6.12: T-SNE of the shadow-removal network G_2 layer L7 output of the shadow-removal network G_2 with artificial shadow (test, blue) and ground truth (gt, green) images from the test data set sampled from the ASD as well as real shadow images (real, red).



(a) without elevation (b) with elevation

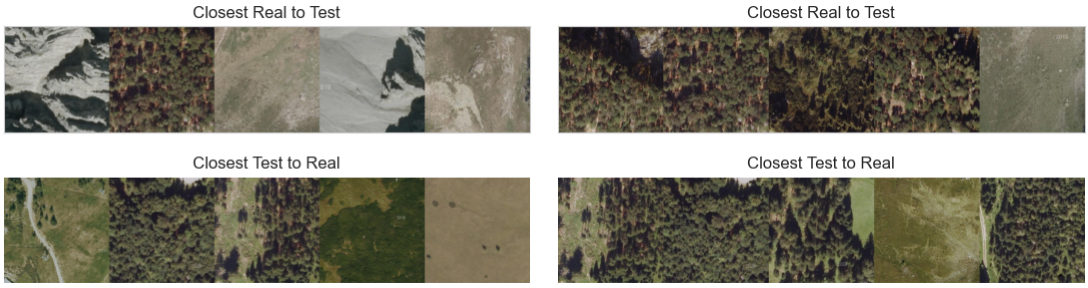
Figure 6.13: Closest and farthest samples from real shadow (real, read) to artificial shadow (test, blue) and vice versa using the Euclidean distance to the class centroids calculated in feature space. The features stem from the L7 layer of the shadow-removal network $G2$. The closest samples are marked with "+" and the farthest with "X".

terrain inevitable contains some shadows, and the test set most likely also contains small shadows within these areas.



(a) without elevation (b) with elevation

Figure 6.14: Most distant samples to the respective other class centroids.



(a) without elevation (b) with elevation

Figure 6.15: Closest samples to the respective other class centroids.

6.4 Training on a Lower Level of Detail

In pursuit of understanding the discriminative behavior of our models concerning real and artificial shadows, observed in the t-SNE analysis (see Section 6.3), two underlying hypotheses were postulated. The first conjecture is that the artificial shadows are not representative. The second assumption regards the persistence of small-scale shadows within the training set serving as a discerning factor for the discriminator, guiding the generator to leave real shadows in the output, which results in the observed discriminative behavior.

To assess which of the aforementioned propositions applies, another experiment was conducted. Larger tiles, measuring 640×640 pixels, were extracted from the ASD and subsequently downscaled to 256×256 pixels. This downsizing to a lower level of detail mitigates the influence of small-scale shadows on the models. The resulting downsized ASD (DASD) only comprises 1,600 images, reduced from the initial 16,000. The retraining of both models was executed over 100 epochs, with all other hyperparameters held constant to maintain consistency with the initial experimental conditions. The epochs are reduced from 150 in the initial training to 100 because the learning curves flatten out after 100 epochs, and there is substantially less training data for this run.

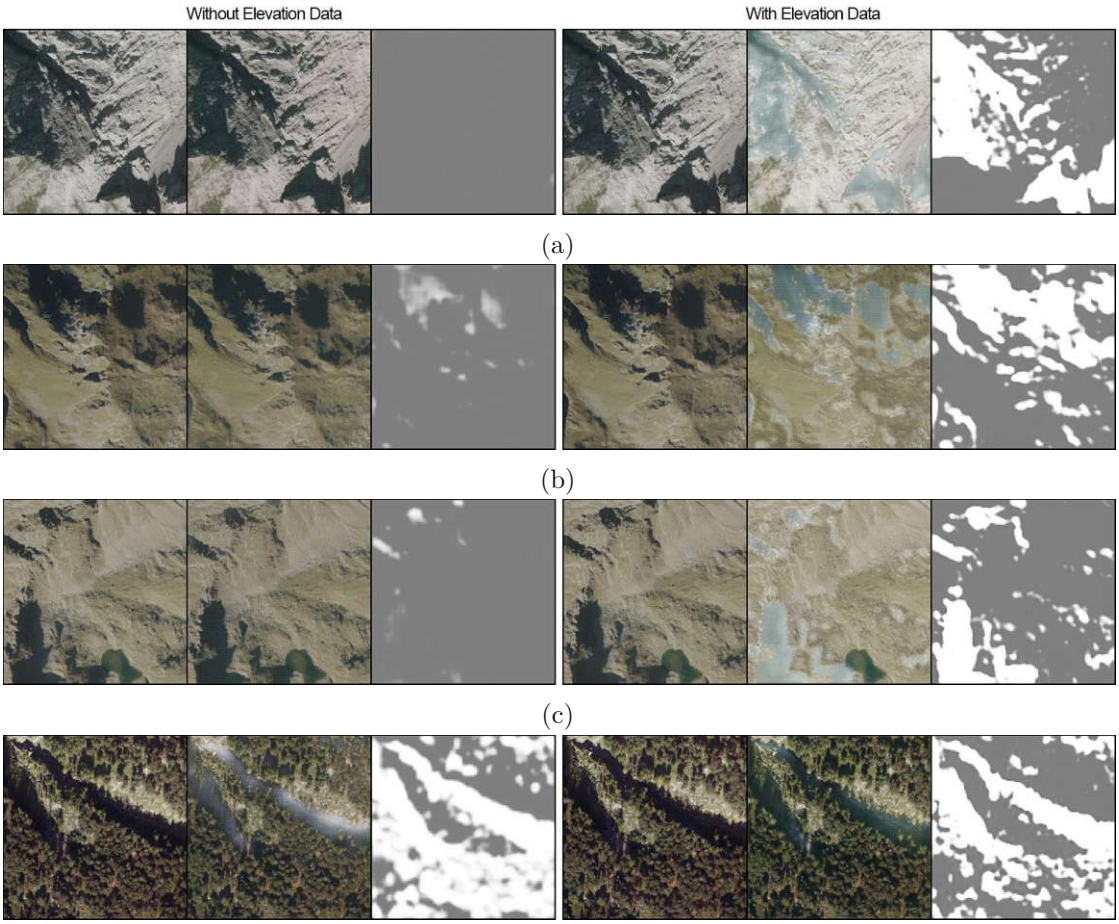


Figure 6.16: Performance comparison with real shadow data. Both models were trained on the downsized version of the ASD. The images depict input, shadow-free prediction and shadow mask prediction from left to right.

Upon visual inspection of the model-generated outputs on real shadow images, a distinct contrast emerged between the two variants. Notably, the model trained with elevation data $M_{elevation}$ outperformed the model without elevation data M_{RGB} in detecting and mitigating real shadows. This can be seen in Figure 6.16 (a), where $M_{elevation}$ accurately identified shadow regions, while M_{RGB} failed to detect any shadows. Similar trends were observed in Figures 6.16 (b) and (c), where M_{RGB} showed suboptimal shadow detection compared to $M_{elevation}$. Impressively, $M_{elevation}$ even correctly classified the pond in Figure 6.16 (c) as a non-shadow area, despite its potential resemblance to a shadow.

The shadow-removal of both models is still not sufficient, mainly due to the limited size of the training set. The training set consists of 1,200 images, with each scene featuring three shadow versions generated with varying parameters, resulting in only 400 unique scenes. However, data can be added with the introduced data generation method.

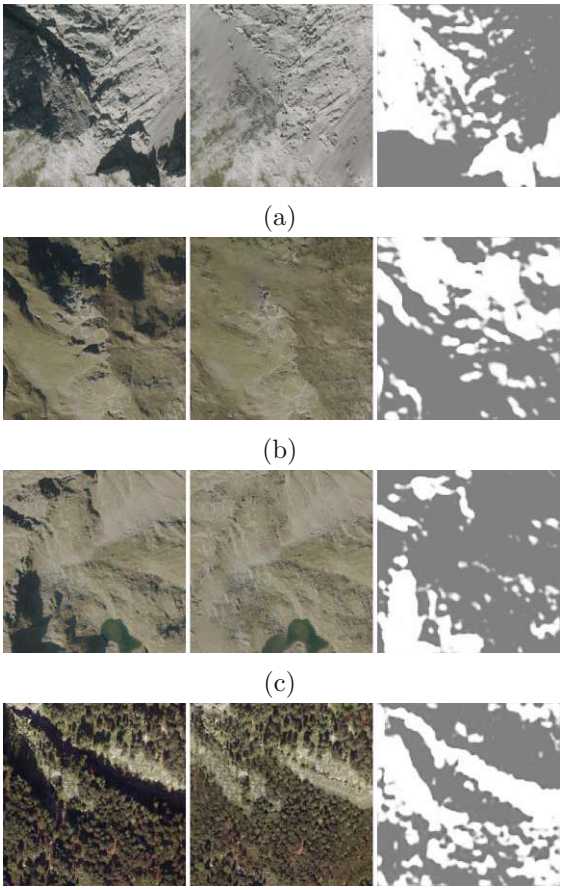
Figure 6.17: Due to their size, shadows of trees remain in the training set even after downsizing the images.

This experiment suggests that remaining small-scale shadows are a contributing factor that hinders the models' ability to generalize to real data. Figure 6.16 (d) shows that in forest areas, the shadow removal is much weaker. This might be due to the persistence of detectable shadow areas in forests occurring in the train set (see Figure 6.17). These shadows were exempt from the "shadow-free" criteria because shadow-free forests are not obtainable from orthophotos, and they are still visible after downsizing the images. This further underlines the above-stated assumption that small-scale shadows are problematic for the model's training process.

6.5 Using an Established Generative Model to Fill Predicted Masks

The shadow-detection performance of $M_{elevtion}$ showed to be quite effective. Therefore, a quick test was done to see if other generative models could use this information to produce a shadow-free version of the images. For this test, Adobe Photoshop's Content Aware Fill[1] (Version 23.5.0) was used together with the predicted shadow masks. Figure 6.18 presents the resulting images of this process. The outcome depicts somewhat good results, especially Figure 6.18 (b), which looks quite convincing. However, Content-Aware Fill is not optimized for shadow removal or the preservation of the underlying data. Many of the deshadowed areas depict terrain that is not really there, for example, the pond in Figure 6.18 (c). Therefore, this method can only be used for artistic visualizations of the terrain. Nevertheless, this experiment demonstrates that $M_{elevation}$ can be used as a basis for a sequential shadow-removal process.

[1] https://helpx.adobe.com/photoshop/using/content-aware-fill.html



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

Figure 6.18: Shadow-removal with Adobe Photoshop's Content-Aware Fill using the predicted shadow masks of $M_{elevation}$. The images show input (left), predicted output (center), and the predicted shadow mask (right).

6.6 Random Shadow Shapes

Upon achieving a model exhibiting visual proficiency on real shadow data through the utilization of a reduced-scale version of the initial dataset, a comparative visual analysis is conducted against a baseline dataset featuring randomly generated shadow shapes. This investigation aims to determine whether the incorporation of rendered realistic shadow shapes offers an advantage over random shapes in fostering generalization to real shadow data and contributes to answering **RQ1**.

The baseline dataset is derived using the identical data and process as the initial dataset but deploys Perlin noise and thresholding techniques, similar to the methodology employed by Morales et al. [MHT19], to produce the shadow shapes. Training procedures for both datasets involve identical hyperparameters applied over 100 epochs, utilizing elevation

data. For this experiment, only one model was trained with elevation data, due to the already subpar performance of the other model on the downsized ASD.

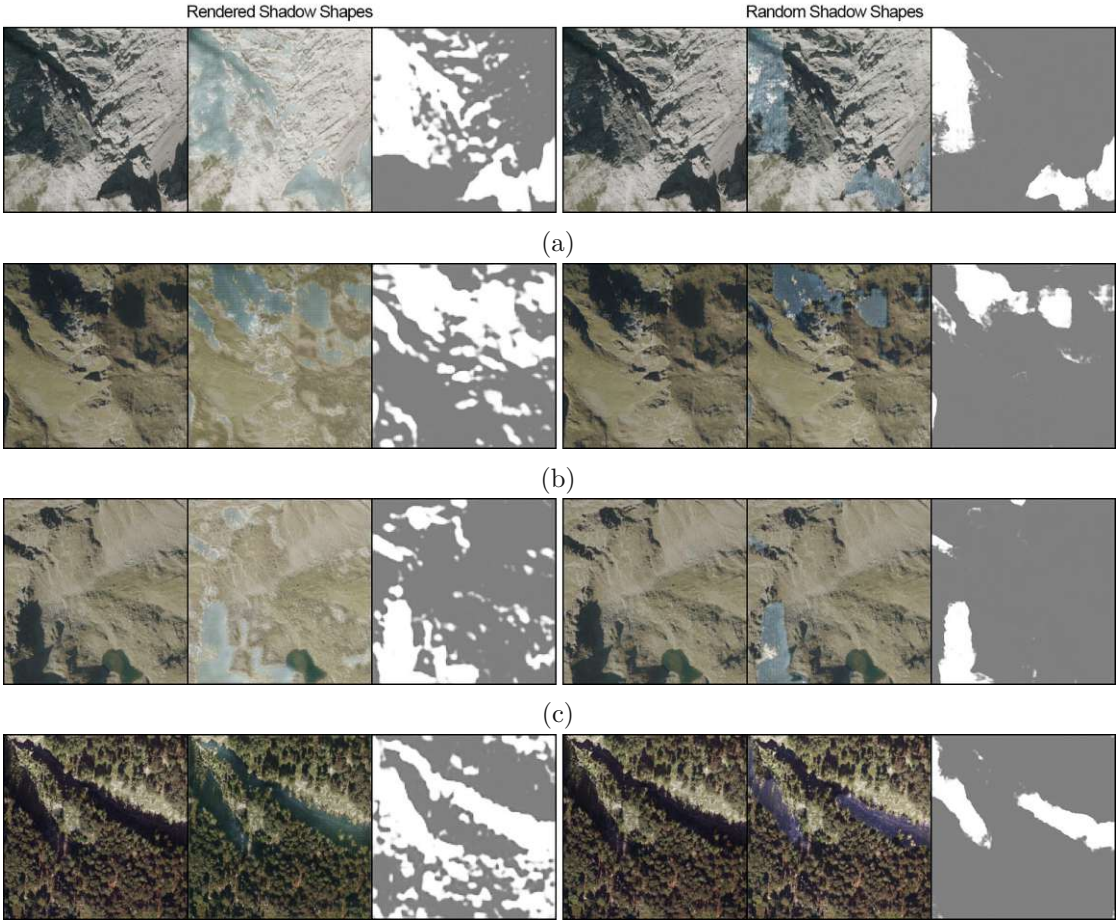


Figure 6.19: Performance comparison between models trained on realistically rendered shadow shapes and random shadow shapes. Both models were trained on the downsized version of the ASD and with elevation data. The images depict input, shadow-free prediction and shadow mask prediction from left to right.

A visual assessment of the model outputs reveals a substantial performance disparity on real shadow data. The model trained on random shadow shapes exhibits considerably inferior results compared to its counterpart trained on realistically rendered shadows using ray-tracing techniques. Figure 6.19 shows visible performance differences throughout all depicted examples. This observation underscores the pivotal role of realistically rendered shadows in enhancing the model's capacity for generalization to real-world data.

6.7 Preliminary Tests

A few preliminary tests were conducted to determine if the models were able to learn sufficiently from the provided data. In these tests, different shadow generation parameters were tested as well as different model architectures. In a first test, shadows were cast onto the orthophotos directly in the rendering engine (see Figure 6.20).

(a) Input (b) Shadowed

Figure 6.20: Artificially shadowed image directly rendered with Blender using the digital surface model of the area and the orthophoto as texture.

This was the fastest method, but it also restricted the adjustability of the shadows. Furthermore, it rendered a shadowed image straightaway without producing a shadow mask. This only allowed for a network where detection and removal were learned in a single CGAN.

Additionally, the rendering engine altered the shadow-free parts of the image, because they were lit by the light source, which brightened exposed areas. This produced subjectively bad results, where many of the artificial shadows were not removed correctly.
Therefore, the rendering engine was only used to produce the shadow masks, leaving the actual shadowing to the introduced algorithm and enabling the use of the ST-CGAN. This also aligns the method more with the already established approach of [MHT19].

The DSM was initially used for the shadow generation and input for the model. However, it had difficulties generalizing to real shadows. It was assumed that this was partially due to the inclusion of elevation data and shadows of ignored shadow areas, like height data and shadows of trees. Therefore, the DTM was used for all subsequent shadow generation methods.

Light-ray bounces were another aspect investigated. The default four ray bounces in Blender produced shadow masks with widely varying shadow intensities due to the reflected rays from the white texture and the topology of the terrain. The resulting shadowed images did not represent real shadows well, because real shadows are much more uniform, as seen in Figure 6.21c. Consequently, the number of bounces was reduced to zero, because all other values in between did not change the output significantly.

(a) 4 bounces (b) 0 bounces (c) Real shadows

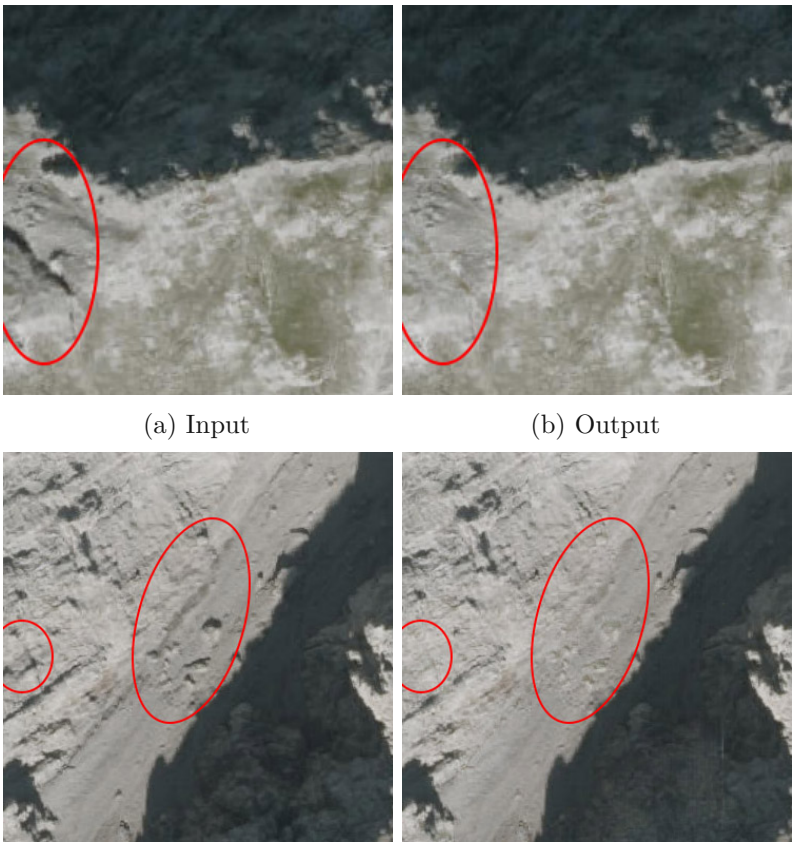
Figure 6.21: Artificially shadowed images rendered with four and zero ray bounces juxtaposed to real shadows.

In spite of the inconsistencies between the four bounce dataset and the real shadows, a full train run was conducted anyway to estimate the number of epochs needed and to see how these models would perform on real shadow data. Figure 6.22 shows the same examples as earlier but processed by the models trained on the dataset with four bounces.

Figure 6.23 displays that in some instances, the models were able to remove small scale shadows, especially the model with elevation data. In example Figure 6.23a the shadow of the trench in the upper right was removed, and in Figure 6.23c similar success can be seen with the shadows of the rocks in the center of the image. However, the removal is very inconsistent, and larger shadows are completely ignored.

(a) input (b) without elevation (c) with elevation

Figure 6.22: Shadow removal examples with real shadows. Models were trained on artificial shadows with four light-ray bounces. Each of the images comprises 36 smaller (256×256) images tiles sent to the network individually and stitched together afterwards.



(a) Input (b) Output

(c) Input (d) Output

Figure 6.23: Examples of small-scale shadow removal from the model trained on artificial shadows with four light-ray bounces and elevation data.

CHAPTER 7

Discussion

7.1 Performance on Artificial Shadows

In chapter 3 the efficacy of DEMs in rendering realistic artificial shadows was demonstrated. The generated shadows directly relate to the real topology of the terrain, presenting a departure from the method of [MHT19] where the shadow shapes are entirely random. Additionally, the inferred 3D geometry of the terrain and ray-tracing contribute to more realistic penumbra regions, simulating the real physical process. Addressing **RQ1**, we can say that DEMs can be seamlessly integrated into shadow-removal pipelines during the training phase as a data generation method. The proposed method of creating realistically shaped shadows also seems to be advantageous for the model's generalization capacity to real data, as evidenced by the visual analyses in Section 6.6.

The experiment on the ASD demonstrated the models' proficiency in handling artificial shadows from the test set, effectively mitigating the introduced darkening in aerial orthophotos. A quantitative evaluation of the experiment revealed that there is a statistically significant difference between the models regarding measures over the whole image. Specifically, model $M_{elevation}$ exhibits improvements over M_{RGB} according to the average RMSE and PSNR scores, even with the naive method of adding the elevation data as a fourth channel to the ST-CGAN. The effective size of the improvement is medium ($0.3 < |r| < 0.5$) measured with RMSE in LAB space and large ($0.5 < |r|$) measured with PSNR in RGB space. More advanced methods to add elevation or depth data to a network could yield even more substantial performance increases. One such method of adding depth data to a generative image network was introduced by Hu et al. [HYFW19]. This approach enhances object segmentation performance with depth data by separating the encoding site of an encoder-decoder architecture into three branches: RGB, Depth and Fusion. The fusion branch takes the combined output of the RGB and Depth branches and learns joint features. However, this network was not developed for

shadow removal and would need crucial adaptations that fall beyond the scope of this thesis.

In respect to **RQ2**, it was concluded that additional elevation data does exert a medium to large statistically significant impact on model performance. The marginal mean divergence between models may be attributed to the ASD's limited complexity, potentially causing convergence towards the upper limits of reconstruction efficacy. The RMSE and PSNR metrics calculated within the shadow regions were inconclusive. The RMSE showed a decrease, while the PSNR showed an increase in performance with elevation data. This inconsistency and the statistical insignificance of these metrics led to their exclusion from the final assessment. Nevertheless, the RMSE and PSNR computed over the whole image are the more important metrics anyway because they comprise false positives as well.

7.2 Generalizing to Real Data

When it comes to real shadow data, all the tested models failed to replicate their previous success on the artificial test set. Most of the time, real shadows were overlooked, and in some instances, the models produced cloud-like structures within shadowed areas. As network layer analysis with t-SNE showed, the models deliberately distinguish between real and artificial shadows. This behavior is especially evident in the bottleneck and L7 layers of the shadow-removal network G_2. In these layers, the most distant samples to the respective other class centroid consistently featured extensive areas of either real or artificial shadows. This exhibits the models' intent to maximize the distance between images with real and artificial shadows. The likely explanation for this lies in the nature of the GAN training framework and the persistence of small-scale shadows in the ground truth of the training set. The remaining real shadows appear to be a discerning factor for the discriminator, which likely considers the presence of real shadows as an indicator of authenticity. Consequently, the generator probably seeks to obtain the real shadows, hence learning to differentiate between real and artificial shadows. Notably, the projection showed overlaps of ground truth and real shadows throughout all layers, which suggests that the selection of ground truth samples is representative.

Training the models on a downsized version of the training images, aiming to mitigate the impact of small-scale shadows, yielded substantially improved results on real data, supporting this hypothesis. However, this approach conflicts with one of the initial goals of the shadow-removal method, which was to preserve as much detail as possible, and downsizing would inevitably result in a loss of detail.

Sanitizing the dataset from real shadows at the highest available level of detail is a really tedious process. While a shadow detection network could aid in filtering shadowed areas, manual oversight may still be required. Of course, the shadow detection network must also be created first, but this seems to be a much easier task, as the already very good shadow detection results of 6.4 demonstrate. Nevertheless, this network most likely has a bias against exactly the shadows that remain in the train set. A detection network trained on one of the available shadow-removal datasets is probably sufficient for the

detection of shadows in orthophotos, but this has to be evaluated first. Areas with vegetation, e.g., forests and rugged terrain, are also never shadow-free, which would lead to an underrepresentation of these areas. We are unsure how the shadow-free ground truth of these areas can be obtained without losing a lot of detail. The shadow-free ground truth from Morales et al. [MHT19] uses satellite pictures with a resolution of 2.8m/pixel which is around 10 times larger than the orthophotos used in this experiment. This resolution greatly reduces the impact of small shadows. At this scale, the defined small shadows (<5m) would not quite span 2 pixels and thus not really be present in the training images.

Qualitative visual evaluation of the outputs from models trained on the downsized ASD revealed that $M_{elevation}$ exhibits noticeably superior performance on real shadow data compared to M_{RGB}. Therefore, we can add to **RQ2** that not only does $M_{elevation}$ demonstrate marginally better quantitative performance on artificial data, but it also exhibits qualitatively superior performance on real shadow data. This observation suggests that elevation data aids the model in developing more robust features.

7.3 Constraints and Drawbacks of Using DEMs

One factor that likely limits the utility of elevation data is the much more detailed scale of orthophotos (16-29cm/px) compared to the lower resolution of DEMs (100cm/px). Downsampling the orthophotos as done in Section 6.4 relatively increases the detail of the DEMs and also includes more terrain for the model to infer information from. For example, a cliff casting a shadow might be on a neighboring tile. Increasing the observed area by downsampling mitigates this problem. The conducted experiment on the downsized ASD also showed better performance on real shadow data, especially with $M_{elevation}$, which might has to do with the relative resolution increase of the DEMs. Nevertheless, as already mentioned, downsampling is lossy and results in less detailed orthophotos. Employing the original resolution and increasing the considered area would demand larger models to handle the larger input images and still retain smaller DEM resolutions.

Considering the broader application of the approach, using elevation data comes with some drawbacks. The additional data leads to more processing requirements and more trainable parameters to accommodate the additional information. This is primarily relevant for training but, to some extent, also for later use. Furthermore, elevation data is not always available, and gathering the required data poses an additional burden. Moreover, this data has to be registered with the orthophotos in the same 2D space, which has to be done during test time as well. Training models with elevation data also restricts the usage of already scarce shadow training datasets like the ISTD [WLY18], ISTD+ [LS19] or SRD [QTH+17]. However, these datasets could be valuable to pretrain or test the models. Pretraining a model on one of these datasets and then fine-tuning it on orthophotos with artificial shadows could be another promising approach to the problem. Another downside of using DEMs as a shadow generation method is that they

restrict the size and shape of the produced shadows. This limits the control over the amount of shadow areas produced and the diversity of the shapes.

7.4 Other Observed Challenges

A number of challenges were observed during the dataset generation and the experiment. First and foremost, not only are there many types of different terrain, but there are also large differences in the capturing process of the images. They have varying quality, resolution, contrast, saturation and hue, which makes the data domain even more diverse and challenging. As seen in the experiment, cGANs can also produce hallucinations. This should be kept in mind if the deshadowed images are used in maps for trip planning. Paths, cottages or rocks visible in the images might not be there in the real world. Sometimes it can be quite challenging to classify a specific region as a shadow. For example, mountain creeks or bushes can be easily confounded as shadows in some cases. This underlines how difficult the problem is and that errors can easily occur even with human oversight.

7.5 Limitations

In pursuit of the overall objective of generating shadow-free alpine orthophotos for scene relighting under varying sun angles, the scope of the shadow-removal domain was confined to addressing large-scale shadows present in orthophotos from the Austrian Alps. The orthophotos, obtained from Basemap [Bas], predominantly comprises images captured during summer, thereby limiting the dataset to scenes from this season. Therefore, the generalizability of the models to completely different geographical areas, such as the Andes, Rocky Mountains, or Himalaya, remains uncertain. Furthermore, the models' performance under different seasonal conditions, particularly winter scenes, remains an open question.

Large-scale shadows were defined as all shadows larger than 5 meters in length. Furthermore, flat landscape regions were excluded from the dataset, as the underlying DEM lacks significant terrain changes conducive to shadow rendering and feature development. The complexity of obtaining shadow-free images in areas populated with bushes, trees, and forests led to the deliberate omission of shadows caused by vegetation. This does not mean that these areas were ignored, but that images containing shadows from vegetation were counted as shadow-free. Completely excluding these areas would lead to an underrepresentation of a large proportion of real-world data.

CHAPTER 8

Conclusion

In conclusion, this thesis proposes a novel dataset generation pipeline tailored for shadow-removal in aerial orthophotos. This pipeline utilizes DEMs and ray-tracing to produce physically plausible shadow shapes aligned with the underlying geospatial topology of the terrain. Sequentially, this pipeline was used to generate the *Alpine Shadow Dataset* which contains RGB and elevation data encoded into the alpha channel of the images. This dataset depicts various scenes featuring alpine landscapes in Austria. A visual analysis substantiated the superiority of artificial shadows generated by this pipeline over random shapes, particularly in terms of their utility for model generalization to real shadow data. Addressing **RQ1**, we demonstrated that DEMs can be effectively used for data generation, enabling the rendering of realistic shadow shapes. Moreover, it was demonstrated that elevation data can be successfully encoded into a fourth input channel of the state-of-the-art shadow-removal model ST-CGAN.

Various experiments were conducted with the adapted version of the ST-CGAN and the *Alpine Shadow Dataset*. The results were evaluated quantitatively with common metrics for shadow-removal and qualitatively through a visual analysis to determine real shadow performance. Concerning **RQ2**, our findings indicate that introducing DEMs as model inputs yields statistically significant improvements with a medium to large effect size on artificial data.

Initially, all trained models failed to generalize to real data. A conducted analysis of the output of each network layer has shown that the models deliberately distinguish between real and artificial shadows in the bottleneck and lowest decoding layer of the shadow-removal network G_2. Through an experiment on a downsized version of the *Alpine Shadow Dataset*, it was argued that the remaining real shadows in the training set are the main perpetrators of this problem. These remaining shadows include shadows from trees, bushes and other smaller shadows excluded from the shadow-free criteria. However, having a very sanitized ground truth data set seems crucial. Otherwise, the

generator learns to leave real shadows in the output images to deceive the discriminator effectively, hence differentiating between real and artificial shadows.

The downsized dataset yielded substantially better results on real shadows, especially when it comes to the model trained with elevation data, which outperformed the model without elevation data noticeably. Therefore, we argue that elevation data fosters noticeable improvements regarding the generalizability of shadow-removal models, additionally contributing to answering **RQ2**.

Although, the shadow detection of the network trained with elevation data already exhibited quite good results, the shadow removal performance was still subpar. The likely reason for this is the small training set size after downsizing the images. However, a small-scale experiment showed that the predicted shadow masks can be used as the basis for other generative model to reconstruct the shadowed area.

CHAPTER 9

Future Work

Many questions raised are beyond the scope of this thesis, and therefore some possible future approaches and experiments to answer these will be discussed in the following sections.

9.1 Training with a Larger Dataset

We showed that the real shadow performance trained with a downsized version of the dataset yielded substantial improvements over the more detailed primary dataset. This is likely due to the reduced impact of small-scale shadows in the dataset and the increased relative resolution of the DEMs. However, the shadow-removal performance was still insufficient, which is likely due to the small size of the data set after downsizing it. Morales et al. demonstrated that it is possible to remove shadows cast by clouds at lower resolutions (2.8 m/pixel) [MHT19]. Therefore, an experiment with a larger dataset should be conducted to elucidate if the models are also able to sufficiently remove shadows cast by terrain. The same proposed data generation pipeline can be used for this process. However, the ”shadow-free“ criteria has to be more strict for this iteration, as remaining small-scale shadows have shown to have a negative impact. Nevertheless, it should be easier to fulfill these criteria at a lower level of detail.

9.2 Propagate Information from Low to High Level of Detail

If training the models on a low level of detail proves to be successful, the produced shadow-free images and shadow masks can offer useful information for shadow-removal at a high level of detail. A network specializing in upscaling and reconstruction could use the masks, and/or shadow-free images together with the detailed shadowed images

to infer a detailed shadow-free image. This process is still hypothetical, and a specialized training algorithm and network needs to be developed to achieve this, but this approach appears to be feasible.

9.3 Dedicated Network for Elevation Data

The experiments demonstrated that additional elevation data improves the performance of the ST-CGAN. Nevertheless, the chosen approach to adding elevation data was quite naive and simple. There are much more sophisticated methods to incorporate elevation or depth into an image processing network. One such method developed for object segmentation with RGB-D data was proposed by Hu et al. [HYFW19]. The so-called ACNet comprises three branches at the encoding site of a U-net like network: a branch for RGB, a branch for depth data, and a fusion branch that combines the features of both branches after each layer. It seems feasible to adapt this strategy to shadow-removal with elevation data, analogous to the ST-CGAN approach of Wang et al. [WLY18].

9.4 More Realistic Shadowing Algorithm

The introduced shadowing algorithm sometimes generates shadows that appear overly dark, because of the varying contrast of the ground truth images. A more sophisticated shadowing algorithm that takes the overall contrast and brightness of the images into account could also help the realism of the training data and the models' ability to generalize. The rendered shadow masks only depict the direct illumination of an area. However, orthophotos could be used as an approximation of the reflectance of the depicted materials to model indirect lighting as well. Furthermore, a sky texture could also be used to approximate the light reflected by the sky. Another aspect disregarded is the albedo and shading of the objects, which do not match the artificial shadowing and cannot be simply created by darkening specific light frequencies. In order to make shadowing and shading consistent, the whole image has to be rendered from a virtual scene. A completely artificial dataset could be created from DEMs and varying, procedurally applied landscape textures. Nevertheless, possible distribution shifts between these artificial and real shadows always remain.

9.5 Domain Adaptation

To overcome domain shifts between artificial and real data, methods like test-time training (TTT) could be applied to the ST-CGAN. This could be especially beneficial for the previously discussed, completely virtually rendered dataset. In the original work of Sun et al., rotation prediction was used as a secondary task for a discriminative model to learn from unlabeled test data [SWL+20]. The authors demonstrated that this method is quite successful in increasing the models' robustness. In our case, the unlabeled data are the real shadow images, and the model is generative. However, predicting rotations

does not make sense for orthophotos because they do not have a canonical orientation. A much better secondary task for the model would be masking areas of the image and letting the model reconstruct them, as proposed by Gandelsmann et al. [GSCE22]. The authors use masked autoencoders (MAE) as the self-supervised secondary task for a discriminative model, which also demonstrated a significant increase in class prediction performance on the test data. In order to incorporate the MAE into the ST-CGAN a second decoder could be added after the shared encoder assigned to the reconstruction task. Masking and reconstructing could be a very beneficial secondary task because, as with shadow-removal, it is a reconstruction process.

List of Figures

List of Tables

Bibliography

[AHO10] Eli Arbel and Hagit Hel-Or. Shadow removal using intensity surfaces and texture anchor points. *IEEE transactions on pattern analysis and machine intelligence*, 33(6):1202–1216, 2010.

[AI22] Ryo Abiko and Masaaki Ikehara. Channel attention gan trained with enhanced dataset for single-image shadow removal. *IEEE Access*, 10:12322–12333, 2022.

[Alp] Alpine maps. `https://alpinemaps.org`. Accessed: 2023-04-18.

[App68] Arthur Appel. Some techniques for shading machine renderings of solids. In *Proceedings of the April 30–May 2, 1968, Spring Joint Computer Conference*, AFIPS '68 (Spring), page 37–45, New York, NY, USA, 1968. Association for Computing Machinery.

[Bas] Base map. `https://basemap.at`. Accessed: 2023-10-11.

[Bon09] Charles Boncelet. Image noise models. In *The essential guide to image processing*, pages 143–167. Elsevier, 2009.

[CZZY17] Yun Cao, Zhiming Zhou, Weinan Zhang, and Yong Yu. Unsupervised diverse colorization via generative adversarial networks. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18–22, 2017, Proceedings, Part I 10*, pages 151–166. Springer, 2017.

[FHLD05] Graham D Finlayson, Steven D Hordley, Cheng Lu, and Mark S Drew. On the removal of shadows from images. *IEEE transactions on pattern analysis and machine intelligence*, 28(1):59–68, 2005.

[Fli] Flight simulator. `https://www.flightsimulator.com`. Accessed: 2023-04-18.

[GJZ+19] Mingliang Gao, Jun Jiang, Guofeng Zou, Vijay John, and Zheng Liu. Rgb-d-based object recognition using multimodal convolutional neural networks: A survey. *IEEE access*, 7:43110–43136, 2019.

[Goo] Google earth. https://earth.google.com. Accessed: 2023-04-18.

[GPAM+14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[GSCE22] Yossi Gandelsman, Yu Sun, Xinlei Chen, and Alexei A. Efros. Test-time training with masked autoencoders, 2022.

[HJFH19] Xiaowei Hu, Yitong Jiang, Chi-Wing Fu, and Pheng-Ann Heng. Mask-shadowgan: Learning to remove shadows from unpaired data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2472–2481, 2019.

[HNL09] Stephan Harvey, Paul Nigg, and Kernausbildungsteam Lawinenprävention. Practical risk assessment and decision making in avalanche terrain. an overview of concepts and tools in switzerland. In *Proceedings of the International Snow Science Workshop, Davos*, pages 654–658, 2009.

[HYFW19] Xinxin Hu, Kailun Yang, Lei Fei, and Kaiwei Wang. Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1440–1444. IEEE, 2019.

[IZZE17] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[LS19] Hieu Le and Dimitris Samaras. Shadow removal via shadow image decomposition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8578–8587, 2019.

[LZY21] Jun Lv, Jin Zhu, and Guang Yang. Which gan? a comparative study of generative adversarial network-based fast mri reconstruction. *Philosophical Transactions of the Royal Society A*, 379(2200):20200203, 2021.

[MHT19] Giorgio Morales, Samuel G Huamán, and Joel Telles. Shadow removal in high-resolution satellite images using conditional generative adversarial networks. In *Information Management and Big Data: 5th International Conference, SIMBig 2018, Lima, Peru, September 3–5, 2018, Proceedings 5*, pages 328–340. Springer, 2019.

[MO14] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[MR03] Sk Sazid Mahammad and R Ramakrishnan. Geotiff-a standard image file format for gis applications. *Map India*, pages 28–31, 2003.

[PEH20] Laurent Polidori and Mhamad El Hage. Digital elevation model quality assessment methods: A critical review. *Remote sensing*, 12(21):3522, 2020.

[QTH+17] Liangqiong Qu, Jiandong Tian, Shengfeng He, Yandong Tang, and Rynson WH Lau. Deshadownet: A multi-context embedding deep network for shadow removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4067–4075, 2017.

[RFB15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[RMML19] Mir Mustafizur Rahman, Gregory J McDermid, Taylor Mckeeman, and Julie Lovitt. A workflow to minimize shadows in uav-based orthomosaics. *Journal of Unmanned Vehicle Systems*, 7(2):107–117, 2019.

[Sha] Shadow map. `https://shadowmap.org`. Accessed: 2023-04-18.

[SL08] Yael Shor and Dani Lischinski. The shadow meets the mask: Pyramid-based shadow removal. In *Computer Graphics Forum*, volume 27, pages 577–586. Wiley Online Library, 2008.

[SP07] David Schobesberger and Tom Patterson. Evaluating the effectiveness of 2d vs. 3d trailhead maps. In *Mountain Mapping and Visualisation: Proceedings of the 6th ICA Mountain Cartography Workshop*, pages 201–205, 2007.

[SWL+20] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pages 9229–9248. PMLR, 2020.

[VST+23] Florin-Alexandru Vasluianu, Tim Seizinger, Radu Timofte, Shuhao Cui, Junshi Huang, Shuman Tian, Mingyuan Fan, Jiaqi Zhang, Li Zhu, Xiaoming Wei, et al. Ntire 2023 image shadow removal challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1788–1807, 2023.

[Wil78] Lance Williams. Casting curved shadows on curved surfaces. In *Proceedings of the 5th annual conference on Computer graphics and interactive techniques*, pages 270–274, 1978.

[WLY18] Jifeng Wang, Xiang Li, and Jian Yang. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1788–1797, 2018.

[ZDH+18] Lei Zhu, Zijun Deng, Xiaowei Hu, Chi-Wing Fu, Xuemiao Xu, Jing Qin, and Pheng-Ann Heng. Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 121–136, 2018.

[ZPIE17] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.