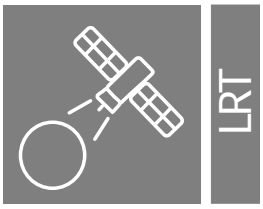

Universität der Bundeswehr München

Professur für
Erdbeobachtung

Interdisciplinary Project (IDP)

Exploiting Single Image Height Estimation for Change Detection from Multi-Sensor Remote Sensing Images

Student: Baran Ekin Özdemir

Datum: 17. October 2023

Supervisor: PD Dr.-Ing. habil. Michael Schmitt

Advisor: Michael Recla

Contents

1 Introduction

One of the most important application fields in remote sensing for earth observation is the change detection. This field in general aims to detect changes in topography and urban areas caused by natural events or human activities such as surveying changes in the elevation of a terrain due to landslides, detecting new constructions or observing changes due to natural or human-made disasters such as earthquakes or wars in urban areas.

For automated detection of changes, there needs to be pre-event and post-event images for comparison. In order to apply classical change detection approaches, these images need to be acquired from the same sensor at the same orbit with a similar observation position however, this is not generally the case especially for some applications with tight time constraints such as responding to disasters, which requires utilization of multi-sensor approach to take advantage of available sensor at the time of need regardless of its configuration thus satisfying the time resolution requirement of the application. For example, in the case of an emergency, only available pre-event might be a synthetic aperture radar (SAR) image where only available post-event image might be an optical image. Even within SAR constellations, there is a high diversity in terms of orbits and observation positions to ensure high frequencies of revisiting to be more responsive to emergency situations. Methods to overcome such problems require detailed geometric knowledge concerning the topography which is often unavailable in cases where change detection is needed. Needless to say, this situation possesses a great challenge for traditional methods.

Recently, deep learning techniques, especially Convolutional Neural Networks (CNNs), have been successfully used opposed to traditional methods to have robust feature representation of scenes as automated change detection heavily requires understanding of spatial context rather than just spectral information. Many of these methods are mainly focused on supervised methods [1] [2], however in the case of multi-sensor and multi-modal applications, it is very costly or even impossible to acquire annotated data for changes. Thus, unsupervised or self-supervised approaches [3] [4] are favored to supervised in such applications.

In this project, an approach to perform self-supervised change detection in multi-sensor and multi-modal remote sensing images using single image height estimation is presented. The proposed approach exploits single image height estimations to project pre-event and post-event synthetic aperture radar (SAR) remote sensing images to a common and comparable space of height maps. This step serves a feature representation of

the scene while ideally mitigating disparities arising from distinct acquisition modes, incidence angles, orbits, configurations, or varying noise profiles. Then change detection is performed on the height maps by a self-supervised change estimation model with Siamese architecture which is very popular for this task. Height maps for SAR images are generated by SAR2Height model proposed by M.Recla and M.Schmitt (2021) [5].

The extension of this concept to allow multi-sensor capabilities by incorporating optical images requires height maps of optical images comparable with SAR height maps. To investigate this possibility, an approach for optical single image height estimation exploring the potential use of synthetic optical data is also presented in this project.

Rest of this report is organized as follows. Section 2 presents the approach for single image height estimation on optical images exploring the potential use of synthetic data in the training. Section 3 introduces the proposed approach for self-supervised change detection on height maps. Results for change detection are presented in the Section 4 and discussed in the Section 5. Final conclusions for the project are drawn in the Section 6. Finally, in the Section 7, an outlook for future improvements are discussed.

2 Optical Height Estimation

2.1 Motivation and Challenges

2.1.1 Differences between Optical and SAR Images

SAR sends microwave pulses and uses the reception of back-scattered signals to generate imagery. Being an active remote sensing technology, SAR emits its own radiation, thereby enabling it to operate irrespective of day or night conditions. It also remains unaffected to variations in natural lighting, weather conditions or cloud cover.

2.1.2 Challenges for Optical Images

Unlike SAR images, optical images are impacted by weather conditions like clouds or fog, which can either obstruct parts of the image or alter how the scene appears. However, the most crucial factor influencing the appearance of an optical image is the sun because it affects how the scene is lit affecting the appearance and illumination, and more notably for the task of height estimation, it dictates the length, strength and direction of the shadows.

In the context of estimating building heights, two primary visual attributes assume significance:

- the extent of visible building facades
- the length of shadows cast by the buildings.

Notably, the former holds heightened importance, particularly when the image acquisition angle is more nadir-looking.

2.1.3 Motivation for Synthetic Data

To develop a deep learning model paying uttermost attention to shadow differences for height estimation, while simultaneously keeping itself resilient to variations in lighting and weather conditions, a prerequisite is the availability of an expansive dataset comprising diverse shadow and lighting scenarios of the same geographical area without actual changes so that the underlying geometry of the scene is preserved. This way, model is expected to effectively distinguish between variations in the imagery stemming from changes in the terrain and those arising from fluctuations in lighting and shadowing conditions. The former is the focal point of the change detection task, while the latter may be viewed as noise.

However, such a dataset of optical satellite images is very costly to acquire. Developed models often suffer from lack of such data as described above. Consequently, for this project, the idea of utilizing a synthetic dataset generated from a 3D model of an urban environment is explored.

2.2 Synthetic Data

In comparison with obtaining real satellite images, producing synthetic data is significantly more cost-effective. This approach also grants the ability to alter certain environmental factors at will, such as viewing angles, sunlight properties, buildings and generate virtually unlimited data. However, using synthetic data to imitate real world imagery comes at a cost of domain gap between real and synthetic data as synthetic data is most likely to suffer from lack of detail, realism, variation of objects present in the scene and natural noise of capturing process.

The ultimate goal of any synthetic approach is to narrow this gap by achieving extreme realism while maintaining the flexibility of manipulation to generate large volumes of data as needed. However, it's important to note that developing such synthetic data methods requires significant skill, effort and incurs various costs. The potential directions for future research in synthetic data approaches are discussed in the outlook section.

For the scope of this project, the objective is to develop a relatively straightforward method capable of generating sufficiently useful data to investigate the feasibility of training a model that can consider shadows for height estimation in scenarios characterized by a lack of fine detail, minimal surface texture, and an nadir-looking perspective, where not many other cues are available.

2.2.1 3D City Generation

In this project, the goal for the synthetic data generation method is to be able to create an urban environment with buildings, roads and vegetation. It is important to have variation building shapes and rooftops, as well as natural looking city layout throughout with curved roads and roundabouts. Otherwise, especially with the lack of realism, model might lean on memorizing structures and patterns, rather than learning underlying features.

To accomplish these objectives, a pipeline which makes use of Blender-OSM [6] tool is employed. This tool uses Open Street Map (OSM) [7], which is an open-source geographic database, to retrieve building footprints and street layouts from real-world geographical locations and generates 3D buildings on top of these footprints in the 3D modelling software Blender. This way, variation in building shapes in real maps is preserved. Furthermore, the tool offers control over distribution of generated building heights, thereby allowing the realization of the desired diversity within the synthetic urban environment.

City Map	Height Levels	Max Height	Default Roof Type
Munich	10	119	Gabled
London	10	317	Flat
Nuremberg	5	135	Gabled

Table 2.1: Synthetic cities for this project.

For this project, three cities are generated with different height distribution and default rooftop types as shown in Table 2.1

As a second step, resulting 3D city buildings are textured with 10 facade textures, 4 gabled roof textures and 3 flat roof textures randomly. This introduces even more variation in building appearances. Concurrently, other existing elements within the scene, including roads, sidewalks, trees, and bodies of water, are also textured to enhance their visual representation.

The scene is illuminated with a simulated Blender sunlight. Blender sunlight object allows control over lighting direction and angle which governs shadow orientation and length, lighting intensity which affects how bright the scene is and lighting color which aids emulating how a scene would look under different weather conditions. The sunlit scene is then rendered with the ray-trace based Cycles engine and denoised with NVIDIA OptiX for realistic lighting and shadows. For a visual representation of the final output, please refer to Figure 2.2

2.2.2 Synthetic Dataset Generation

City scenes are depicted from the viewpoint of 5 nadir-looking orthogonal cameras to mimic the appearance of satellite imagery. The cameras and output images are scaled to match 1 pixel per meter resolution similar to the available optical images. In order to produce a synthetic dataset that closely emulates and is directly comparable to the existing optical dataset which is single channel panchromatic, output images are then converted to grayscale. For the visualizations, please refer to the Figure 2.3

Ground truth height maps are generated by texturing the objects in the scene with a custom material that normalizes global height of each point in the scene by the maximum height of the scene. For the material implementation in Blender, please refer to the Figure 3.4

Final training and validation datasets are created by capturing two cities, Munich and London, from 5 different cameras and under 5 different lighting and ground texture combinations, resulting in 50 scene images. Test dataset on the other hand, is created by capturing another city, Nuremberg, employing a single camera and under 2 different lighting and ground conditions, resulting in 2 scene images. From each scene image, 200 non-overlapping patches of size 512x512 pixels to be used as input to the model are

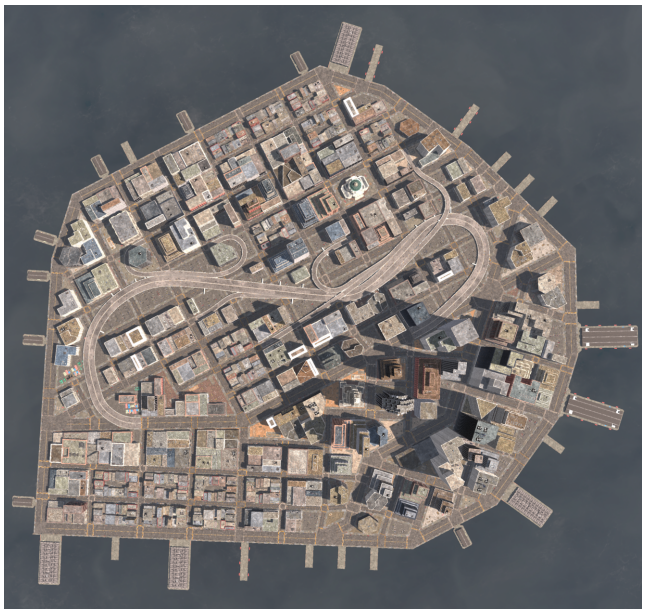

Figure 2.1: Example 3D urban environment offered by Unreal Engine 5 as the City Sample project. Even though graphics are detailed and realistic, there is not enough variation in the building shapes and street layout. Besides, the area is limited to given city boundaries, preventing generation of high volume synthetic data.

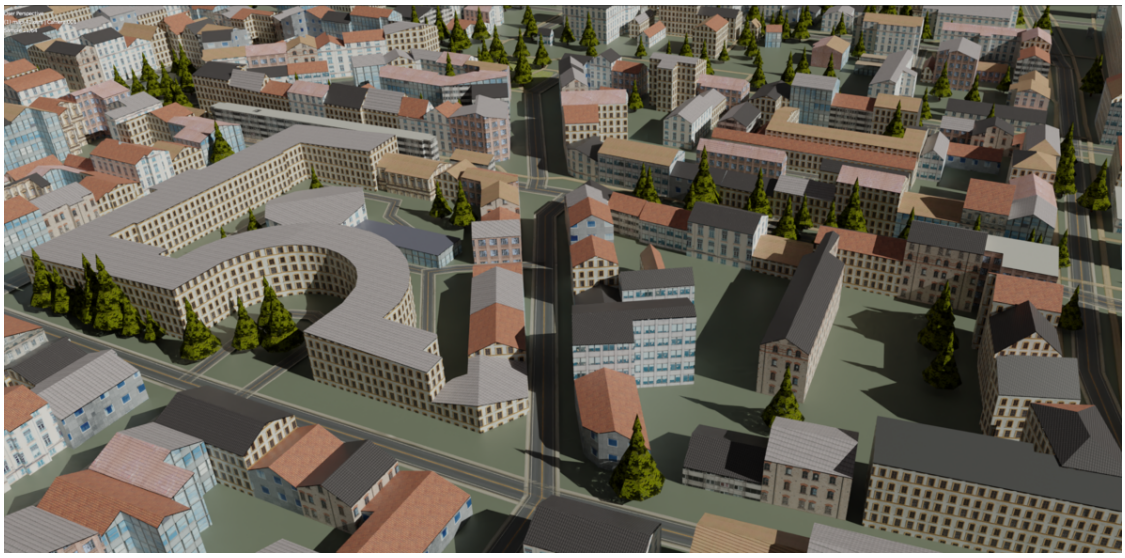

Figure 2.2: Close-up visualization of 3D urban environment based on Nuremberg map realized by our method. With this method real city maps can be directly imported, offering very large city models with high variation in building shapes and sizes in a natural looking layout.

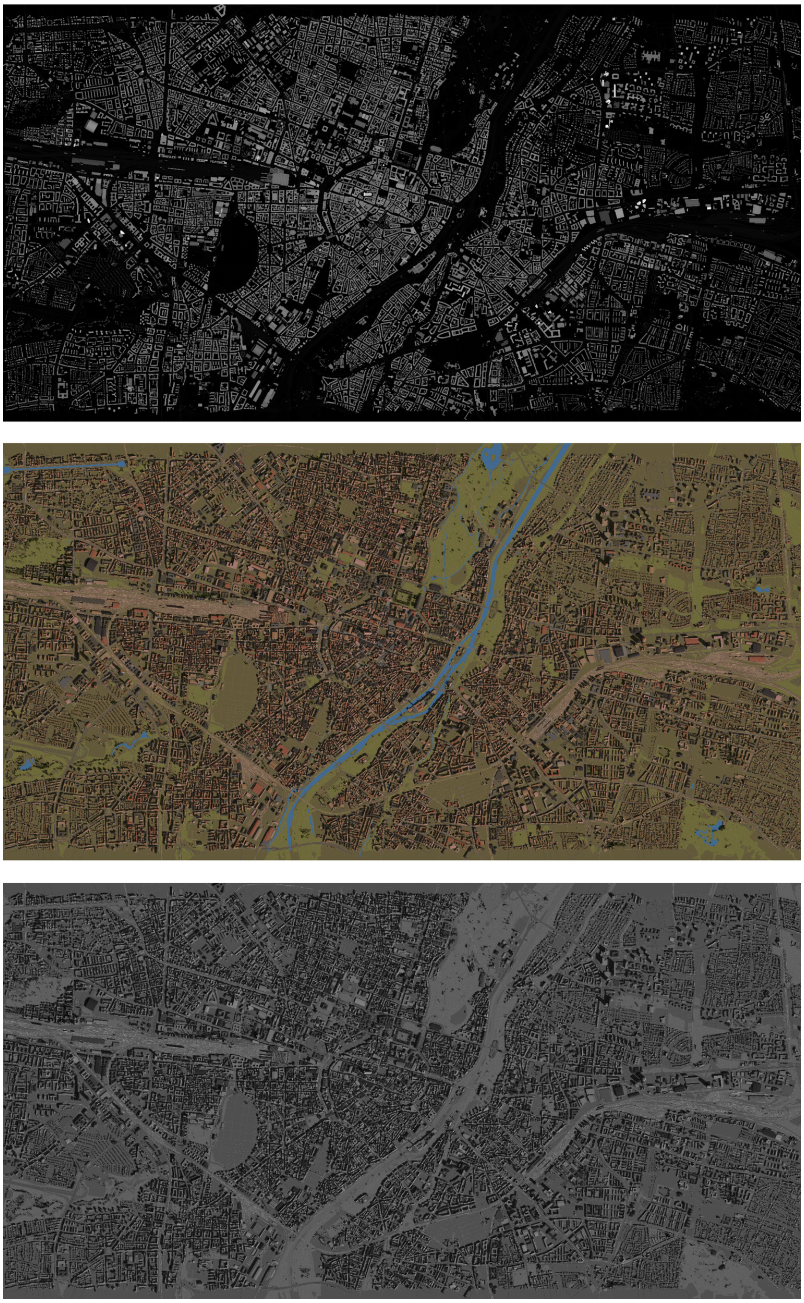

Figure 2.3: Views of the Munich based city. Top is the height map of the city, used as a ground truth for height estimation. Center is the city with textures and lighting. Bottom is the grayscale counterpart of the center image, used as an input to the model.

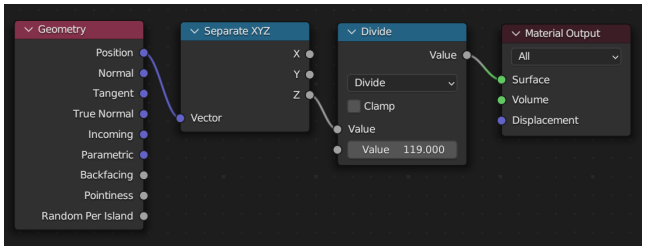

Figure 2.4: Material implementation in Blender to generate height maps. Z component of global position for each point is normalized by the maximum height of the scene to have floating point values in [0,1] range.

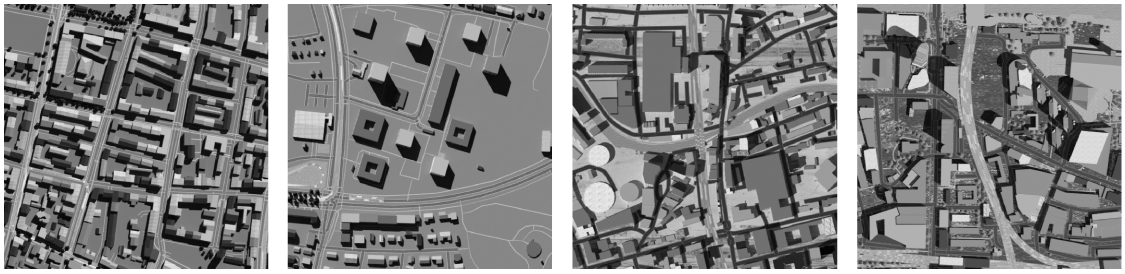

Figure 2.5: Input image examples.

extracted in a grid pattern. In total, this results in a dataset of 10,000 images designated for training and validating purposes, while the test dataset comprises 400 images. Each image is bundled with its corresponding height map as a ground truth for loss calculation. The example patches used as input image can be seen in Figure 2.5

2.3 Height Estimation

2.3.1 Model

In this project, optical version of the network proposed by M.Recla and M.Schmitt (2021) [5], which is an adaptation of IM2HEIGHT [8] is employed to investigate possibilities with the synthetic data. In essence, this network is structured with an architecture of an encoder-decoder paradigm, consisting of two sub-networks, namely convolutional and deconvolutional networks, both incorporating residual blocks. The convolutional sub-network utilizes pooling layers to create a bottleneck as in encoder-decoder architectures while the deconvolutional sub-network utilizes unpooling layers and skip connections to bring the output dimensions to match the input dimensions, estimating a height value for each pixel in the input image.

2.3.2 Training

For the training of the model, 85% of 10,000 input pairs of images with their corresponding height maps is used, while remaining 15% is reserved as validation set. Model is trained in mini-batches with a batch size of 3 over 42 epochs. For optimization, Adam optimizer with a learning rate of 0.0001 is used. Mean absolute error (MAE), also referred as L1 loss, is employed as the loss function. With x_i being the estimated height value and y_i being the ground truth height value, loss value over N samples can be expressed as:

$$L(x_i, y_i) = \frac{1}{N} \sum_{i=1}^{N} |x_i - y_i| \tag{2.1}$$

2.4 Results

The evaluation of the model's performance is conducted using the test dataset derived from Nuremberg based city, a dataset held separate from the training data. It is worth noting that although the images within this test dataset have very similar appearance and style to those encountered in the training dataset, given that they are generated through the same synthetic pipeline, there are two important distinctions:

- Novel City Layout: The city layout featured in the test dataset is entirely novel, resulting in distinct building configurations, roads, and vegetation patterns not present in the training data.
- Unique Lighting Conditions: The test dataset introduces novel lighting conditions, resulting in unprecedented shadow orientations, lengths, and intensities that differ from those encountered during the training phase.

Figure 2.6 shows the estimations of the model on two different example synthetic optical images from the test set and one example real optical image alongside the corresponding ground truth height maps.

2.5 Discussion

2.5.1 Synthetic Images

In the case of synthetic images, the accuracy of estimations allows us to infer that the model is effectively incorporating shadow information into the height estimation process. This inference holds especially true given the almost-perfect nadir perspective of the images, which deprives the model of alternative cues for determining building heights. As illustrated in the first example presented in Figure 3.6, the oddly shaped building positioned in the lower-left is significantly taller than its neighboring structures, accurately estimated by the model. A similar accurate estimations can be seen in the three clusters of horizontally arranged buildings in the upper-left region of the first image.

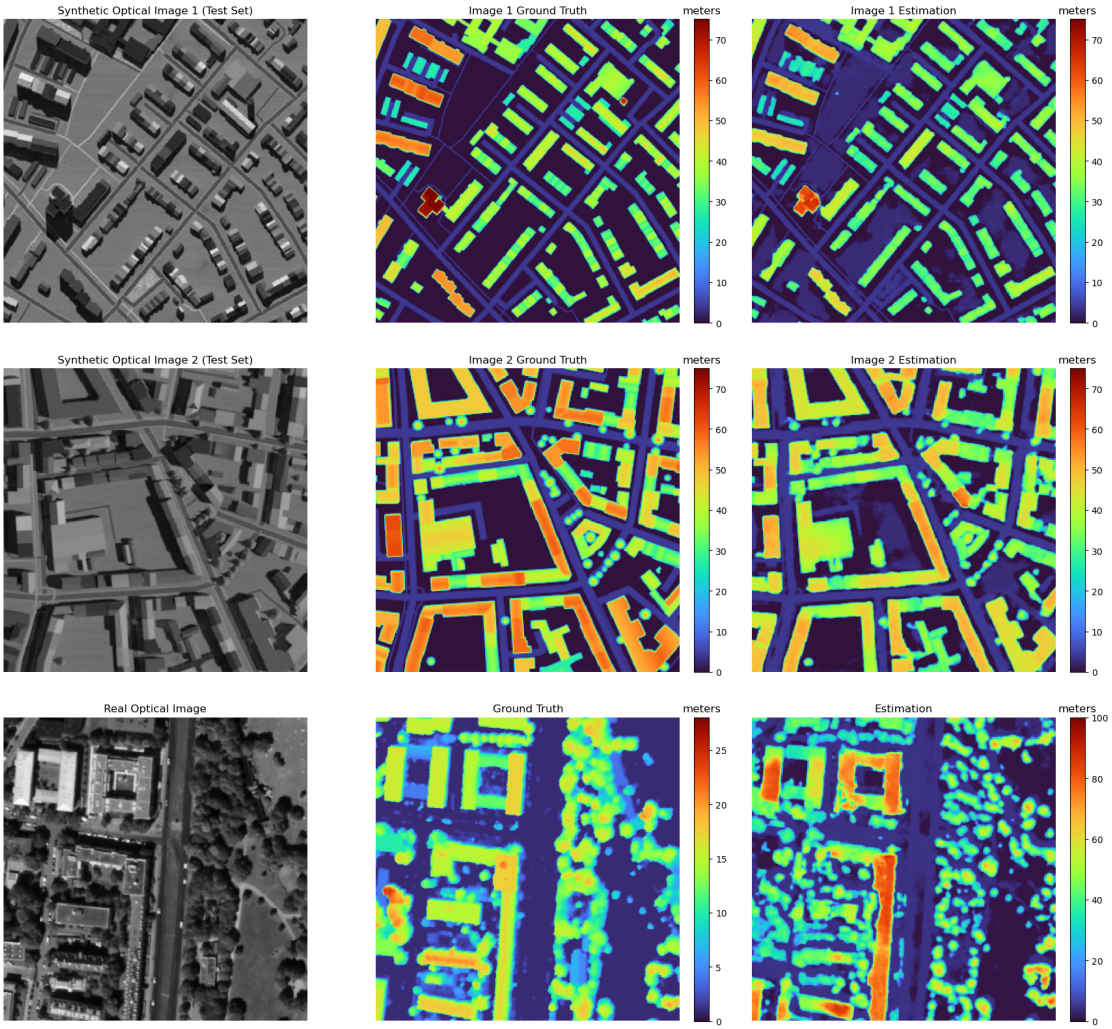

Figure 2.6: Estimations of the model. First column is the input image, second column is the ground truth height map, third column is the height estimation of the model. Top and center rows are showing results on synthetic test images while bottom row shows results on real optical satellite image.

Similarly, the very central portion of the second image offers a good illustration of the model's achievement in recognizing shadow-related nuances, showing distinct variations in shadow levels among adjacent buildings. This example shows that the model successfully aligns building tops with corresponding shadows to derive height estimates. It is also important to note that the model reliably distinguishes between buildings and non-building areas, even in cases where building roofs are almost as dark as the shadows.

Regarding its shortcomings, it's noticeable that the model accurately distinguishes height differences between buildings but tends to underestimate the overall heights of structures within the entire scene. This is a common problem in single image height estimation models due to the heavy tailed distribution of the data. There are always more pixels near zero height, representing ground level, compared to pixels corresponding to tall buildings.

Another limitation arises when the ratio between building height and its area is either exceptionally high or excessively small. For instance, consider the building in the upper right region of the initial image in Figure 2.6, which is exceptionally tall yet has a tiny footprint. In such cases, the shadow's coverage area significantly exceeds that of the building's footprint, posing challenges for the model in establishing accurate correspondences between the building and its shadow, resulting in incorrect height estimations.

A similar issue arises when the height-to-area ratio is exceedingly low, leading to potential confusion between shorter, larger buildings and the surrounding ground. This issue is amplified by the lack of realistic texturing, which hinders the difference between the building and the ground. It is important to note, however, that these challenges due to the limitations of the synthetic data and do not extend to real-world scenarios where building heights are correlated with their footprint size and many environmental details aid in distinguishing objects.

2.5.2 Real Images

In the case of real images, the obtained results are noticeably less accurate as anticipated, but still promising. The third row of Figure 2.6 shows outcomes derived from a real optical satellite image, one that also is nadir-looking perspective. It is evident that the model tends to greatly overestimate building heights, highlighted by the different color scales of the ground truth and the model's estimations. This difference can be attributed to the fact that, on average, the synthetic data buildings are significantly taller compared to those encountered in the example real image. Additionally, the model encounters difficulties in detecting some buildings, as their shadows blend with surrounding trees and vegetation.

Nevertheless, in the square building blocks in the upper-left corner of the real image, it is apparent that the model successfully distinguishes height differences between horizontally and vertically oriented blocks.

Despite being purely trained on synthetic data, the model shows an ability to transfer its height estimation capabilities to real optical images, proving that synthetic datasets can be utilized for single image height estimation task on optical satellite images. Possibilities for enhancing the method to bridge the domain gap between real and synthetic data to improve performance on real optical images, are discussed in the Outlook section.

2.6 Outlook

Improving the accuracy of single-image height estimations primarily relies upon having the volume of available data as described in Section 2.1.4, as with any deep learning method. This project demonstrates the utility of synthetic data generation in addressing this issue. Nevertheless, in order to narrow the existing domain gap between synthetic and real data, there is a need to improve the synthetic data generation process. The primary shortcomings of the proposed method are:

- Limited Object Variety: The current can not generate diverse object categories encountered within real-world scenes. Notably absent are the representations of distinctive structures such as churches, landmarks, statues, bridges, stadiums, and factories, among others. Moreover, the method fails to generate numerous smaller objects frequently observable in satellite imagery, including vehicles, containers, street lights, electric and telephone poles, making it vulnerable to novel, unanticipated objects.

- Sparse Tree Density: In contrast to real scenes, our method features a considerably lower density of trees. This pronounced dissimilarity in the distribution of trees make synthetic and real scenes very dissimilar.

- Limited Tree Variation: The method relies exclusively on a solitary tree model, namely the pine tree, and maintains a narrow range of tree heights. This limitation hampers the model's capacity to adequately represent diverse tree species and their varying heights, a crucial aspect since vegetation often serves as a valuable indicator of the scale of nearby structures and objects.

- Lack of Detail: Objects generated within our synthetic scenes consist of basic shapes, failing to represent the rich variety that is in real world scenery.

Addressing these shortcomings will undoubtedly improve the effectiveness of the synthetic data generation process for improved height estimation accuracy. In future, this shortcomings may be addressed in two domains or in combination of both:

- Improvements in the 3D Domain: One plausible approach involves addressing these issues within the domain of 3D modeling. During the course of this project, the availability of the Google Maps Tiles API [9] has emerged. This API empowers software applications utilizing its capabilities to import high-quality and realistic city models from Google Maps. One such application is Earth Modeler for Blender

Figure 2.7: Model of London imported by the Earth Modeler empowered by Google Map Tiles API.

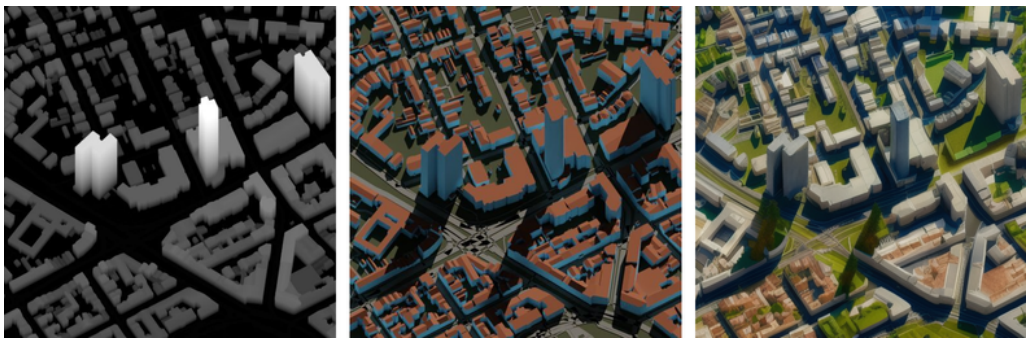

Figure 2.8: Example illustration of an experimentation of image domain manipulation idea by using Stable Diffusion [11] and ControlNet [12]. Resulting right image is created by height map on the left, non-textured basic 3D city model on the center and text prompts.

by Imagiscope [10], although its release status at the time of this report remains pending. For an illustration, please refer to the Figure 2.7 for an example model imported through this software as advertised by them.

- Improvements in the Image Domain: An alternative approach toward achieving greater realism revolves around the manipulation of resulting images, rather than focusing solely on 3D model improvements. Recent advancements within the generative artificial intelligence domain offer promising prospects in this regard. Leveraging techniques such as style transfer applied to satellite imagery on top of basic 3D models presents an opportunity to enhance the level of detail and realism within the output images. The Figure 2.8 provides an example of the potential of this approach.

3 Method

3.1 Problem Statement

Let I_1 and I_2 denote multi-temporal remote sensing images of the same area at the same spatial resolution and H_1 and H_2 denote georeferenced, rasterized height maps associated with their respective images. It is crucial to highlight that these height maps are estimated by a network, thus have inaccuracies in height estimations and suffer from shift and distortions, as anticipated. If H_1 and H_2 were perfectly aligned and correct height representations of I_1 and I_2, a direct pixel-wise comparison would suffice for change detection purposes. However this is not the case, this step treats height maps as input feature representations and further tries to distinguish changes on the ground from the changes stemming from other factors.

Furthermore, let I_1^i and I_2^i denote the i^{th} pixel in each image, representing the same position on the ground. Finally, let $y(I_1^i, I_2^i)$ be the ground truth for change detection in i^{th} pixel, can be described as:

$$y(I_1^i, I_2^i) = \begin{cases} 1 & \text{if } I_1^i \text{ and } I_2^i \text{ represent a change on the ground} \\ 0, & \text{if } I_1^i \text{ and } I_2^i \text{ represent no change on the ground} \end{cases} \tag{3.1}$$

Ideally, the objective of this project is to have a function f with parameters θ that estimate one change value per pixel i on the height maps, that can be defined as:

$$\theta = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{N} |f_\theta(H_1^i, H_2^i) - y(I_1^i, I_2^i)| \tag{3.2}$$

However, for this project, no annotated data for change detection is available, thus we don't have access to ground truth $y(I_1^i, I_2^i)$ to realize above objective directly. As a workaround, a self-supervised technique is employed.

3.2 Self-supervised Learning

Self-supervised learning is a machine learning paradigm, prominently applied in scenarios where labeled training data is scarce or costly to acquire. Unlike traditional supervised learning, which relies on explicit annotations for training, self-supervised learning autonomously supervises itself. As mentioned in Section 1, annotated data for multi-sensor or multi-modal change detection is costly to acquire or even unavailable, thus developing a self-supervised method is very important for this task.

To address the challenge of the unavailability of annotated pixel-wise ground truth data for change detection, the proposed method utilizes a dataset comprising pairs of height map patches. These patches are sourced either from the same geographical location and temporally aligned, hence assumed to be indicating no change on the ground, or from distinct locations and different time instances, indicating actual ground-level changes. These patches are automatically labeled in accordance with their origins. Process of data preparation is explained in detail in Section 3.3.1. Then, a self-supervised Siamese Patch Change (Dissimilarity) Estimation Model is trained on the prepared dataset with the objective of regressing a change (dissimilarity) score between pairs of height map patches. Utilization of Siamese networks for self-supervised dissimilarity learning is discussed below in Subsection 3.2.1

Finally, the trained Patch Change Estimation Model is applied to pairs of very large height map images using a sliding window approach. This process yields a densely populated change map, achieving pixel-wise change detection. Details of change detection process via the model is explained in detain in Section 3.4.

3.2.1 Siamese Networks

Siamese networks are a specialized class of neural network architecture specifically designed for learning similarity or dissimilarity between pairs of input data instances. The key characteristic of a Siamese network is that it consists of two identical sub-networks that share the same architecture and parameters. These sub-networks process each data instance in a pair independently and are trained to project the data into a feature space in such a way that similar instances are mapped closer together, while dissimilar instances are pushed farther apart.

Siamese networks have a wide range of applications, from face recognition to signature verification, and they are particularly relevant in areas where labeled data is scarce or costly to obtain. They play a crucial role in self-supervised learning by facilitating tasks like face recognition or similarity ranking and they are being ubiquitously used for change detection because of their capacity to learn meaningful and discriminative feature representations.[1] [2] [3] [4]

In this project, a Siamese Patch Change Estimation model is trained for the task of dissimilarity learning in between pairs of height map patches.

3.3 Patch Change Estimation Model

3.3.1 Data

For this project, a set of height maps estimated for various multi-temporal SAR images from different cities are available. For each available city, two height maps derived from SAR images captured within the same geographical area and in closest temporal proximity to each other are selected, as shown in Table 3.1. Since no ground is truth

Image	City	Mode	Orbit	Incidence Angle	Acquisition Date
bar112	Barcelona	ST	A	35	05.02.2023
bar239	Barcelona	ST	A	48	11.02.2023
ber452	Berlin	ST	A	30	17.09.2018
ber612	Berlin	ST	D	36	17.08.2018
fra104	Frankfurt	ST	A	21	19.11.2014
fra339	Frankfurt	ST	D	34	21.11.2014
muc131	Munich	HS	A	23	03.07.2010
muc136	Munich	HS	A	23	10.10.2010

Table 3.1: Acquisition details of image pairs assumed to represent no change.

available for change detection, an assumption is made that these image pairs represent no change on the ground. The reasoning behind this assumption lies in the fact that the images have been acquired in very close temporal proximity, thereby minimizing the likelihood of significant ground-level changes and attributing any observed variations to other factors.

Height Map Patches

In order to train a self-supervised Siamese network for dissimilarity learning we need a set of positive (indicating change on the ground) and negative (indicating no change on the ground) height map pairs. However the height maps estimated from the SAR images as elaborated in Table 3.1, as shown in Figure 3.1, are way too large to be processed by a model and cover extensive land area that makes assigning a single label of positive or negative regarding the change, infeasible. Thus, we need to sample smaller height map patch pairs from them, that are small enough to have a very low probability of a significant ground change when sampled from a negative pair. These patches also need to be small enough to be quickly processed by the model and to be used in a sliding window later for pixel-wise change detection of the whole height maps while large enough to be robust against changes stemming from other factors and provide meaningful amount of spatial context to represent features that reflect changes. It is empirically found that patches of size 128x128 are best suited for the training of the model.

Data-loader Preparation

Height map pairs derived from the image pairs presented in Table 3.1 are first geographically aligned. Subsequently, height map pairs are divided into a non-overlapping grid of 256x256 tile pairs. 25% of tile pairs are held back for a validation split for while the rest is reserved for training. During model training, each time the model requests for a sample, a negative or a positive patch pair request is assigned randomly at the same probability. In the case of a negative pair assignment, one 128x128 patch centered at a randomly selected pixel is extracted from both tiles of a 256x256 tile pair in training set. Thus a negative pair always represents the same exact location in very close temporal

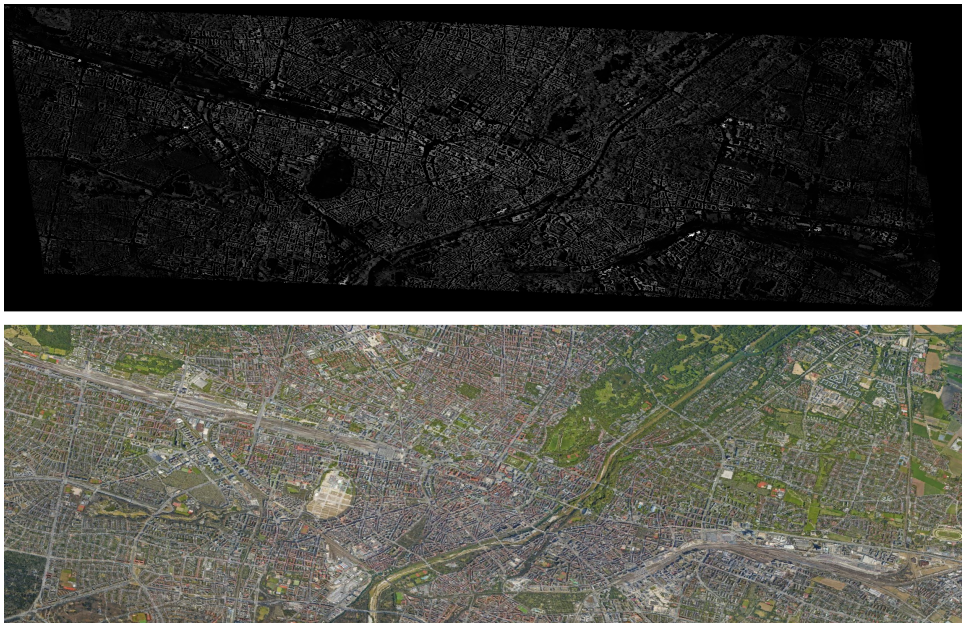

Figure 3.1: Top is the illustration of estimated height map of Munich from SAR image "muc131" in Table 3.1. Bottom is the image of the same area of around 60 km^2 on Google Maps.

proximity, assuming no change on the ground. Conversely in the case of a positive pair assignment, two 128x128 patches centered at independently random pixels are extracted from two totally random tiles from the training set. This way it is ensured that don't represent the same location, hence change on the ground is certain. Either way, both patches go through the augmentation step explained in the next section and bundled alongside a autonomously assigned binary label indicating change, 0 for a negative pair and 1 for a positive pair. Patch pair and the label is passed to the model as a sample. While the randomness introduces variation over epochs during training, tile split ensures prevention of information leak from training set to validation set.

Data Augmentation

Data augmentation is a technique used in machine learning to increase the diversity of a dataset by applying various transformations to the existing data. Introducing variations that are imitating the fluctuations in the data helps making the model robust against them. In the scope of training Patch Change Estimation model, two augmentation transformations are applied to each individual height map patch to address two challenges working with height estimations:

1. Random Translation: Despite being geographically aligned, estimated height map patches suffer from shift in between negative pairs which is caused by re-projection error in height estimations. In order to make the model robust against shift, each

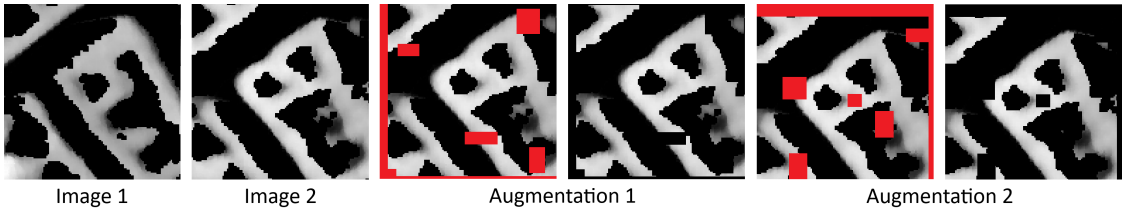

Figure 3.2: Visualization of example augmentations. Image 1 and Image 2 are an example negative patch pair. Two example augmentations for Image 2 are illustrated. For each augmentation example, left image highlights the transformed parts in red while right image shows the result after transformation.

height map patch is randomly translated by up to 10 pixels in both directions of X and Y axes, while keeping the patch size same and the center invariant. This helps the model to neglect small global shift in the patch regarding change detection. Furthermore, keeping the center invariant results in information loss on the edges, which suggests the model to discount edges compared to the center. This helps ignoring the case of a structure in the edge of one patch being outside the other patch due to the re-projection error.

2. Random Erasing: Great amount of pixels within height map patches have no height value, mainly caused by shadowing effect in SAR images. These areas with no height value often cover more than half of a patch. Larger shapes formed by such areas are great indicators of ground-level change while many smaller no value areas are just due to image acquisition and estimation challenges. Pixels within these areas are filled with zeros. To make the model robust against fluctuations of these smaller areas, random erasing transformation is applied. For each patch, five rectangles of random sizes ranging from 0.5% to 2% of the total area of the patch with the aspect ratios ranging from 0.3 to 3.3 at random locations are erased from the patch. These values are selected to better imitate small fluctuations of such areas stemming solely from insignificant noise, while preventing good amount of information in a patch.

Examples regarding augmentations on height map patches are visualized in the Figure 3.2.

3.3.2 Model Architecture

Being a Siamese network, the Patch Change Estimation Model consists of two identical sub-networks with shared weights, as shown in Figure 3.3. The objective of each sub-network is to create a refined embedding of a given height map patch that is representative of information significant to assessing it's dissimilarity to other embeddings regarding ground-level structure of the patches.

Sub-networks consist of back-to-back blocks of convolution, ReLU activation and max

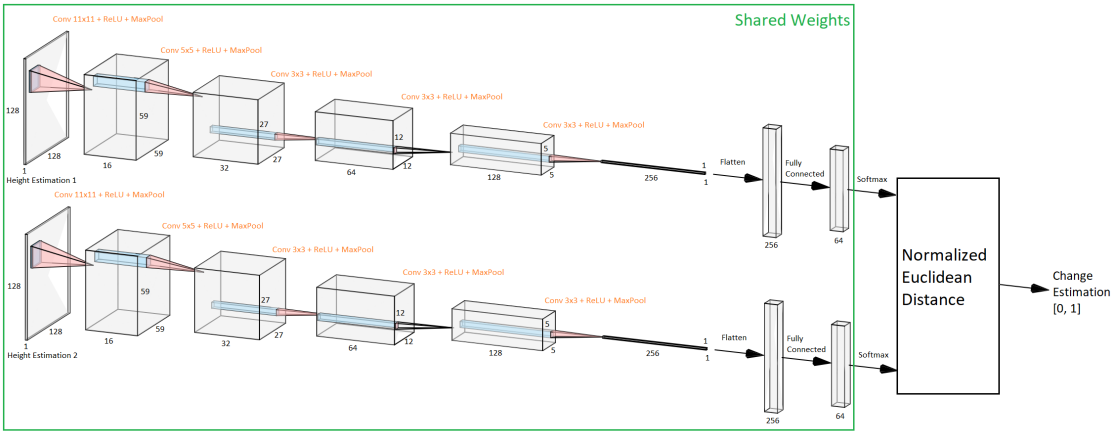

Figure 3.3: Overview of the Patch Change Estimation Model architecture. Top and bottom sub-networks are identical and share the same weights. Input to each sub-network is a single channel height map of size 128x128. Model estimates a single change estimation for a pair of height maps, while learning 64 dimensional embedding vectors.

pooling layers, reducing spatial resolution to 1x1 while increasing the channel depth to 256. After flattening the 1x1 filters into 256 nodes, a fully connected layer is applied to reduce embedding dimension to 64. Finally, softmax layer is applied, resulting in an embedding vector that is essentially a probability distribution over 64 classes. One benefit of applying softmax is to have all values sum up to 1, which makes normalizing the distance in between vectors easier. Another benefit is the regularizing effect while pushing embeddings of positive pairs apart from each other, due to the fact that each element in the embedding vector is also converted into (0,1) range.

3.3.3 Training

Let s_1 and s_2 be the 64 dimensional output vectors of height image patches h_1 and h_2 after the softmax layer. Then, normalized Euclidean distance in between these vectors can be expressed as:

$$p = \frac{1}{2} \left\| s_1 - s_2 \right\|_2 = \frac{1}{2} \sqrt{\sum_{i=1}^{64} (s_{1i} - s_{2i})^2} \tag{3.3}$$

where i denotes dimension. $\frac{1}{2}$ normalizes p into [0, 1] range as Euclidean distance in between s_1 and s_2 can be at most 2 since the output after the softmax is essentially a probability distribution over 64 embedding categories where each component is a probability in range (0, 1) and the components add up to 1. p is interpreted as probability of change in between the patches, treating the problem as a binary classification task.

Thus the model is trained via Binary Cross Entropy (BCE) as a loss function. This loss function can be expressed as:

$$\text{BCE} = -(y \log(p) + (1 - y) \log(1 - p)) \tag{3.4}$$

where y is the binary ground truth label which is either 0 (no change) or 1 (change) depending on the pair. This way, loss assumes the role of a discriminative mechanism and model learns to minimize the distance between output vectors that belong to a positive pair, while simultaneously maximizing the distance for negative pairs.

Training is executed over 40 epochs with 1068 training and 356 validation pairs per epoch with a batch size of 1. Patch pairs from validation set are not seen during training and are used for early stopping mechanism. Augmentation step is only applied to patches from training pairs, introducing novel variation and act as a regularization to prevent over-fitting. For optimization, Adam optimizer with a learning rate of 0.001 is used.

3.4 Change Detection

The Patch Change Estimation Model offers a valuable metric for assessing change/ dissimilarity when comparing two small patches of height maps. Nevertheless, as stated in Section 3.1, the primary objective main objective is to achieve pixel-wise change detection across complete height maps. In pursuit of this goal, the trained model is deployed using a sliding window approach over the height maps.

One approach involves dividing the height maps into a grid of patches of size 128x128 for comparison with the model, subsequently fusing the change values to construct a change map. However, such a map would exhibit very low resolution, yielding a singular change value for an area covering multiple buildings.

To enhance the resolution of the estimated change map, an alternative method is employed by sliding a window of size 128x128 over both height maps for comparison. Height map patches under these sliding windows are then processed by the trained model to estimate a change score. Subsequently, the window gets shifted by 16 pixels, and the process is reiterated, similar to a convolution operation in image processing. However, in this case, it is a deep learning model rather than a kernel matrix that convolves over the image with a stride of 16. The outcome is a matrix of change values, corresponding to the change within the sliding window positions.

The matrix of change values undergoes an upscaling process via interpolation to restore it to the original dimensions of the input height maps. The application of nearest neighbor interpolation would yield a change map where each 16x16 pixel area retains the same change value, as determined by the stride value. However, for the final outcome in this project, bilinear interpolation is used as the upscaling method to yield smoothed continuous values, thus achieving pseudo pixel-wise results for change detection.

4 Results

Proposed method is tested on two height map pairs, estimated from the SAR images obtained from London, Frankfurt and Munich. Table 4.1 shows the acquisition details of these images. These images are selected for testing to ensure that they exhibit significant ground-level changes, since there is years of temporal gap in between them.

Figure 4.1 illustrates example results of the Patch Change Estimation Model on negative and positive patch pairs.

Figure 4.2 shows various change maps for the London test height map pair, demonstrating the results obtained with the approaches mentioned in the Section 3.4. Figure 4.3 shows four example highlights of change detections in Frankfurt and Munich, magnified for clearer view.

For demonstration purposes, in Figure 4.2 and 4.3, colored change maps are blended over the buildings of height maps. Minimum change value threshold for the color ramp is set to 0.4 better highlight significant changes.

Image	City	Mode	Orbit	Incidence Angle	Acquisition Date
lon533	London	HS	A	23	29.04.2011
lon504	London	ST	A	37	05.05.2020
fra339	Frankfurt	ST	D	34	21.11.2014
fra903	Frankfurt	ST	A	47	18.02.2023
muc749	Munich	HS	D	49	18.06.2008
muc121	Munich	HS	A	37	28.01.2023

Table 4.1: Acquisition details of image pairs to test the proposed method.

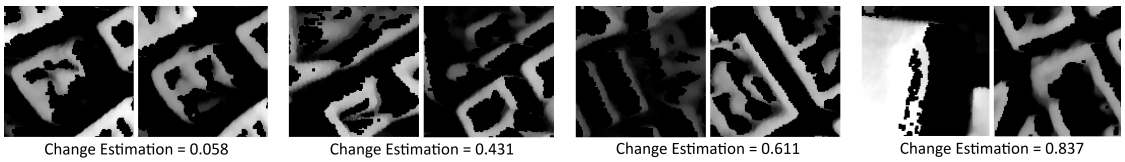

Figure 4.1: Estimated change values on pairs of patches. Two pairs on the left are negative pairs while two pairs on the left are positive pairs.

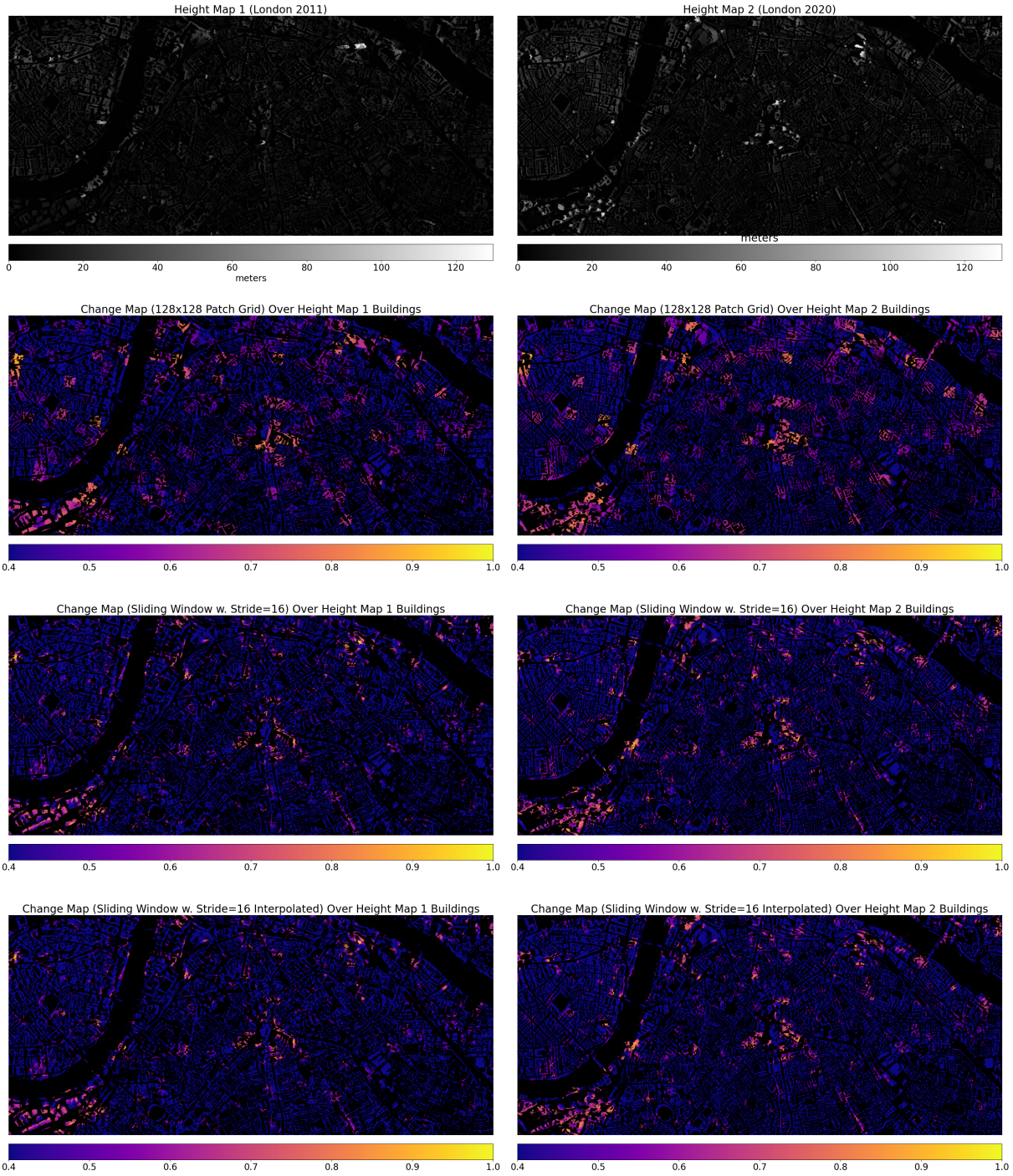

Figure 4.2: Illustration of change detection results in London image pair. In the first row, left column shows the height map from London in 2011 while right column shows the height map of the same area in 2020. Second row shows the change map acquired from applying the model in a grid approach. Third row shows results of the sliding window approach with a stride of 2. Finally the bottom row shows the results after bilinear interpolation.

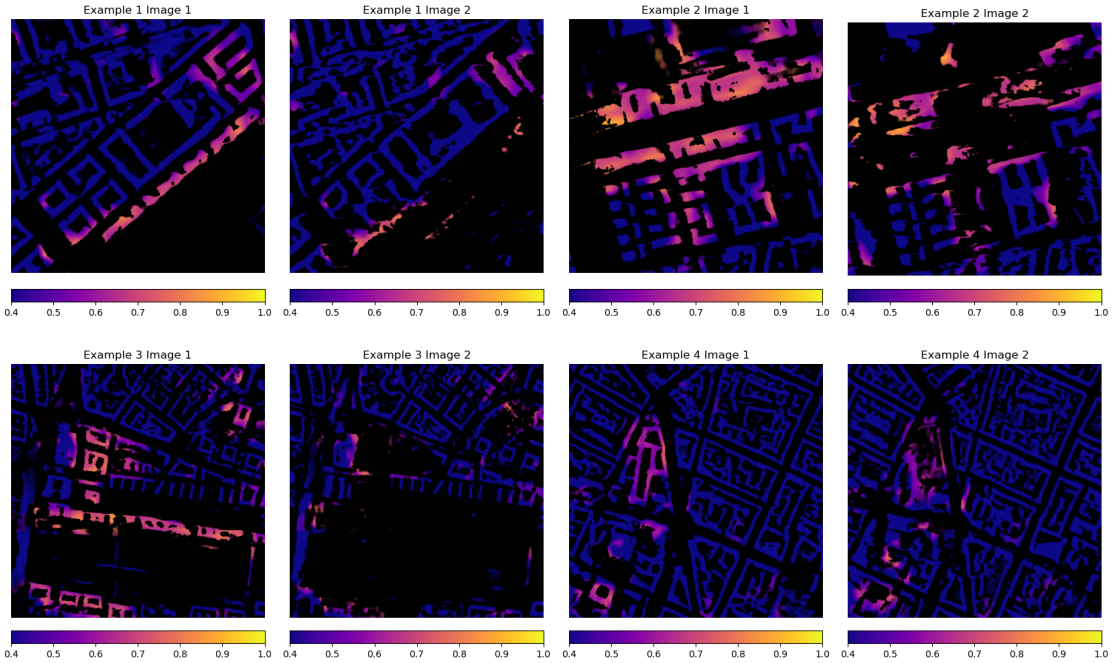

Figure 4.3: Examples from change detection results of Frankfurt (top) and Munich (bottom).

5 Discussion

The Patch Change Estimation Model provides a valuable metric for estimating when comparing among patches of height maps as illustrated in Figure 4.1. In the initial example, a notably low change score of 0.058 is adeptly estimated for a negative pair of patches, effectively mitigating minor noise in the height estimations, altering the appearance of the patches. Likewise, in the second example in another negative pair, change score remains below 0.5, despite strong disparities in the appearance of the patches, mainly caused by variations in image acquisition configurations. Examples of two positive pairs are both estimated above 0.5, successfully representing changed scenes.

As depicted in the second row of Figure 4.2, the application of the model to the height map tiles arranged in a grid generates a change map that already holds significant information. Although localized very coarsely, this map adeptly identifies regions exhibiting potential changes. In contrast, the proposed change detection methodology, which employs the model with the sliding window approach, yields a change map with enhanced accuracy in pinpointing the precise locations of changes, extending down to the level of individual buildings, as highlighted in more examples in Figure 4.3. Method excels at detecting demolished structures or severely changed building footprints, while staying robust against small perturbations and re-projection shift in the height estimations. On the downside, it detects considerable amount of false positives of various sizes throughout the map. Larger ones among these false positive detections can be primarily attributed to skyscrapers or other tall buildings and vegetation heavy areas which are weaknesses of height estimation model as well.

The final stage of the bilinear interpolation transforms the discrete boundaries within the map into continuous gradients. This process adds a smoothing effect to the change map, and provides pseudo pixel-wise estimations. Notably, using a stride value smaller than 16 for the sliding window operation introduces a significant degree of noise into the change estimations. This renders going further down to a stride of 1 with the goal of pixel-wise estimations totally unfeasible. Hence, interpolation emerges as a viable alternative in this scenario.

6 Conclusion and Outlook

In this project, potential applications of single height estimations for change detection across multi-sensor images, with a motivation mainly rooted in the recent success of SAR2Height network, are explored. Towards this objective, a proposal made to address the challenges posed by the lack of annotated data for change detection and the unavailability of extensive optical data for the extension of this concept to multi-sensor images. The proposed approach introduces an optical height estimation method, investigating the feasibility of leveraging synthetic optical data, while concurrently presenting a self-supervised method for change detection, harnessing the benefits of height estimations.

While the optical height estimation method delivers results that fall short of the performance achieved with SAR image-derived height estimations on real data, thereby limiting the extension of the concept to multi-sensor images, it effectively illustrates the potential of synthetic data in the training process. By implementing an improved synthetic data generation method, as discussed in Section 2.6, a similar approach holds promise for producing results that can be integrated alongside SAR estimations for multi-sensor change detection.

On the change detection side, proposed idea employing a Siamese Patch Change Estimation Model achieved yielding a dense change map from height estimations, detecting significant ground-level changes while staying robust against minor perturbations in the height estimations. Some future research cues for further improving the model can be:

- Advancing data augmentation techniques beyond simple translation and erasing, to achieve better generalization and robustness.
- Integration of hard positive and hard negative pair samples in the training data. Hard negative pairs are dissimilar images with no ground-level changes, that acquired at the same location and at a similar time but in strongly different imaging configurations such as various orbits, acquisition modes and incidence angles. Hard positive samples on the other hand are very similar images consisting of subtle ground-level changes that need to be detected.
- Jointly training the height estimation models with a more sophisticated change estimation model to enable the height estimation models to fine-tune themselves with the objective of enhancing change detection.

As of this project, proposed method is able to successfully detect changes on the ground down to the level of individual buildings with a reasonable accuracy and recall by providing

pseudo pixel-wise estimations. It is worth noting, however, that these estimations are still very distant from achieving genuine pixel-wise outcomes, in terms of certainty of estimations of each individual pixel. Nevertheless, the utilization of the sliding window approach successfully advances the method closer to the ideal case in a self-supervised fashion, without the need of an annotated change detection data.

Bibliography

[1] R. C. Daudt, B. L. Saux, and A. Boulch, *Fully convolutional siamese networks for change detection*, 2018. arXiv: 1810.08462 [cs.CV].

[2] F. Rahman, B. Vasu, J. Van Cor, J. Kerekes, and A. Savakis, "Siamese network with multi-level features for patch-based change detection in satellite imagery," in *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, IEEE, 2018, pp. 958–962.

[3] Y. Chen and L. Bruzzone, "Self-supervised change detection in multiview remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2022. DOI: 10.1109/tgrs.2021.3089453. [Online]. Available: https://doi.org/10.1109%2Ftgrs.2021.3089453.

[4] S. Saha, P. Ebel, and X. X. Zhu, "Self-supervised multisensor change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–10, 2022. DOI: 10.1109/tgrs.2021.3109957. [Online]. Available: https://doi.org/10.1109%2Ftgrs.2021.3109957.

[5] M. Recla and M. Schmitt, *Deep-learning-based single-image height reconstruction from very-high-resolution sar intensity data*, 2021. arXiv: 2111.02061 [cs.CV].

[6] *Blender-OSM*, https://prochitecture.gumroad.com/l/blender-osm, 2023.

[7] *OpenStreetMap*, https://www.openstreetmap.org, 2023.

[8] L. Mou and X. X. Zhu, *Im2height: Height estimation from single monocular imagery via fully residual convolutional-deconvolutional network*, 2018. arXiv: 1802.10249 [cs.CV].

[9] *Google Map Tiles API*, https://developers.google.com/maps/documentation/tile/overview, 2023.

[10] *Imagiscope Blender Earth Modeler (To be released)*, https://youtu.be/80GrFXFOayE?si=3CHEdp2KuY76lwVF.

[11] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, *High-resolution image synthesis with latent diffusion models*, 2022. arXiv: 2112.10752 [cs.CV].

[12] L. Zhang, A. Rao, and M. Agrawala, *Adding conditional control to text-to-image diffusion models*, 2023. arXiv: 2302.05543 [cs.CV].