ISPRS Journal of Photogrammetry and Remote Sensing 211 (2024) 104–120

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

The SAR2Height framework for urban height map reconstruction from single SAR intensity images

Michael Recla *, Michael Schmitt

Department of Aerospace Engineering, University of the Bundeswehr Munich, 85577 Neubiberg, Germany

ARTICLE INFO

Keywords:
Single image height estimation
Synthetic Aperture Radar (SAR)
3D reconstruction
Radargrammetry
Urban areas
Deep learning

ABSTRACT

Recently, it was shown that a detailed reconstruction of urban height maps is possible from single very high resolution (VHR) synthetic aperture radar (SAR) images with deep convolutional neural networks. Being merely a proof-of-concept so far, the potential of this approach has not been fully exploited yet. With this work, we present an optimized deep learning model for height estimation from single VHR SAR images, which incorporates sensor knowledge into the estimation. We embed this model into a SAR-specific processing chain that allows the generation of seamless georeferenced digital surface models (DSMs) with geodetically defined heights in an orthometric map coordinate system. Extensive experiments are carried out with a custom-generated dataset including over 50 TerraSAR-X images from 8 different cities. They confirm that our workflow generalizes well across different locations while being robust to different properties of the input data. Thus, our workflow provides the unique ability to produce elevation models of urban areas quickly, regardless of weather, around the clock, and at low cost. This can be of immense benefit when time is critical, e.g. in disaster response scenarios or in the context of reconnaissance activities.

1. Introduction

Measuring elevation data is a time-consuming and expensive task. Urban areas in particular pose a challenge with their dense complex man-made structures. Moreover, it is also just these zones that are subject to the most and fastest changes. At the same time, these changes affect the most people, since cities are where the majority of the population lives. Thus, efficient, fast, and cost-effective methods for a regular update of urban topography data bear great relevance, as different stakeholders could use them in many applications, e.g. change monitoring. Since most conventional 3D reconstruction methods, e.g. LiDAR, photogrammetric stereo, InSAR, or TomoSAR require pairs or even stacks of images, a method relying only on single images would lower the required acquisition time and data costs significantly. A single satellite overflight would be able to provide an updated urban height map for a detailed 3D analysis of the situation, e.g. after a disaster.

1.1. Related work

Pioneering and well-cited works in this context are suggesting single-image height reconstruction from optical aerial images (Amirkolaee and Arefi, 2019; Ghamisi and Yokoya, 2018; Mou and Zhu, 2018). Since their publication, ever more sophisticated model architectures and deep learning paradigms for depth estimation from the field of computer vision spilled over into the world of remote sensing and height estimation, driving up the performance scores (Li et al., 2022; Sun et al., 2022a; Xing et al., 2022; Karatsiolis et al., 2021; Liu et al., 2020). While being very interesting and strong methodological approaches, most of these works still use the same small-scale datasets and sensor modalities, optical orthophotos, or nadir-viewing satellite imagery, consisting of two to at most four different scenes. Train and test sets are drawn from the same scenes so that the distribution of the test data strongly resembles the one known from training. Although the methods can be compared among each other, such a practice distorts the achieved error metrics compared to an application using an actual unseen test area and thus being closer to a real-world scenario. Also, using true orthophotos for testing is generally pointless from a practical point-of-view, as the elevations of the scene are already a prerequisite for their generation in the first place. Overall, the methods often lack a certain practicality and awareness of remote sensing-specific problems, such as spatial generalization capability, the mosaicking of small model outputs into coherent, seamless maps, or accurate georeferencing of fairly raw sensor data.

A substantial leap in terms of the amount of data used for training was provided by Cao and Huang (2021). Panchromatic multi-view data (forward, nadir, backward) and multispectral data were collected over 42 Chinese cities, from which building footprints and heights are jointly

* Corresponding author.
E-mail addresses: michael.recla@unibw.de (M. Recla), michael.schmitt@unibw.de (M. Schmitt).

https://doi.org/10.1016/j.isprsjprs.2024.03.023
Received 14 September 2023; Received in revised form 28 March 2024; Accepted 28 March 2024
Available online 8 April 2024

derived within a multi-task network. The method was then applied to U.S. cities as well. With the Chinese Ziyuan 3-03 satellite, all the necessary data can be acquired in one satellite pass. Testing was still done in a random split, but the method bears the potential to be applied in practice. Another interesting choice of sensor for height estimation was taken by Müller et al. (2023), who used Sentinel-2 data as input, creating a theoretically global system with high temporal coverage. The results are surprisingly good considering the coarse resolution of the input data. Some tweaks such as attention gates in the skip connections and parallel encoders with different sized receptive fields are used to achieve these results. While it is only tested on a small scale as well, it holds potential for large-scale applications. Regarding the problem of only patch-wise relative estimates, for instance, Amirkolaee and Arefi (2021) separate object pixels from ground pixels in a rule-based manner and place these relative height estimates on an SRTM-DSM as an approximated DTM, thus creating a coherent higher-resolution DSM. Another nice remote sensing-specific turn in the methodology was done by Chen et al. (2021) which estimate heights per building instance instead of per pixel by using an adapted form of Mask R-CNN in which a height prediction head is incorporated into its region proposal network. For spatially lower-resolution data, instance-wise height estimation can often be more reasonable. Li et al. (2023) also estimates heights at a building level. Very interesting is the paradigm introduced to determine the heights not by a direct regression, but by a detour over 3D centripetal shift representations. These represent the distances between the visible roof points to the visual center of the building at ground level. From these, the corner points can also be estimated and thus the footprints be determined.

What all these methods have in common, however, is the use of optical or multispectral data as inputs. The related work using SAR as input data remains very scarce to this day. This is probably due to the unfamiliar imaging geometry of SAR and its more difficult coregistration with other geospatial data. At the same time, however, more and more commercial actors are entering the SAR business, which has recently given the technology a fair amount of hype. SAR is offering a set of advantages as an active system. It can be used at any time of day or night independent of weather conditions, making it a perfect system for reconnaissance and disaster management. Some works address forest height estimation (like Zhang et al. (2022b)), or the estimation of large-scale mountainous areas using L- and C-band, like Xue et al. (2022), but the reconstruction of urban areas with just a single image is very rare, with our earlier publication (Recla and Schmitt, 2022) being the first of its kind to the best of our knowledge. In this work, we showed that it is generally possible to estimate patch-wise relative heights from slant range SLC SAR intensity data using a dataset consisting of four images from two cities. Sun et al. (2022b) take a different, higher-level approach and use building footprints from OpenStreetMap and locate a bounding box around the visible echo of the corresponding facade of each building in the SAR image. Through the length of these facades, the heights of the buildings can be determined.

1.2. Contributions

In this work, we build upon (Recla and Schmitt, 2022) and develop a fully operational framework for the generation of geodetic height maps of urban areas from single VHR SAR images solving several remote sensing-specific challenges. For that purpose, we propose an enhanced neural network architecture, which takes physical auxiliary knowledge into account. In addition, we complement the neural network with SAR-specific pre- and post-processing steps so that eventually not only relative heights are generated in a patch-by-patch manner, but geodetically meaningful, seamless, and coherent height maps in a projected orthometric reference system are provided. This enables the derivation of the structure of an urban area from a single SpotLight scene, quickly and around the clock, with high accuracy.

1.3. Paper structure

The remainder of this paper is structured as follows: Section 2 contains the description of the methodology, which consists of the necessary preprocessing steps, the description of the used Deep Learning model and its training, and the postprocessing, which consists of the mosaicking process, the conversion from slant range data to ground range and some filtering techniques of the result. In Section 3, the data used is described and the extensive experimental results are summarized, including ablation studies and cross-validation. Finally, the results are discussed and put into perspective in Section 4, before the findings are summarized in Section 5.

2. The SAR2Height framework

The overall workflow of the SAR2Height Framework is depicted in Fig. 1. It can be divided into a training phase, in which the parameters for the deep convolutional network residing at the core of the framework are determined, and an inference phase, which uses the trained model and applies a series of necessary postprocessing steps to its results to generate a coherent DSM in an orthometric reference system. The framework consists of different submodules for each of these steps which will be described in the following.

2.1. Data preparation

To run the SAR2Height processing chain, several steps are necessary to prepare the required data. To apply the method, calibrated and resampled SAR data serves together with its approximative looking angle as input, while a coarse-resolution terrain model is necessary to georeference the result. The preprocessing steps for these data types are described in the following.

The complex pixel values of VHR SAR level 1B data are calibrated with the calibration factor from the metadata and converted into a logarithmic scale. The values obtained in dB correspond to the so-called radar brightness β_0 and are approximately normally distributed. The aim of the calibration is to counteract different recording geometries and to provide the model with data as homogeneous as possible. A further step to meet this goal is histogram matching. From all the data available during training, a common histogram is created. Each further image to be evaluated is then matched to this histogram. The core of the process of histogram matching is generating cumulative histograms for both the actual and reference images, which are then used to identify the unique pixel values in the reference image that most closely correspond to the quantiles of the distinct pixel values present in the image to match. Linear interpolation is utilized to carry out this mapping between the two sets of values. It should be noted that this histogram matching step only provides value if the areas under consideration are of a similar nature. If, for example, a large part of the area is covered with water, histogram matching would even be detrimental to the intention. However, if, as is the case here, it is exclusively similar urban regions, this procedure can help to make images from different viewing angles and/or other sensors more comparable. Subsequently, the slant range data is resized to a square pixel size on the ground of 1 m. This scaling is performed with the mean projected pixel spacings over the entire scene and thus corresponds only to an approximation, which is however sufficient for mostly flat urban space.

As auxiliary data, the model uses the approximate local looking angle of the image section shown as an additional parameter. The central pixel of the patch is geolocated on a globally available, coarse-resolution digital terrain model, giving us the target's position $\vec{x_T}$. Together with the position of the sensor $\vec{x_S}$, the local looking angle θ can be determined by

$$\theta = \pi - \arccos\left(\frac{\vec{x_S}\cdot\left(\vec{x_T}-\vec{x_S}\right)}{|\vec{x_S}|\cdot|\vec{x_T}-\vec{x_S}|}\right). \tag{1}$$

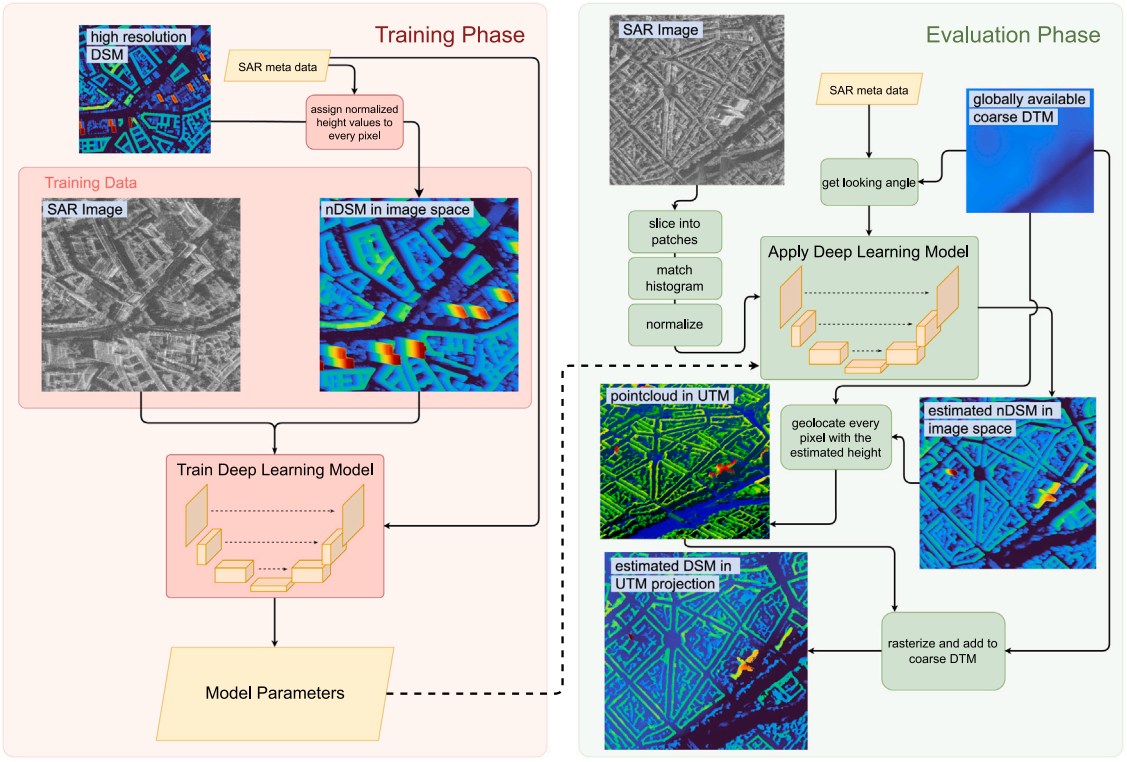


Fig. 1. The workflow of the entire SAR2Height framework can be divided into a training phase (left) and an evaluation phase (right). For training the model, high-resolution nDSMs are projected into the image geometry of the SAR images. These pairs of SAR and relative height are used to train the model in a supervised manner. The so-learned model parameters are then used in the evaluation phase to interpret an unseen SAR image and estimate its pixel-wise heights. A globally available coarse-resolution terrain model is used on the one hand to determine the local looking angle as an auxiliary parameter for the model, and on the other hand to reproject the estimated nDSM in slant range into an orthometric (projective) coordinate system (like UTM). The resulting point cloud is filtered and rasterized to finally obtain the estimated DSM in ground range.

This angle is then converted into a metric measure related to the length of the layover that appears in the image by applying the cotangent. This should help the model to distinguish whether longer appearing layovers are coming from taller buildings or are a consequence of a different viewing geometry.

Lastly, for the transition from normalized elevation values coming from the deep learning model to geodetic absolute elevations after post-processing, a digital terrain model is necessary. Due to the tendentially low-frequency characteristics of bare topography, a coarse-resolution terrain model is already sufficient for our purpose (<30 m). Such datasets are globally (and in some instances freely) available. However, a potential error in these elevation data will directly affect the final geoproduct. It is important to note that all heights used here have to be geometric heights (if desired, a geoid model can be reintroduced to the final DSM).

2.2. CNN-based single-image height estimation

The core of the approach presented here is the Deep Learning model. It is the part where the actual estimation of the height above ground value for each pixel of a SAR backscatter intensity image takes place. This represents a supervised regression problem and is achieved with a modified form of a U-Net (Ronneberger et al., 2015). The architecture and the training strategy are described in the following.

2.2.1. Generating training data

This section describes the process of obtaining ground truth data for the training process since we are dealing with a supervised method. For the plain application of an already trained model, these steps are of course omitted. For the annotation of the SAR images with elevation values, a high-resolution surface model (DSM) and terrain model (DTM)

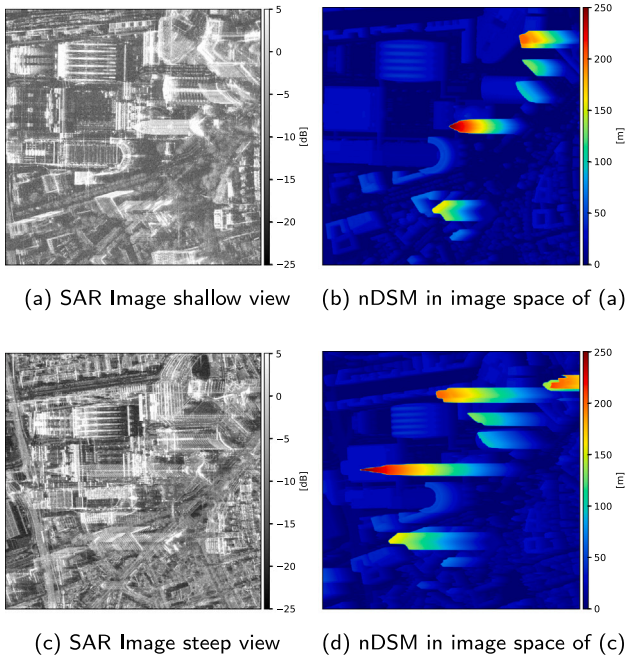
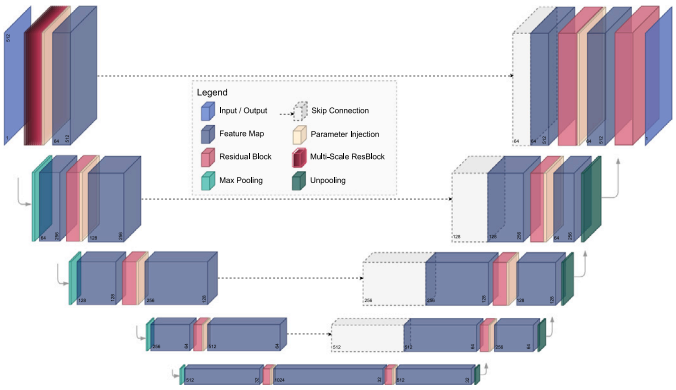
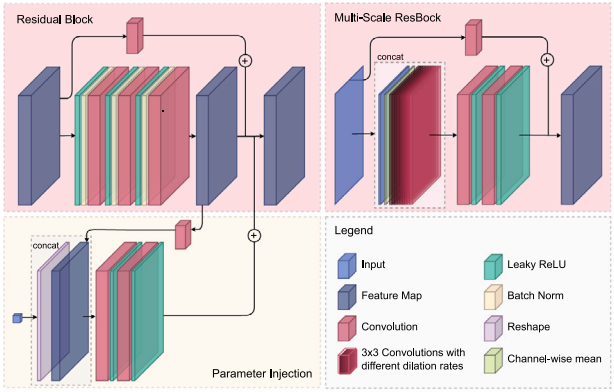


Fig. 2. The lines each form a matched pair of SAR image and nDSM projected in its image geometry. The image at the top was acquired at an incidence angle of 47°, which corresponds to a very shallow viewing setting. The bottom row shows an image with an incidence angle of 21°, which is already close to the TerraSAR-X lower operational limit. It can be seen that one and the same building produce layovers of very different lengths.



(a) U-Net-like network architecture "SAR2Height". In three points of the network, the looking angle is injected as an auxiliary parameter. The "Multi-Scale ResBlock" at the beginning broadens its receptive field. Instead of upconvolutions as in the original version of U-Net, unpooling using the indices from the encoder is applied to restore the resolution of the input.



(b) The conventional ResBlock in the upper left, the Parameter Injection Block in the lower left and the Multi-Scale ResBlock in the upper right. These three represent the basic building units of the network architecture.

Fig. 3. Overview of the model used for estimating the relative heights per pixel in slant range intensity SAR images. The right figure describes the blocks used, the left shows where they are applied in the network. The input tile of size 512 × 512 pixels is reduced to a size of 32 × 32 in the encoder, which in turn is restored to the original dimensions in the decoding part of the network.

are required. By subtracting the DTM from the DSM, the normalized surface model (nDSM) is created, whose height values can then serve as our target during the training. In essence, nDSMs contain heights above ground, i.e. they represent the relevant heights of elevated objects such as buildings, masts, trees, etc., which we seek to reconstruct eventually. These heights are then projected into the image space of the SAR acquisition in question. This process uses the sensor model defined in the metadata to locate each pixel on the surface model using geometric relations. Basically, a high-resolution surface model is intersected with the respective Zero-Doppler plane per image line, resulting in a two-dimensional terrain profile (or slice) for each image line. All point targets of the pixels of one image line must lie on the corresponding terrain profile. By intersecting these profiles with the range circles of the individual image columns, the position and height of the possible scatterers are obtained for each of the slant SAR pixels. In the case of layover effects occurring, where multiple scatterers from different locations (e.g. ground, wall, roof) fall within one resolution cell, the highest lying point is projected, for instance, the one from the roof. However, the pixels are not examined for visibility, which means that heights are also predicted in the radar shadow. These will be filtered out in postprocessing, see Section 2.3.3. For a detailed look at this annotation process, refer to Recla and Schmitt (2022). In Fig. 2 you can see two examples of a height-annotated image pair used for training the model.

2.2.2. Network architecture

For our earlier proof-of-concept presented in Recla and Schmitt (2022), we relied on the U-Net-like `IM2Height` architecture (Mou and Zhu, 2018) with some minor tweaks. In this work, we present the `SAR2Height` architecture, which again belongs to the U-Net family, but contains several enhancements specifically designed for height reconstruction from single VHR SAR images. Fig. 3(a) shows the network architecture. The single-channel input tiles of size 1 × 512 × 512 pixels are brought by max-pooling operations in each of the four consecutive stages of depth to a spatial dimension in the bottleneck of 1024 × 32 × 32. In the decoder, these feature maps are then converted back to the output size by unpooling, as it is done by Mou and Zhu (2018). Unpooling describes the reverse process of (max) pooling by restoring the spatial dimensions of the feature maps. In this process, the pooling indices from the encoder are used again to place the values at their original position. Also, the feature maps from the encoder are reintroduced to bring back the spatial details from the earlier stages of the network. The model is based on what is known as a residual block (originally introduced by He et al. (2016)), which is used in every level of the network. Compared to the original ResBlock, the ones used here consist of one more convolutional layer and a reversed arrangement of activation, batch norm, and convolution (compare Fig. 3(b) upper left). The residual blocks help to robustly train a deep network by learning only a residual part of the function instead of a complete mapping. This greatly reduces the problem of vanishing gradients.

The entire model is built on 3 × 3 convolutions. The maxpooling operations increase the actual receptive field, but this happens only in deeper parts of the network. In order to allow the model to get a more global impression of the scene also at an earlier stage, a so-called Multi-Scale ResBlock is introduced. The Multi-Scale ResBlock is shown in Fig. 3(b), top right. It consists of a sequence of dilated convolutions with different dilation rates (namely 1, 2, 3, 4, 6, 8, 16, 32 and 64). The feature maps from these dilated convolutions are concatenated with the input and a channel-wise mean value. They are then scaled to the size of the output in two activated convolutional layers and finally added to the input in the sense of a ResBlock. While using such high dilation rates it is important to set the padding mode in the convolutions to *reflect* to avoid distorting the result towards a fill value. The idea of using atrous convolutions is inspired by *DeepLabv3* (Chen et al., 2017).

Another particular feature of the model used here is the utilization of sensor parameters as auxiliary knowledge. In the age of data-driven models, this is often forgotten and missed potential, especially when it comes to remote sensing data. In particular SAR sensors, as active systems, provide a wealth of very precisely known and well-calibrated metadata. For a lot of problems in Computer Vision, all the necessary information lies in the pure pixel values of the input image (e.g. for classifying a cat). However, a land use classifier, for example, could strongly benefit from the knowledge of the position of the image in question (information that is contained in every satellite image and mostly ignored). A noteworthy instance can be found in the work of Zhang et al. (2022a), wherein the authors leverage not only a multispectral image but also its associated longitude and latitude coordinates. This additional spatial information speeds up the training process of a land use classification system while also improving its

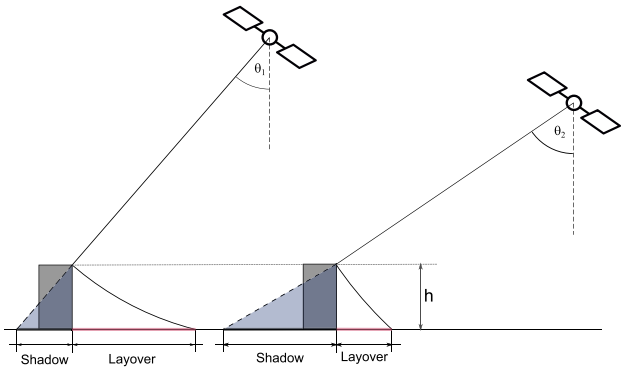


Fig. 4. Visual comparison of the influence of the viewing angle on the length of the layover and the shadow. A smaller looking angle results in a longer layover, but a shorter shadow.

generalization capabilities, even when limited training data is available. To enable the fusion of pixel feature vectors and spatial details within the model, a dual-branch network is employed to ensure that both components are appropriately aligned to a uniform size so that they can be stacked together.

In the context of height estimation, as discussed here, incorporating supplementary sensor data can prove highly beneficial, too, as the primary indicators of height in a slant range SAR image are likely to be the geometric phenomena of layover and shadow, which directly correlate with building or target height. Fig. 4 illustrates how the extents of layover and shadow regions are influenced by varying viewing angles. When a model is provided with a bare SAR image devoid of any supplementary details, an elongated layover effect may originate from either a taller structure or a more acute viewing angle. In this regard, compare Fig. 2 left and right. A steep viewing setting leads to very long layover artifacts, but the building in question keeps, of course, the same height. If prior knowledge regarding the viewing angle is supplied, the model's determination of the height of the building in question can be considerably simplified. Integrating additional data in the form of scalar numbers (like the looking angle) into a U-Net, as the one in the case at hand, can be accomplished in various ways, although it is comparatively rarely described in the existing literature. In Kang et al. (2021), the authors employ a small fully connected subnetwork to adapt the scalar input to the spatial dimensions of the feature maps in the bottleneck. However, this approach has the limitation that the dimensions of the input image cannot be freely chosen after training, as the number of nodes in the subnetwork would also need to change accordingly. To overcome this constraint and retain the flexibility of accepting arbitrarily shaped inputs, we have devised a more dynamic method for injecting additional parameters. We described a first draft and proof-of-concept of this idea in Recla and Schmitt (2023b). The method works regardless of the size of the feature maps. The fundamental idea involves feeding the auxiliary data (in this case, the looking angle) into the outputs of the residual blocks, as illustrated in Fig. 3(b). We refer to this construct as the "Parameter Injection" block. It takes the scalar input s and reshapes it to match the spatial dimensions of the feature maps within the corresponding ResBlock, resulting in $S[1 \times m \times n]$. To provide this block with information about the current features, the residual feature map of the ResBlock F $[c_{out} \times m \times n]$ is passed through a convolutional layer, compressing it to a channel count of 3, creating F_{compr} $[3 \times m \times n]$. S is then concatenated with F_{compr} and forwarded through two activated convolutional layers. The number of channels already matches the output of the ResBlock $[c_{out} \times m \times n]$ that it can be added to the output of the residual block F. The network can thus utilize the additional information only where necessary, with the option to set regions of no interest to zero. As it is challenging to determine precisely where the model learns what and requires the additional information, the infusion block is incorporated into multiple locations throughout the network, specifically within each ResBlock, instead of only one particular point of the model. In theory, this approach can accommodate multiple additional parameters as well. Further scalar values (in a number k) can be straightforwardly added to the infusion block as additional channels, resulting in dimensions of $[k \times m \times n]$. But for this work, we injected the looking angle as the only additional parameter in the manner as it is described in Section 2.1.

2.2.3. Training strategy

The model is trained in a conventional supervised manner. The input images and the corresponding ground truth are cut into trainable patches of 512 × 512 pixels. A fixed number of individual patches are randomly sampled from the input images. This is to ensure a certain balancing between the different available recordings. In practice, 400 patches were drawn from each image. For the smaller scenes, this results in oversampling, but two identical patches are still unlikely. For each epoch, not all patches are then used in training, but a random subset. Thus, images with the same content but with some random offset/shift are shown in the different epochs (due to the random spatial sampling). As the loss function, $\mathcal{L}_1$ was used to train the model. Even though the MSE loss (Mean Squared Error) would reduce the final achievable RMSE, we decided against it, since many small spatial details in the output height maps would get lost using it. An alternating approach was also tested, but ultimately did not perform better than the $\mathcal{L}_1$ loss alone. Also, a batch size of 24 and the Adam optimizer were used with a learning rate of 0.0005. The input data is clipped between −30 and 10 dB and afterwards min–max-normalized to a range between 0 and 1. The target data (heights) are also scaled down in a similar fashion with 50 m becoming 1 but without clipping. This helps the model to train faster and more reliably. Before evaluation, the outputs are denormalized to represent meters again.

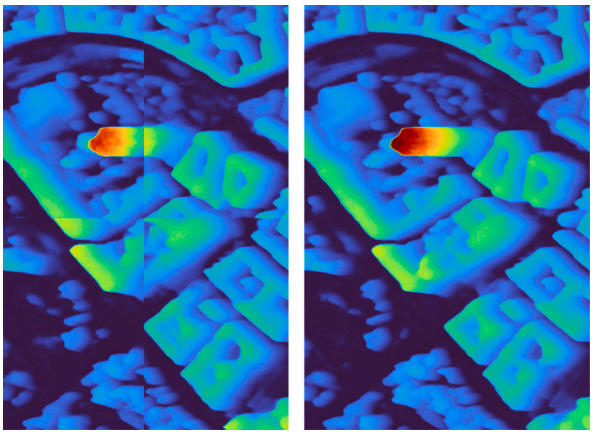
Like all datasets for Single Image Height Estimation in remote sensing, ours suffers from a long-tailed (or heavy-tailed) distribution, too. By their very nature, most pixels, even in urban areas, represent ground pixels and thus carry a value of zero in an nDSM. Roof pixels, and especially those of skyscrapers, on the other hand, do not appear in large numbers. This imbalance would result in a model that always tends to underestimate the height values just because of the statistical prior of more probable small heights. To counteract this long-tailed regression problem to some extent, all patches in the dataset are classified into "height classes" by the maximum height that occurs in each image section. These classes are then weighted against each other and drawn using a RandomWeightedSampler during training. This thus ensures that the model is presented with an equal number of patches per height class within an epoch, by under-drawing over-represented classes and over-drawing under-represented classes. The bounds of the used classes are as follows: 30 m, 60 m, 100 m, 150 m, and 200 m. This means that per epoch, the same number of patches with a maximum height of between 100 and 150 m and less than 30 m, for example, are provided to the model for training.

2.3. Postprocessing

A significant portion of the SAR2Height framework is the post-processing of the now patch-estimated height values. Most works from the literature end already at this point. However, the previously predicted height maps are neither georeferenced nor in a usual mapping geometry. The necessary steps are described in the following sections.

2.3.1. Generating coherent mosaics

The model estimates the heights above ground for each pixel of the input patches, in our case of the size 512 × 512 pixels. However, the patches, as they come out of the model, do not yet form a coherent image. Fig. 5(a) shows an example of how adjacent patches were simply stitched together. Unsightly artifacts appear at the borders and



(a) mosaic after simply stitching the outputs together (b) mosaic using the described technique

Fig. 5. The figure on the left shows the boundary area of four distinct height estimates, i.e. model outputs. Unsightly edges appear at the borders of the individual patches, especially in range direction. In the image on the right, the individual outputs have been merged using the method described, resulting in a seamless appearance.

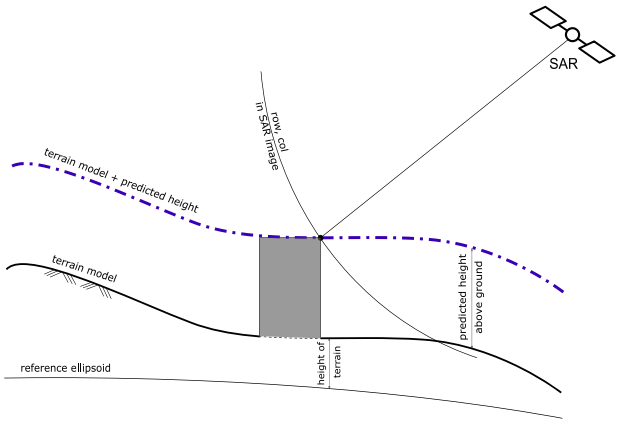


Fig. 6. Schematic illustration of the projection process from relative heights in slant range to absolute heights in ground range. An additional terrain model is needed to supply the absolute reference and to locate the pixel's target in 3D space. For every pixel, the task is to determine the intersection of the range circle (representing the pixel) and the purple dashed line, which is the terrain including the estimated height for this pixel.

buildings cut off at the edges are underestimated. Reasonable, because in such a case the base of the building is not shown in the input and therefore only a part of the layover can be interpreted. This makes it impossible for the model to estimate the building height correctly. To overcome this problem, the mosaic is formed in a two-step process. First, a 100-pixel-wide area at the edges of each output patch is discarded (to ensure that only partially mapped buildings are not included in the final result). The remaining core of the patch is merged with the adjacent one by weighting a small overlapping area against each other. The weights applied here decrease linearly with the distance from the center of each patch. This results in an artifact-free composite, as can be seen in Fig. 5(b). Also, the tall building in the middle of the example is no longer underestimated.

This patch-wise approach together with the oblique view of the SAR system limits the maximum detectable building height. As described in Section 2.2.2, and shown in Fig. 4, the object's height and the incidence angle influence the length of the mapped layover effect of raised objects such as buildings. Of course, we can only expect a model to estimate its height correctly if there is a complete view of the object. Taking the padding into account, the following limitations can be identified for different incidence angles: While a 150 m high building at an incidence angle of 20° already represents the maximum height that can be mapped, at an incidence angle of 50°, the buildings can be almost 500 m tall to be still covered within the area of the patch. A medium-range incidence angle of 35° leads to a maximum building height of approximately 300 m.

2.3.2. Projecting to UTM

From the model and after the mosaicking process, we obtain the estimated relative height values above ground for each pixel of the supplied SAR intensity image, but still in image space, which in the case of SAR means the so-called slant range geometry. In slant range, the image columns directly correspond to the distance to the sensor. This means that the pixels of an image column contain the echoes of all backscatterers at the same distance from the sensor. As a result, building facades, for example, are imaged as lying towards the sensor (compare Fig. 2). In order to correct these geometric peculiarities of a SAR image and to obtain a familiar orthometric projection, the absolute heights for each pixel must be known. Together with the known acquisition geometry of the sensor, namely the position and velocity of the phase center of the antenna in an earth-fixed coordinate system, each pixel can thus be georeferenced and an image in the so-called ground range geometry is generated. A first study on this approach can be found in Recla and Schmitt (2023a). In order to derive a DSM in ground range from an nDSM in slant range, we adopt the long-known method for the orthorectification of SAR images (Curlander, 1982): Finding the position of a pixel's target in a reference frame is often referred to as *geocoding* and involves solving a system of three non-linear equations: First, the range equation

$$R = \left| \boldsymbol{P}_S - \boldsymbol{P}_t \right| \tag{2}$$

with $\boldsymbol{P}_S$ as the sensor's and $\boldsymbol{P}_t$ as the target's position in an earth-centered frame. The range value R is defined by the column of the SAR image, the position of the sensor is known from the metadata. When satisfied, the equation ensures that a target at the position $\boldsymbol{P}_t$ would have ended up in the correct column (range gate) of the SAR image in question. The next equation to solve is the so-called Doppler equation

$$f_{Dc} = \frac{2}{\lambda R} \left(\boldsymbol{V}_S - \boldsymbol{V}_t \right) \cdot \left(\boldsymbol{P}_S - \boldsymbol{P}_t \right), \tag{3}$$

where f_{Dc} is the Doppler centroid frequency, λ the signal wavelength, $\boldsymbol{V}_S$ represents the sensor's and $\boldsymbol{V}_t$ the target's velocity, which can be derived from $\boldsymbol{V}_t = \boldsymbol{\omega}_e \times \boldsymbol{P}_t$ with $\boldsymbol{\omega}_e$ acting as the Earth's rotational velocity vector. This equation includes all possible $\boldsymbol{P}_t$s that would be mapped in the corresponding image line, i.e. in the direction of azimuth or flight direction of the sensor. To limit the multitude of now still possible solutions for $\boldsymbol{P}_t$, the last equation is introduced: the so-called world equation, which draws the solution to the Earth's surface. The equation should model the shape of the Earth. Usually, an oblate ellipsoid is used for this purpose:

$$\frac{x_t^2 + y_t^2}{\left(R_e + h\right)} + \frac{z_t^2}{\left[(1-f)\left(R_e + h\right)\right]^2} = 1 \tag{4}$$

with the radius of the Earth as R_e at the equator, the target height as h above the chosen reference model, the flattening factor f, and x_t, y_t and z_t as the single components of $\boldsymbol{P}_t$.

This method requires knowledge of the absolute altitude h above the reference ellipsoid. However, we estimate only a normalized value, i.e. the relative heights above ground. So, Eq. (4) cannot be used as it is. To provide this missing absolute component to the process, we have to introduce a piece of external information in the form of a digital terrain model. To make the method work worldwide, this terrain model should be globally available. Also, it should actually feature bare ground, in other words, no man-made objects or vegetation,

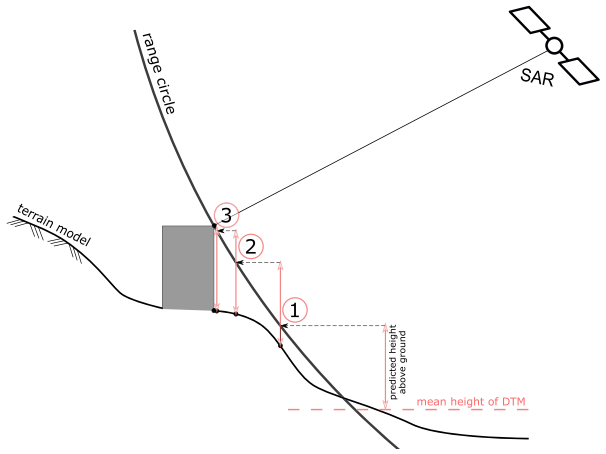


Fig. 7. Illustration of the iterative process to forward geocode a pixel together with its estimated height above ground and a DTM. Position ① on the range circle is found by the evaluation of the RPC model together with an initial height assumption. At this location, the height of the terrain together with the estimated building height is used to get the next approximation ②. This is repeated until the position no longer changes ③.

since that is what we are estimating with the model. The process of incorporating the absolute component is described in the following: For each individual pixel of the model's output, the digital terrain model (DTM) is elevated by adding its respective estimated height, denoted as $\mathrm{DTM}' = \mathrm{DTM}, +, h_{est}$ (as illustrated by the purple dashed line in Fig. 6). Then, the algorithm searches for the precise location on the modified DTM' where both the range and Doppler equations are fulfilled. Notably, the Earth model Eq. (4) is inherently satisfied due to the search restricted to the surface of DTM'. Consequently, only two equations are optimized through this approach. The accuracy, however, remains sub-pixel precise, as continuous interpolation between the DTM cells takes place. To practically implement this method, the *fsolve* optimizer from the *scipy* Python Library can be utilized. This optimizer determines the roots of a system of equations starting from a given point, by minimizing the sum of squares of its components. The outcome of this process yields a three-dimensional point cloud in an absolute reference system, with each pixel of the image corresponding to a single point.

The use of numerical solvers inevitably entails a considerable computational effort, which is reflected in the runtime of the algorithm. Therefore, an alternative approach is proposed in the following: By approximating the Range-Doppler model described above with a so-called RPC sensor model (rational polynomial coefficients), the runtime can be substantially reduced. To set up an RPC model, a set of ground control points (GCPs) with their corresponding SAR image coordinates is required. These points can be obtained in any number by the known range-Doppler model and are therefore often referred to as virtual GCPs (vGCPs). Using the Eqs. (2)–(4), for the vGCPs following a grid in world space (X_i, Y_i, Z_i), their corresponding image coordinates (row_i, col_i) are determined. With this set of coordinate pairs, the 78 coefficients describing the RPC model can be defined by a least squares fit (Akiki et al., 2021). However, in order to combine the predicted relative heights with the absolute heights from the DTM, an iterative process is again necessary: In a first step, the absolute height H we are looking for is initiated as the mean height of the DTM $\overline{h}_{DTM}$ together with the estimated relative height h_{est}:

$$H_{init} = \overline{h}_{DTM} + h_{est} \tag{5}$$

This allows the pixel (row, col) to be localized using the RPC model by

$$X, Y, Z = RPC(row, col, H). \tag{6}$$

The X and Y coordinates obtained can be used to further refine the height H by means of interpolation on the DTM:

$$H = DTM(X, Y) + h_{est} \tag{7}$$

These two steps, i.e. the localization of the pixel with the current H (6) and the subsequent re-evaluation of H (7), are repeated until the position in X and Y stops changing (within a specified tolerance). Fig. 7 schematically shows the sequence of such an iteration in very difficult terrain to illustrate the concept. The process runs in 3 iteration steps. The mean height and the estimated building height are used to determine position ① on the range circle. Position ② is identified by adding up the terrain height under position ① and the building height to get an updated H to put into the RPC model. At some point, the algorithm converges at the location of the building being searched for, and the next iteration results in no, or only a minimal, change in its position.

Using the RPC-based approximation for the geocoding process gives the method a massive boost in terms of runtime. For example, compared to the range-Doppler approach, the reprojection of a HS scene (approx. 45 km^2) now takes 34 s (single-core) instead of 67 min (using 30 CPU cores). This is mainly due to the vectorizability of the polynomial functions. In addition, in flat urban areas very rarely more than three iteration steps are necessary to converge with the method presented above. The residuals between the results of the Range-Doppler approach and the RPC method are in the centimeter range and are therefore negligible regarding the achievable accuracy of the overall task.

2.3.3. Filtering the point cloud

If the predictions of the model were perfect, after the former described process of georeferencing each pixel, the resulting point cloud would already be ready to use. However, as the prediction of 3D information from a 2D image remains a mathematically ill-posed problem, estimation errors cannot be avoided. Deviations in the estimated heights lead to misplaced points in the point cloud — not only in the height component but also in the positional coordinates. This reallocation happens during the projection process. In Fig. 6, a wrongly predicted height would move the resulting point of the point cloud along the range circle away from its true position in world space. Fig. 8 shows this behavior in more detail. The gray block is meant to symbolize a potential building. At its upper right corner, the range circle's tangent is indicated as an approximation for it. Along the range circle, a solution of the above equation system is searched for. If the height of the building is overestimated, the resulting georeferenced point will move away from the sensor and its actual position. If the opposite is the case, the point comes closer to the sensor. This means that the error, which in slant range only occurs in the vertical direction, is distributed between the location and the height components after the transition to ground range. The displacement in range direction Δl can be approximated by

$$\Delta l = \Delta h \cdot \cot\theta, \tag{8}$$

with Δh representing the error in the height prediction and θ as the looking angle.

A well-known problem of monocular depth estimation is that the model is trained using a 2-dimensional loss function, i.e. on the depth maps. Converting these depth maps to points in 3D space can result in very different accuracies. A small deviation in depth around edges leads to dispersed points floating around in the air. This would not happen if the model had the opportunity to train directly in 3D space, but this is computationally very expensive and till now oftentimes not possible. Pınar Örnek et al. (2022) did an extensive investigation on this problem. The model used here also suffers from this shortcoming. In our case, we do not estimate depth maps but height maps in slant range, which is a 2-dimensional representation of the 3-dimensional world, too. As a U-Net, the model struggles to generate clear edges in the height map when the input image shows blurry edges as well

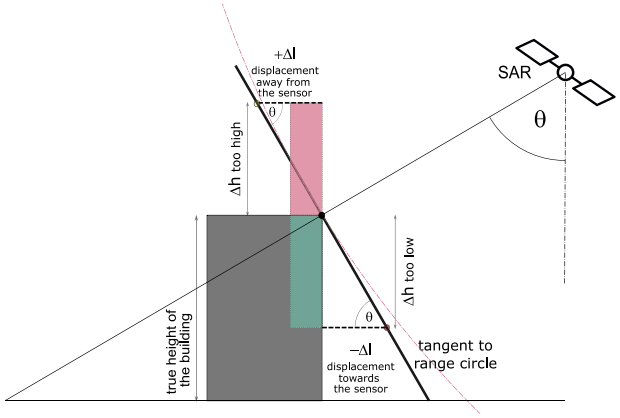


Fig. 8. The impact of an inaccurate height prediction, denoted as Δh, on the displacement in the range direction Δl of a target in world space is dependent on the viewing angle θ. When the height (for instance of a building) is overestimated, the projected point target will move farther away from the sensor in the georeferenced point cloud; conversely, if it is underestimated, the target will approach the sensor. This shift is happening in range direction upon the range circle (here approximated by its tangent).

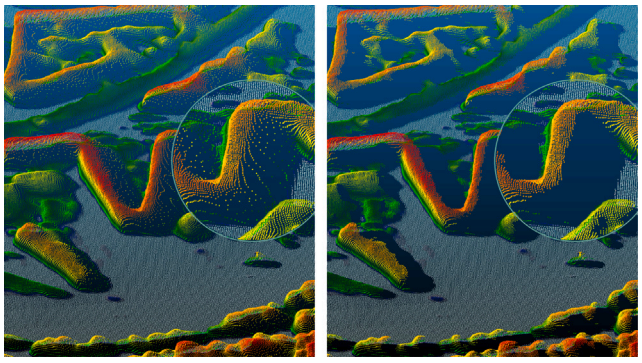


Fig. 9. Comparison between the raw point cloud after projection to 3D space (left) and the filtered point cloud (without shadow filtering) (right). Blurry edges in the height estimation lead to levitating isolated points after the conversion to ground range. They can be seen in the blank areas of layover. The sensor captured the scene from the right (as the direction of the layover effect suggests).

due to speckle effect and limited spatial resolution. It is in its nature to stay on the safe side in such cases and to generate a blurred edge, with which it achieved a better loss in training. In slant range this is no further disturbance, but after projection to ground range, washed-out building edges lead to unsightly free floating, non-contiguous points. And, due to the effect described above, these points do not spread only in vertical direction around the building wall, but also in horizontal direction (compare Fig. 9(a)). For filtering out these erroneous points, we can take advantage of the property that the error component in the point's position is limited to the range direction, i.e. along the sensor's line of sight. Considering this aspect, the filter algorithm operates as follows:

First, the range direction α_{range} is determined on the ground, i.e. in world coordinates. This can be done using two points originating from the same image line j but different columns with

$$\alpha_{range} = \arctan\left(\frac{y_{j,k} - y_{j,l}}{x_{j,k} - x_{j,l}}\right). \tag{9}$$

In addition, the three-dimensional point cloud is thinned and reduced to the two spatial dimensions (x and y) to speed up the subsequent filtering operations and thus make the process more efficient. The thinning sieves out all those points that are not far enough from their neighboring points in their 2-dimensional position using a specified threshold. For each of the remaining points, the relative direction $\Delta\alpha_i$ between the connections to their 10 nearest neighbors and the range direction is calculated through

$$\Delta\alpha_i = \alpha_{range} - \alpha_i, \tag{10}$$

with α_i being the directional angle for the neighbor i. The corresponding distances d from the examined point to its neighbors can then be divided into their range and azimuth components by

$$d_{range\,i} = d_i \cdot \cos\left(\Delta\alpha_i\right) \tag{11}$$

and

$$d_{azimuth\,i} = d_i \cdot \sin(\Delta\alpha_i). \tag{12}$$

Each point is considered a valid point only if at least one of the neighboring points meets the specified distance tolerances in both the azimuth and range directions, i.e. any $\left(d_{direction} < tol_{direction}\right)$. Otherwise, the point is regarded as isolated and removed from the final point cloud. In this way, the tolerances can be set differently in the two directions, which is necessary because these mislocated points are frequently close to each other in the azimuth direction, but far from each other in range direction. In Fig. 9, an example view of a point cloud before and after the filtering process can be seen. The scattered points behind the buildings are removed in the filtered point cloud (Fig. 9(b)).

As mentioned in Section 2.2.1, the CNN predicts height values for every pixel, even if it contains no echoes and therefore represents radar shadow. Because these areas were not visible to the sensor, even a hypothetically perfect model could not make any reliable assumptions about them. For this reason, these pixels are filtered out in the height image and appear there as *nodata* values. For that purpose, the intensity image is converted into a shadow map using a threshold: All low-intensity pixels will be discarded from the predicted height map based on this shadow map. By adding an acquisition from a different viewing angle, these gaps could be filled.

2.3.4. Rasterizing the point cloud

After filtering the point cloud, only one step is left, which is rasterizing the result and thereby generating an actual DSM. In an equidistant grid, the maximum occurring height value of all points located within each cell is mapped. This results in a 2.5D DSM or 2.5D nDSM (knowing the location of each pixel from the projection procedure, both values can be projected). Small holes with a maximum size of 3 pixels are filled by interpolation. The areas of layover and shadow lead to large gaps without data. These can optionally be filled with values coming from the DTM.

3. Experiments & results

In order to test and validate the SAR2Height framework described in Section 2, a series of experiments is carried out. We compare our adapted deep convolutional neural network architecture to the classic U-Net baseline, investigate the influence of different geometry-related acquisition parameters, quantify the reprojection error due to the conversion to ground range, and finally apply the full framework to all available data in a cross-validation manner.

3.1. Data

For training and evaluation of the model, a dataset was compiled consisting of SAR intensity images and their relative elevation data projected into slant range geometry. A total of 51 TerraSAR-X Level 1B SLC images from 8 different cities were available for that purpose. The

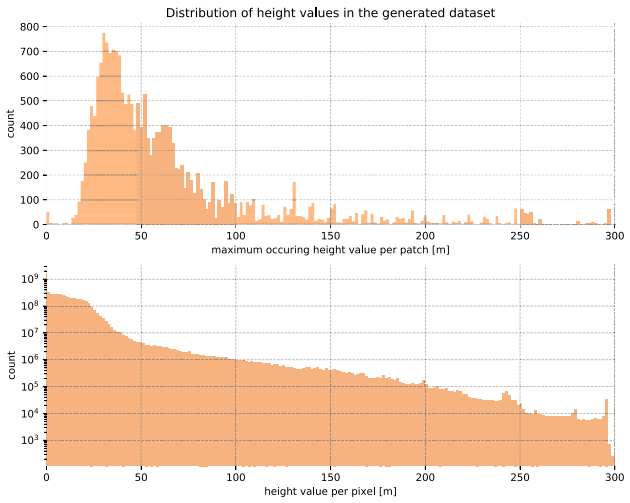


Fig. 10. The upper graph shows the histogram of the maximum occurring height values per each patch in the generated dataset. The figure at the bottom, on the other hand, shows the height distribution on a pixel-wise level throughout the dataset. The y-axis is here in a logarithmic scale. There are way more small height values present in the nDSMs. This represents a heavy-tailed distribution and is challenging for neural networks.

images vary with respect to orbit direction, incidence angles, imaging modes, i.e., High Resolution SpotLight (HS) and Staring Spotlight (ST), and polarization, with the majority acquired horizontally–horizontally (HH), and only a small portion in vertically–vertically (VV). For the most part, the acquisitions were archival data. The most descriptive parameters of the available SAR acquisitions are summarized in Table 1. The two imaging modes differ in their azimuth resolution. In HS mode, the antenna is steered using a squint angle to elongate the acquisition time and thus the synthetic aperture and achieve a resolution in azimuth of 1.1 m. In ST mode, the radar antenna is pointed at the target during the fly-by for an even longer period of time, which reduces the azimuth resolution to about 25 cm, but also reduces the area covered. The range resolution remains for both the same at 60 cm. Even if the pixel spacing of the model inputs is resized to 1 m in both directions, a distinct difference can still be observed. Due to the enlarged resolution cells, the speckle effect is also more pronounced in the HS images and fine spatial details are therefore more difficult to perceive. The data was collected over the cities of Munich, Berlin, Frankfurt, London, Barcelona, Vienna, Melbourne, and St. Louis. The different scenes from the same cities overlap spatially but differ in their recording geometry and/or spatial resolution.

The primary constraint for deep learning-based single-image height reconstruction approaches pertains not so much to the availability of satellite data, but to the accessibility of high-resolution urban elevation data, which is required to provide the supervisory target signal. However, more and more governmental geospatial data providers, mostly land surveying agencies, subscribe to an open data policy, which makes it easier to collect the needed training data even if these data are still very heterogeneous between different governmental districts and thus need to be preprocessed regarding their format and height systems. In addition, if no DTM is available, a terrain model must be generated from the surface model, since this is required to determine the nDSM containing only relative heights above ground. This can be done by removing buildings and vegetation from the surface model and then interpolating the gaps.

Fig. 10 shows both the distribution of height values in the entire data set per pixel (bottom, note the logarithmic scaling of the y-axis) and the distribution of the maximum height values occurring per patch (top). It can be seen that there is a difference of several orders of magnitude between the number of pixels with zero, pixels with heights of around 20 m, and pixels with height values greater than 50 m. To counteract this imbalance, height-aware sampling is applied during training (see Section 2.2.3).

3.2. Error metrics

In order to classify the performance of the model and the entire pipeline numerically, various error metrics are used. These are kept to a minimum since the statement usually remains the same even with a vast number of metrics. A very simple and straightforward metric is the Mean Absolute Error (MAE). It is used to quantify the average absolute difference between predicted values $\hat{y}$ and actual target values y. It provides a simple and interpretable measure of the overall prediction accuracy. The formula is given by

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \tag{13}$$

with n as the number of data points, y_i is the actual target value for pixel i, and $\hat{y}_i$ as the predicted value for pixel i.

A similar measure, but penalizing larger errors more heavily through the square, is the Root Mean Squared Error (RMSE). It is widely used to assess modeling errors. With the same nomenclature as above, it is given by

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}. \tag{14}$$

The Pearson Correlation Coefficient (or Pearson's r) is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It ranges from -1 to 1, where 1 represents a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 corresponds to no linear relationship. It is given by

$$r = \frac{\sum_{i=1}^{n}(y_i - \mu_y)(\hat{y}_i - \mu_{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \mu_y)^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \mu_{\hat{y}})^2}}, \tag{15}$$

with y_i and $\hat{y}_i$ representing the target and predicted value for pixel i and μ_y and $\mu_{\hat{y}}$ as their corresponding mean values.

The Structural Similarity Index Measure (SSIM) is a metric used in image processing and computer vision to quantify the perceptual similarity between two images. It takes into account luminance, contrast, and structure to provide a comprehensive assessment of image quality that aligns more with human perception than traditional pixel-based measures (like PSNR). It is determined in a local windowed fashion (using Gaussian kernels), averaging the local results to get the final measure. It is defined as

$$\mathrm{SSIM} = \frac{(2\mu_y\mu_{\hat{y}} + C_1)(2\sigma_{y\hat{y}} + C_2)}{(\mu_y^2 + \mu_{\hat{y}}^2 + C_1)(\sigma_y^2 + \sigma_{y\hat{y}}^2 + C_2)}, \tag{16}$$

where y and $\hat{y}$ are the predicted and target images, μ_y and $\mu_{\hat{y}}$ their average pixel intensities, σ_y^2 and $\sigma_{y\hat{y}}^2$ the corresponding variances, $\sigma_{y\hat{y}}$ representing the covariance, and C_1 and C_2 as constants to stabilize the division (Wang et al., 2004).

3.3. Ablation study for the single-image height prediction network

In order to substantiate the merits of the deep convolutional network architecture proposed in this article, our SAR2Height model is compared to a plain U-Net without any of our presented modifications. This baseline U-Net, like introduced by Ronneberger et al. (2015), has only been adapted so that Softmax is not used as the last activation anymore, because that originally was designed for segmentation tasks, and so that the number of convolutional filters per block matches the number used in the proposed model to ensure a fair comparison (64–1024 instead of 32–512). The two models were trained in the exact same fashion, using the same data set including all available

Table 1
Table with some meta parameters of the 51 individual SAR images comprising the dataset used in all the experiments. The cells are color-coded to facilitate the impression of the distribution of individual data characteristics in the data set.

	Image	Mode	Pol	Orbit	Incidence Angle [deg]	Acquisition date
Berlin	ber_432	HS	VV	A	30	10.03.2015
	ber_553	HS	VV	D	36	12.03.2015
	ber_426	ST	VV	D	22	30.05.2016
	ber_312	ST	VV	A	41	23.12.2016
	ber_609	ST	VV	D	36	15.07.2018
	ber_612	ST	VV	D	36	17.08.2018
	ber_452	ST	VV	A	30	17.09.2018
Munich	muc_749	HS	VV	D	49	18.06.2008
	muc_634	HS	VV	D	38	19.06.2010
	muc_507	HS	VV	D	25	24.06.2010
	muc_131	HS	VV	A	23	03.07.2010
	muc_136	HS	VV	A	23	10.10.2010
	muc_045	ST	VV	A	37	14.07.2017
	muc_714	ST	VV	D	39	17.07.2017
	muc_942	HS	HH	A	48	23.07.2020
	muc_121	HS	HH	A	37	28.01.2023
London	lon_533	HS	HH	A	23	29.04.2011
	lon_919	HS	HH	D	47	01.11.2015
	lon_431	HS	HH	A	37	16.02.2016
	lon_634	ST	HH	D	23	08.11.2016
	lon_442	ST	HH	A	37	17.11.2016
	lon_504	ST	HH	A	37	05.05.2020
Barcelona	bar_238	HS	HH	D	33	18.02.2009
	bar_406	ST	HH	D	33	29.01.2023
	bar_241	HS	HH	A	35	31.01.2023
	bar_112	ST	HH	A	48	05.02.2023
	bar_239	ST	HH	A	35	11.02.2023
St. Louis	stl_430	HS	HH	D	25	23.08.2012
	stl_557	HS	HH	D	41	29.08.2012
	stl_012	HS	HH	A	47	31.08.2012
	stl_724	HS	HH	D	53	04.09.2012
Melbourne	mel_001	ST	HH	A	49	18.07.2022
	mel_658	ST	HH	D	27	18.07.2022
	mel_300	ST	HH	A	20	19.07.2022
	mel_957	ST	HH	D	53	19.07.2022
	mel_131	ST	HH	A	37	24.07.2022
	mel_828	ST	HH	D	42	24.07.2022
	mel_659	HS	HH	D	27	29.07.2022
	mel_959	HS	HH	D	53	30.07.2022
	mel_829	HS	HH	D	42	04.08.2022
	mel_003	HS	HH	A	49	09.08.2022
	mel_302	HS	HH	A	20	10.08.2022
	mel_133	HS	HH	A	37	15.08.2022
Frankfurt	fra_931	ST	HH	A	36	29.12.2013
	fra_104	ST	HH	A	21	19.11.2014
	fra_339	ST	HH	D	34	21.11.2014
	fra_903	ST	HH	A	47	18.02.2023
Vienna	vie_403	HS	HH	A	26	27.02.2008
	vie_242	HS	HH	A	40	04.08.2008
	vie_034	HS	HH	D	36	01.02.2023
	vie_159	HS	HH	D	48	07.02.2023

Table 2
Comparison in the performance of an original U-Net against the here described SAR2Height architecture, once with and once without the so-called "Parameter Injection" (PI). All models were trained in the same fashion on the same data set (excluding images from Berlin). The models were applied to all acquisitions from Berlin. The mean absolute error is split up into different "height classes" according to the ground truth. The difference between the models becomes larger while concentrating on tall structures. The values are computed in slant range geometry, so that no reprojection errors distort the comparison.

Image	Model	MAE ↓ [m]				Pearson ↑	SSIM ↑
		Overall	<10 m	10–30 m	>30 m		
ber_312	UNet	3.71	2.64	4.38	18.38	0.74	0.84
	Ours w/o PI	3.48	**2.39**	4.25	16.13	0.76	0.85
	Ours	**3.46**	2.55	**4.06**	**15.33**	**0.77**	**0.85**
ber_426	UNet	4.23	3.27	4.17	19.34	0.69	0.84
	Ours w/o PI	4.20	**2.92**	4.54	17.79	0.71	0.84
	Ours	**4.04**	3.10	**4.10**	**16.93**	**0.71**	**0.85**
ber_432	UNet	4.12	3.18	4.37	16.59	0.73	0.83
	Ours w/o PI	3.96	**2.91**	4.41	14.57	0.76	0.83
	Ours	**3.79**	2.96	**4.08**	**13.56**	**0.76**	**0.85**
ber_452	UNet	3.63	3.01	3.54	17.05	0.74	0.87
	Ours w/o PI	3.60	**2.62**	3.86	15.56	0.76	0.87
	Ours	**3.46**	2.70	**3.58**	**14.77**	**0.76**	**0.88**
ber_553	UNet	3.85	2.91	4.34	14.81	0.76	0.83
	Ours w/o PI	3.68	**2.68**	4.35	12.65	0.78	0.83
	Ours	**3.51**	2.73	**3.96**	**12.06**	**0.79**	**0.85**
ber_609	UNet	3.34	2.72	3.48	13.66	0.82	0.81
	Ours w/o PI	3.18	**2.40**	3.58	11.52	0.84	0.82
	Ours	**3.13**	2.55	**3.34**	**11.01**	**0.84**	**0.82**
ber_612	UNet	3.32	2.66	3.55	13.07	0.82	0.77
	Ours w/o PI	3.13	**2.34**	3.59	10.64	0.85	0.78
	Ours	**3.10**	2.48	**3.41**	**10.37**	**0.85**	**0.78**

cities except Berlin. Table 2 shows the error scores achieved for the baseline, the SAR2Height model with and without using the parameter injection on all of the remaining SAR images covering different areas in Berlin. The results are shown for all of the available Berlin images. The method presented here performed better than the baseline in every category. However, looking only at the partly small numerical increases in performance scores averaged over the entire scene, the cost-benefit ratio of the increased complexity of the SAR2Height model could be questioned. The usefulness of the added modules is revealed more explicitly in a visual comparison. Fig. 11 shows this comparison as a side-by-side examination of the baseline's and the complete SAR2Height model's outputs against ground truth. Tall buildings in particular seem to benefit from the new architecture. To reflect this in the error metrics as well, the mean absolute error was not only calculated over the entire scenes but was divided into "height classes" according to the pixel's height values. The height classes, or height ranges, are determined via the ground truth data. Only those pixels that belong to the corresponding height class in the ground truth are subsequently used for the error determination. The classes used were pixels with height values below 10 m, between 10 and 30 m, and everything above 30 m in height. Buildings taller than 30 m are considered high-rise and represent an extraordinary challenge for the models. This is where the difference to the baseline also becomes larger in the achieved MAE. It seems as if it is particularly these special cases that are covered by the increased complexity in the SAR2Height model.

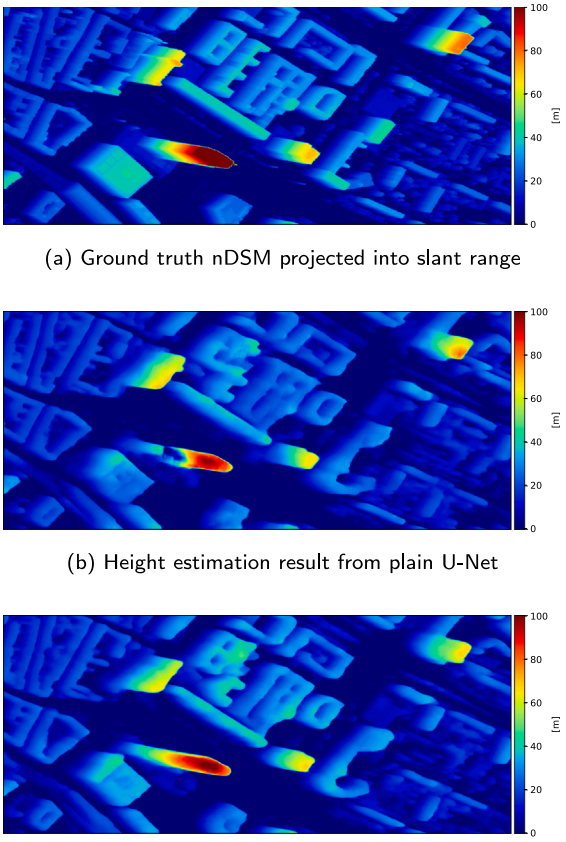
To highlight the influence of the Parameter Injection Blocks (PI-Blocks), an exceptional example is shown in Fig. 12. It shows the outputs of the models once with and once without the use of the looking angle as an auxiliary quantity within the PI-Blocks in comparison to the ground truth in slant range. The largely free-standing skyscraper was estimated to be far too high without using the PI-blocks, as demonstrated by the elevation profile shown in Fig. 12(e). Although the numerical error metrics do not suggest a large difference, the use of the looking angle can have a significant effect on the quality of the height estimate in individual cases. However, not every building shows such a strong and significant effect. Numerous buildings are estimated correctly even without the PI-Blocks, regardless of the geometry of the image. Thus, there must be viewpoint-invariant features that allow the model to infer height, such as the number of rows of windows that can be converted to stories and thus a height.

3.4. Reprojection error

As mentioned in Section 2.3.2 and as shown in Fig. 8, an inaccurately estimated height in slant range affects not only the height component (z) of the result after projection in ground range but also its position (x and y). The shift here takes place mainly in range direction, which of course does not have to correspond exactly to an axis of a (projective) coordinate system like UTM. To quantify this effect, the SAR2Height method was applied to a SAR acquisition showing an area in Berlin. The model used for this purpose was trained, like before, on the remaining cities exclusively. Table 3 shows the mean absolute error between the predictions and the ground truth in three different phases of the system, namely after the effective height estimation in slant range ($\mathbf{MAE}_{slant\ range}$), after the conversion of these estimates into a georeferenced point cloud in the UTM system ($\mathbf{MAE}_{point\ cloud}$), and yet in the final grid in UTM (i.e. the nDSM in ground range, $\mathbf{MAE}_{ground\ range}$). In the 3-dimensional point cloud, it is possible to decompose the error into its components x, y, and z, because, for each pixel of the SAR image, not only the relative height from the ground truth is known, but also its 3-dimensional position in space. So, after the projection to UTM, an error vector per pixel can be determined, which in turn is decomposed into its components. The height dimension z can be further distinguished into absolute z_{abs} and relative heights z_{rel}. The relative heights are those above ground, which are also estimated by the model. By the projection to ground range and through the external DTM, however, the absolute height for each pixel is known as well. This absolute height additionally contains the error of the used DTM. The values were calculated for a complete scene (not tile by tile) and averaged. It should be noted, that the point cloud was filtered as described in Section 2.3.3. The error metrics are thus already adjusted for shadow pixels and obvious faulty reprojected targets. It can be seen

Table 3
Error evaluation in different stages of the method: after the height estimation in slant range ($MAE_{slant\ range}$), after the georeferencing to UTM as a point cloud ($MAE_{point\ cloud}$) and after the rasterization of the point cloud to a nDSM in ground range ($MAE_{ground\ range}$). In the point cloud, the error was separated into its 3 components along the axes of the coordinate system. The same scene was projected to ground range using two DTMs of different resolutions and accuracies.

Image	DTM for reprojection	$MAE_{slant\ range}$ [m]	$MAE_{point\ cloud}$ [m]					$MAE_{ground\ range}$ [m]	
			norm	x	y	z_{abs}	z_{rel}	DSM	nDSM
ber_609	WorldDEM™ DTMlite	3.13	8.42	5.84	1.34	4.38	3.19	6.52	5.08
ber_609	HighRes LiDAR DTM	3.13	6.35	4.28	0.97	3.19	3.20	4.53	4.53
ber_312	WorldDEM™ DTMlite	3.46	8.73	5.54	0.79	5.04	3.57	6.54	4.74

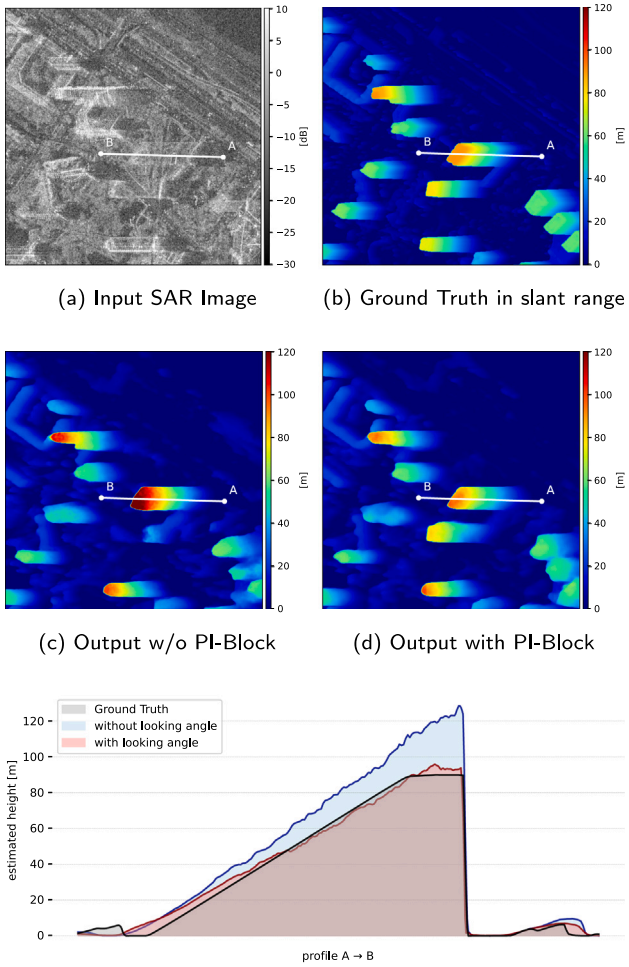


(a) Ground truth nDSM projected into slant range

(b) Height estimation result from plain U-Net

(c) Height estimation result using our SAR2Height Network

Fig. 11. Qualitative comparison between the outputs of different models. The top figure shows the normalized height values projected into the slant range geometry of the SAR acquisition in question. In the middle row, the output of the original U-Net can be seen. The figure at the bottom is our model's output. The two models were trained the exact same way for an equal number of epochs on the same training set. It can be seen that, despite the similar-looking numerical error metrics, the differences are clearly visible, especially for taller structures.



(a) Input SAR Image

(b) Ground Truth in slant range

(c) Output w/o PI-Block

(d) Output with PI-Block

(e) Profile through one of the shown buildings

Fig. 12. Comparison between the outputs of the model once with and once without using the Parameter Injection Blocks (PI-Blocks) in slant range. The corresponding ground truth can be seen in the upper right. The elevation profile at the bottom shows the intersection right through a reconstructed skyscraper (A to B). The model without the PI-Blocks overestimated the height of the shown building heavily.

that the final error after rasterization is a mixture primarily of the error in the x and z-directions. The x-direction (west-east) corresponds for TerraSAR-X approximately to the range direction through its polar orbit.

This also includes the error introduced by the use of the coarse-resolution terrain model. A terrain model in urban space always represents an artificial product derived algorithmically, which is subject to uncertainties. If a higher resolution DTM were available for the city in question, this could be used and thus reduce the final error budget. For a comparison of the impact of this effect on the results in ground range, the scene from Berlin was also projected to ground range with a high-resolution DTM from a LiDAR campaign (grid size of 1 m). The results are shown in Table 3 as well. As can be seen, the accuracy of the relative height z_{rel} remains approximately the same, these values are not affected by the reprojection. The absolute height, on the other hand, depends on the DTM used, and with it, the displacement of the pixels in range direction (approximately x) changes too. The error introduced by the terrain model can be directly quantified by comparing the MAEs in DSM and nDSM in ground range. While these remain the same using the precise DTM, the mean error in the estimated DSM is almost 1.5 m larger than in the nDSM when the lower-resolution terrain model was used for the projection. Fig. 13 shows a visual comparison between the two generated DSMs. The error maps (13(b) and 13(d)) quantify the

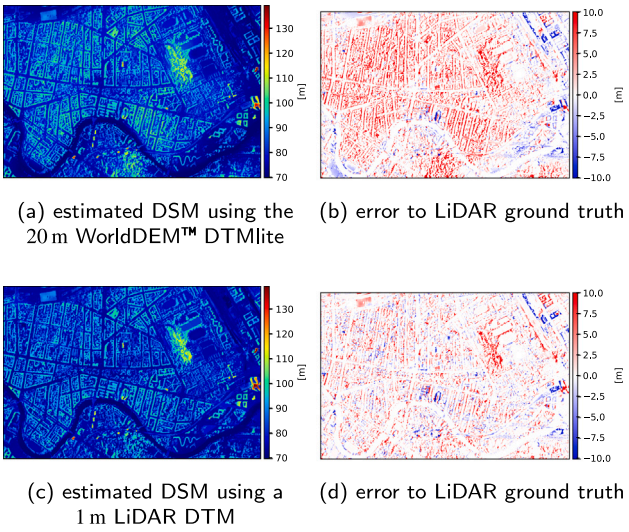


Fig. 13. Comparison between the estimated DSMs using a low- and high-resolution DTM for reprojection. The low-resolution version (top row) leads to larger errors (see error maps on the right) because of the errors present in the DTM and the consequential shift of elevated pixels in range direction after the projection to ground range.

difference between the estimated DSMs and ground truth from LiDAR. In a pixel-wise comparison like this, the pixel's shift in x-direction affects of course the error map. The edges of the buildings are visualized as areas of high errors because they are not at their true position. This effect is greatly reduced using the high-resolution DTM for reprojection (Fig. 13(d)). Overall, a high-resolution DTM should thus be used if such is available. For a globally applicable version, however, there is so far no higher-resolution alternative of better accuracy. This is the reason why all experiments shown here were done with the WorldDEM™ DTM-lite, even if higher resolution alternatives would have been at hand. The table also shows that different scenes result in different final achievable accuracies, even if the error in z is similar in the point cloud. This is due to the different acquisition geometries of the images and will be discussed in the next section.

3.5. Evaluation across different acquisition settings

The goal of the SAR2Height framework is to be as generically applicable as possible and therefore produce comparable (n)DSMs, regardless of the geometric setting of the image, i.e. orbit and angle independent. Fig. 14 compares two SAR images of the same area but taken from different orbits at different angles. The orthometric nDSMs in ground range (rightmost column) should be perfectly georeferenced and thus comparable. However, the accuracies to be achieved do depend on the acquisition geometry and the imaging mode. Since the area to be estimated has a substantial impact on the error metrics as well, all SAR scenes used here were cropped to the same overlapping area. This prevents, for instance, a difficult-to-estimate high rise from being present in one image and not in the other, which would bias the outcome. The results of this investigation can be seen in Table 4. The mean absolute deviation has been determined both in slant range (i.e. directly after the DL model) and in ground range (after post-processing). Refer to the middle and right columns in Fig. 14 for a visual comparison between slant and ground range.

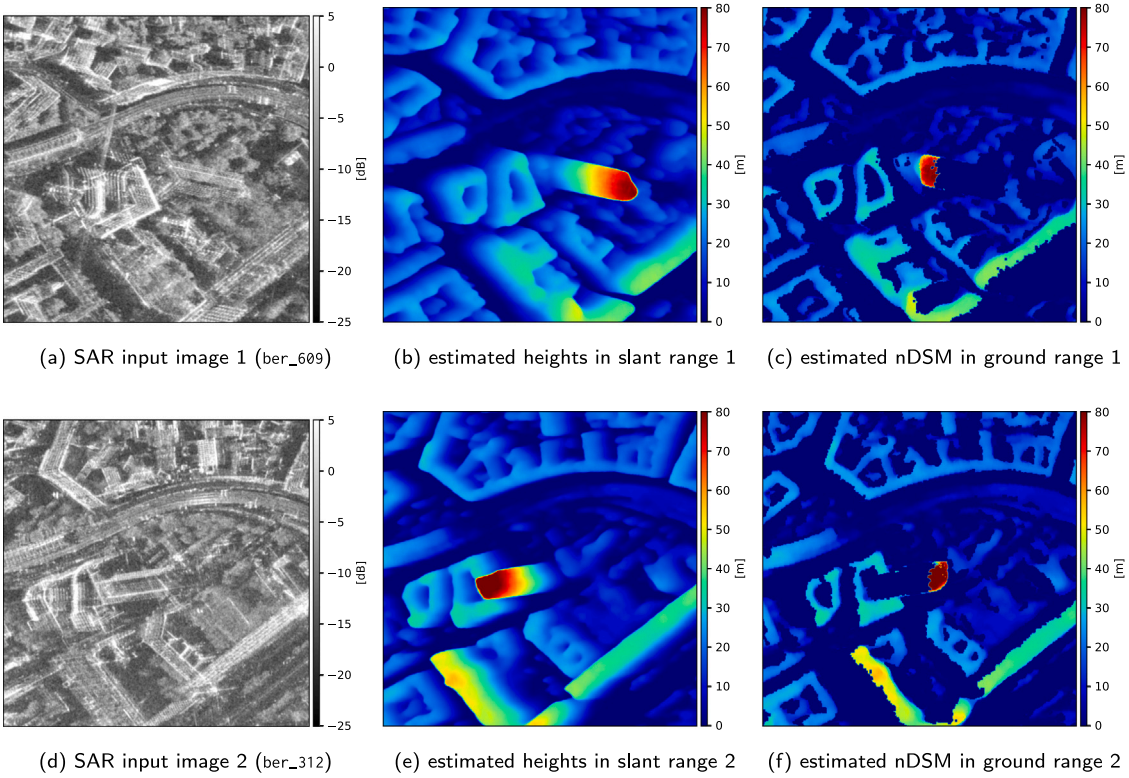
As in Section 3.3, the error was separated into different intervals. While the errors in slant range (overall) seem to be quite comparable, these changes in ground range and clear differences appear. Smaller incidence angles, corresponding to a steeper acquisition setting, lead to a stronger effect of the projection error from slant to ground range (see Section 2.3.2 and Fig. 8). This effect emerges clearly from the experiment. The image with the smallest incidence angle (`ber_426`) yields the largest error in ground range. However, this is not the only explanation for the reduced reconstruction capability. A steeper acquisition geometry yields longer layovers (see Fig. 4), which thus overlap more. Road ditches are also difficult to distinguish as a result. Separating these layovers from each other and correctly assigning them to only the tallest building is a challenge for the model. Even if the numerical error values, at least in the slant range, do not suggest any major differences between the various scenes, the results obtained using images with larger incidence angles are clearly preferable from a purely qualitative or visual point of view. Fig. 15 shows such a comparison. The model outputs in slant range of two different images, one steep (`ber_426`) and one flat (`ber_312`), are compared. The one with the larger incidence angle leads to much clearer and sharper edges, less misclassification, and overall a better fit to ground truth. Thus, in application, incidence angles in the mid to high range are preferable. It should be kept in mind that with a larger incidence angle, the shadow areas also grow. Even if this has a detrimental effect on the completeness of the resulting DSMs, the advantages of the reduced layover and the associated smaller influence of the reprojection errors discussed above still outweigh the disadvantages.

In contrast, the reduced resolution of the High Resolution SpotLight (HS) acquisitions compared to the Staring SpotLight (ST) images did not significantly degrade the results. However, the choice of orbit (ascending or descending) plays a further role in the result. Since the roof surface of a building always merges with the echo of the facade in the layover, the edge of the building further away from the sensor can be better identified in the SAR image than the edge closer to the sensor. Consequently, this edge is localized more precisely in the final DSM. Another circumstance to be considered is that a building perfectly estimated in relative height will appear in a different position in an ascending and descending image (with constant viewing direction) in ground range, if the DTM used for the projection presents an inaccuracy since this will affect the two images in exactly the opposite way (e.g. the pixels will be shifted towards the sensors, but they were looking at the scene from different directions).

3.6. n-Fold cross-validation

To validate the model's performance also across different test sites, cross-validation is performed with all the data available. The split between train and test set is done very strictly, namely city-wise. This means that the city on which testing is performed is completely excluded from the training. This test set also has no effect on the choice of the checkpoint of the model weights. For all results shown here, a fixed number of epochs was trained, regardless of the fact that there may have been a more optimal point in time during the training. This is to create a real-world scenario in which the previously trained model is applied to an unseen scene. So, eight different models were trained here, each with the data from the city being tested excluded from the train set. Training on images coming from the same city would obviously improve the performance of the method in this city drastically. Fig. 16 shows the MAE, RMSE, and median of the prediction error of the ground range nDSMs for all of the 51 SAR acquisitions as a bar chart. The dashed horizontal lines indicate the corresponding average value over all the scenes coming from one city. It can be seen that the achievable error depends not only on the acquisition geometry of the SAR image in question but also on the area covered and the nature and arrangement of its objects. A very densely built-up inner city with high-rise buildings like Melbourne's leads to increased error values. Not only is the task more difficult due to the heavy overlapping of the different layovers, but the absolute error budget is also much larger for high-rise buildings. A city like Munich with primarily low- to mid-rise buildings and wide streets is easier to reconstruct, which is also reflected by the error metrics.

Fig. 17 shows exemplarily two reconstructed nDSMs, compared to the given ground truth from LiDAR. The area in Barcelona 17(a) is



(a) SAR input image 1 (ber_609) (b) estimated heights in slant range 1 (c) estimated nDSM in ground range 1

(d) SAR input image 2 (ber_312) (e) estimated heights in slant range 2 (f) estimated nDSM in ground range 2

Fig. 14. To see are two SAR images showing the same scene from different orbits with their corresponding height estimation in slant range (middle column) and the final nDSMs in ground range after reprojection. The pixels containing no data were set to zero for better visualization.

Table 4
Comparison in MAE of all available SAR scenes from Berlin. The relationship between error in slant and ground range highly depends on the looking/incidence angle of the acquisition in question. All the images were cropped to the same area to create a fair comparison.

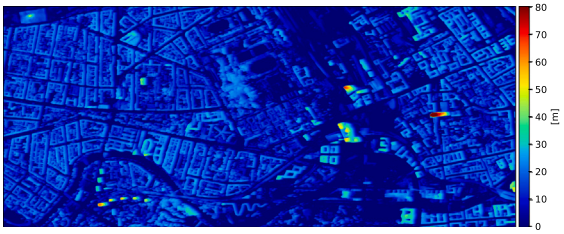
Image	Mode	Inc. Angle	$MAE_{slant\ range}$ [m]				$MAE_{ground\ range}$ [m]			
			Overall	<10 m	10–30 m	>30 m	Overall	<10 m	10–30 m	>30 m
ber_426	ST	22°	3.70	3.47	3.61	11.55	8.36	11.31	4.24	8.95
ber_452	ST	30°	3.06	2.54	3.29	8.11	5.97	7.93	3.59	6.36
ber_609	ST	36°	2.81	2.45	2.96	8.21	5.19	6.94	3.28	6.40
ber_612	ST	36°	2.87	2.48	3.04	8.28	5.14	6.81	3.36	6.56
ber_312	ST	41°	3.08	2.32	3.62	9.42	4.29	4.90	3.51	8.61
ber_432	HS	30°	3.60	3.02	3.84	9.45	6.03	7.46	4.12	8.98
ber_553	HS	36°	3.32	2.76	3.62	8.96	5.43	6.65	3.90	8.20

reconstructed very well. Most of the buildings are around 30 m high and they are clearly distinguishable from each other in the SAR image. However, the building outlines are mostly too narrow, as it is difficult for the model to delineate the roof area from the rest of the facade. The areas of the non-reconstructed roof surfaces do in fact not include any values but were set to zero here for better visualization. Fig. 17(b) shows the area around the financial district in Frankfurt, consisting of a series of skyscrapers built closely together. When reconstructing this type of urban design, the method reaches its limits. High-rise buildings are partially recognized as such, but the height is no longer very reliable. In Fig. 18, absolute DSMs are shown for two areas, namely Berlin 18(a) and Munich 18(b), which contain the elevation of the terrain as well. As a product of the estimated reprojected nDSM and the coarse-resolution DTM, the influence of the DTM can be seen. It appears smoother and more washed out than its LiDAR-derived counterpart. Using a higher-resolution terrain model would eliminate this effect.

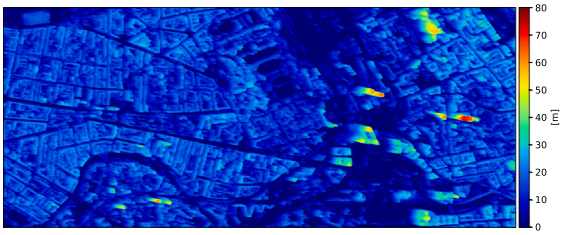
4. Discussion

From the results in the previous section, some key insights can be drawn:

- While even an ordinary U-Net is able to derive the heights of most buildings quite accurately, the adapted network architecture presented in this paper, using the looking angle for a parameter injection, is particularly beneficial for unusually tall buildings.
- SAR images with larger incidence angles tend to yield more accurate height estimates both in slant and ground range: in slant range due to reduced layover effects and clearer building separation and in ground range because of the reduced impact of a wrong height estimation and/or an error in the DTM on the resulting projected DSM.
- As higher incidence angles reduce the length of the layover effects, the maximum building height that can be mapped under very oblique viewing directions (i.e. higher incidence angles) is higher. This is another reason why high incidence angles are advantageous for the type of height mapping presented here.
- However, larger incidence angles cause enlarged radar shadow. This leads to more occluded areas and thus nodata pixels in the DSM. These gaps could be filled using another acquisition from a different aspect, using a data fusion approach.
- The reprojection error depends on the accuracy of the height estimates on the one hand and on the quality of the DTM on



(a) model output using a SAR image with an incidence angle of 41° (ber_312)



(b) model output using a SAR image with an incidence angle of 22° (ber_426)

Fig. 15. Comparison between the model outputs from an image with a large incidence angle (top), and a small incidence angle (bottom). The more shallow acquisition geometry (top) leads to sharper edges and a better reconstruction capability.

the other. Due to the oblique view of the system, pixels with an incorrect height are also shifted in their position in ground range within the process of the projection. Therefore, if a high-resolution DTM is available for the area in question, this should be used to achieve the best possible result. This effect is smaller for images with a large incidence angle.

- The difference in resolution between High Resolution SpotLight and Staring SpotLight imaging modes affects the accuracy of the method only slightly.
- SAR data as a product of an active system, in contrast to optical images, represents a very controlled commodity that allows a DL model to generalize quite easily. Even though the data set used here consists of only eight areas and also has a clear bias towards westernized cities, the model already generalizes surprisingly well. In our opinion, expanding the data set would provide a further enormous boost in performance.
- Cross-validation experiments demonstrated the method's adaptability across various urban environments, with achievable accuracy influenced by city characteristics, building density, and object arrangement. In very densely built-up cities consisting of many skyscrapers, the method reaches its limits. In the intensity image, the facades of the individual buildings are too strongly mixed to reliably assign them to the respective objects. Compare the SAR image in Fig. 19(a) with the corresponding height image in 19(b). Even knowing the ground truth, it is difficult to recognize the facades.
- As can be concluded from the results, unusually large and/or unusually shaped buildings tend to cause larger errors. This is most likely due to the fact that they rarely occur in the training data set. The same holds for forested or heavily vegetated areas. The model has learned the concept of forest exclusively from urban parks, which represent only a small proportion of the data set compared to built-up areas. Closed canopies in particular represent a limiting factor, as they allow no or only very limited evaluation of layover effects.
- Especially the edges of buildings often look a bit fuzzy and wobbly in the reprojected DSM after filtering. Here, smart learned post-processing filters could be added to correct these deficiencies, as found in Stucker and Schindler (2022) or Bittner et al. (2018).
- The provision of height estimates, even in slant range, significantly enhances the interpretability of synthetic aperture radar (SAR) scenes, particularly for users lacking specialized training. Refer to Figs. 14(a) and 14(b), and 14(d) and 14(e) for comparison. While discerning objects within SAR images typically necessitates the expertise of seasoned users, the height image, even in the hands of a non-expert, offers readily comprehensible insights.

Furthermore, the achieved absolute accuracy of the method remains to be assessed. It is important not to be misled by the numerical values of the error metrics: On the one hand, the SAR recordings and ground truth data are not temporally aligned. This means that newly built or already demolished buildings are included in the ground truth, but not in the SAR image. This artificially raises the error metrics. Furthermore, there are certain characteristics of buildings that make them almost invisible to a SAR sensor. Figs. 19(c) and 19(d) show an example of such a case. The Diagonal ZeroZero skyscraper in Barcelona has a glass facade without any horizontal struts. In the SAR image, only the roof surface can be made out, but no facade. Single-image height estimation from SAR imagery understandably reaches its limits here, since the input data does not include the necessary information. From the literature on optical single image height estimation, lower achieved RMSE values are known. This is very likely due to the fact that no reprojection errors further influence the predictions of the models when using orthophotos as input and also due to the sharper edges of optical images compared to the noise-rich SAR data. Not least, the datasets used also affect the model's performance. These datasets are mostly very small-scale and of rather rural character, like the ISPRS dataset of Vaihingen, a small German village mostly consisting of free-standing single-family homes, which is easier to reconstruct than metropolises. Additionally, the frequently found random split between train and test set artificially drives up the performance scores compared to a truly unseen scene since the distributions of train and test data match strongly.

In summary, our results show that our SAR2Height framework can provide a viable alternative to conventional 3D reconstruction methods. While LiDAR scanning or optical stereo certainly provide the best accuracies, they are relatively complicated, time-consuming, and expensive, not least due to their dependency on favorable weather conditions. In the SAR realm, classic InSAR comes to a limit in urban areas due to phase mixtures in the layover areas, while TomoSAR requires very large stacks of coherent images which makes it most time-consuming and expensive. Thus, especially when speed is critical and accuracy is less important, single-image-based height reconstruction from SAR will shine and can no longer be ignored.

5. Summary and conclusion

In this paper, the SAR2Height framework was described that allows to derive a georeferenced digital surface model in an orthometric projective coordinate system of urban areas from a single SAR intensity image. For this purpose, a new convolutional neural network architecture based on the well-known U-Net was proposed, which incorporates sensor knowledge. SAR-specific pre- and postprocessing methods were introduced which make the framework operational. Using it, seamless digital surface models can be generated. The dataset comprises over 50 images from 8 different cities. The method was extensively tested, examined for shortcomings, and its performance was put into perspective. It was shown that the model is able to generalize well to unseen areas and to reconstruct cities with reasonable accuracy. Too dense arrangements of skyscrapers represent a limitation since individual layovers can no longer be divided in the intensity data. The method has the unique ability to derive the geometry of an urban area from just one acquisition independent of light and weather, adding immense value to time-critical applications.

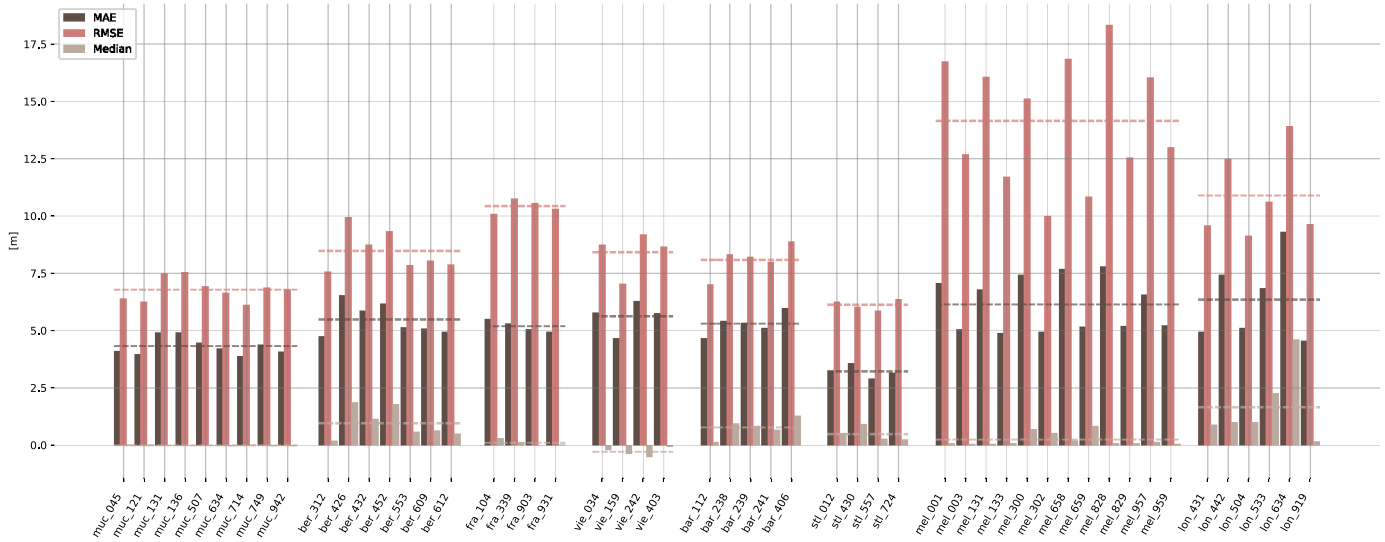
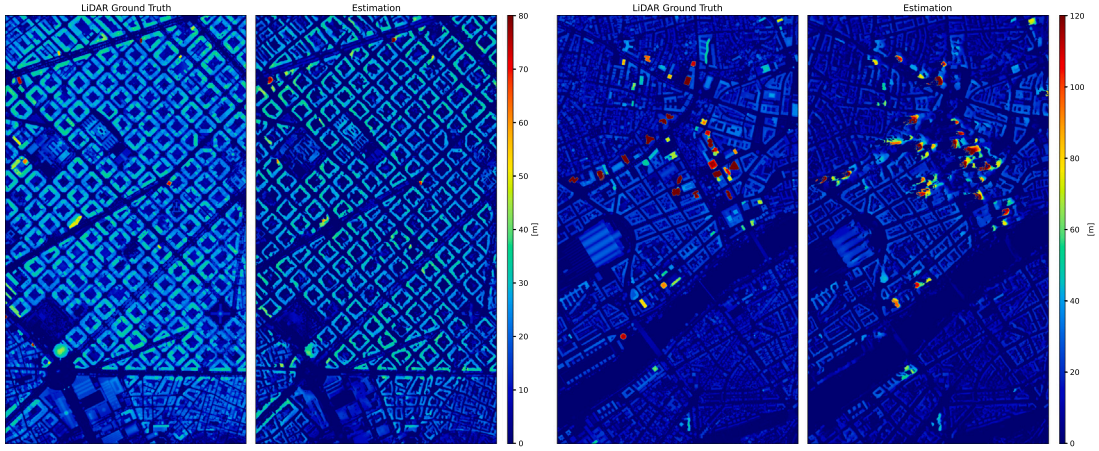
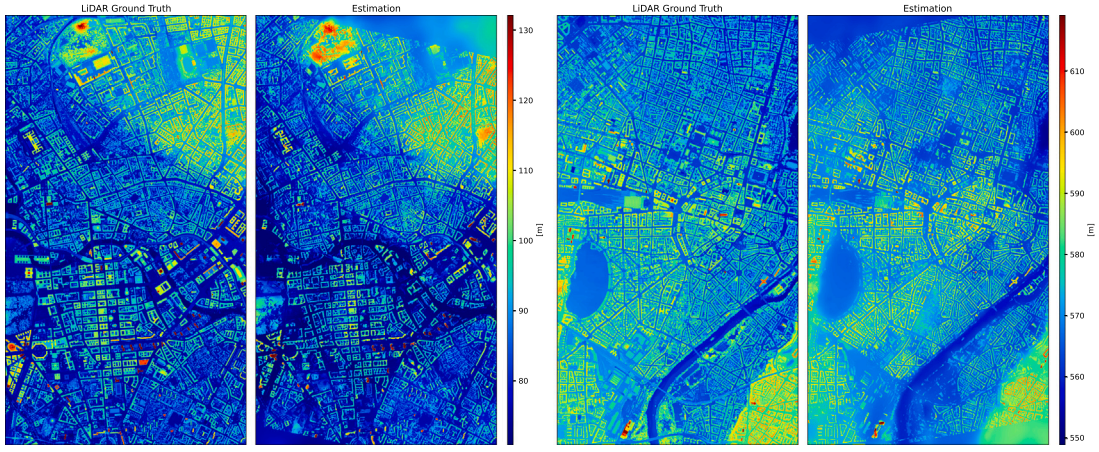


Fig. 16. Error metrics for all available SAR images in the data set. Shown are the individual RMSE, MAE, and the median of the pixel-wise errors of the scenes in question. The dashed lines represent the respective mean over all scenes in a city. It can be seen that the achievable accuracy of the method is highly dependent on the area being estimated. A city like Melbourne, with numerous skyscrapers built close together, increases the error significantly.



(a) nDSM for an area in Barcelona (b) nDSM for the financial district in Frankfurt

Fig. 17. Two generated normalized surface models (nDSMs) for Barcelona (left) and Frankfurt (right) with their corresponding ground truth from LiDAR. While Barcelona can be reconstructed very well, the method reaches its limits with the skyscrapers in Frankfurt. The individual buildings can no longer be accurately mapped. Pixels with no data are filled with a height value of zero for the purpose of better visualization.



(a) DSM for an area in Berlin (b) DSM for the inner city of Munich

Fig. 18. Two generated Surface Models (DSMs) for Berlin (left) and Munich (right) with their corresponding ground truth from LiDAR. The smoother or more washed-out character of the estimation is due to the coarse-resolution DTM underneath. The height values represent absolute measures above a reference ellipsoid. Pixels with no data are filled with values from the DTM.

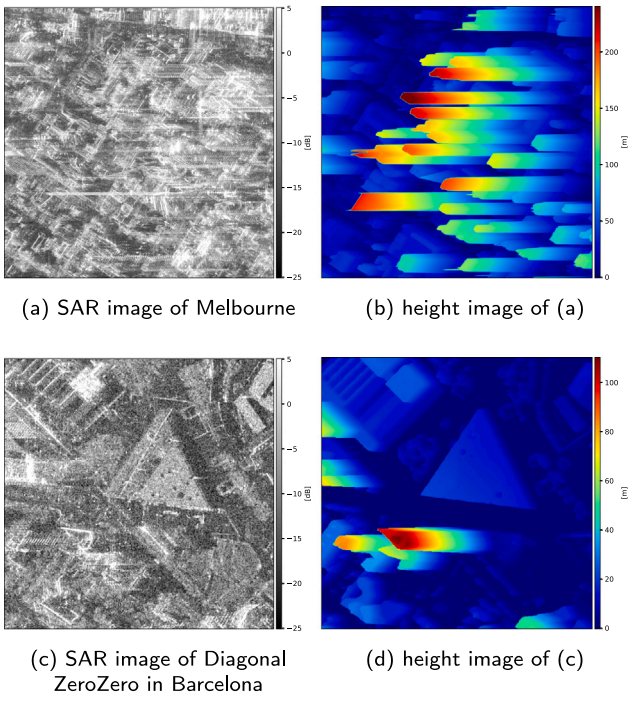


Fig. 19. Two example pairs of SAR image and ground truth height image to showcase the limitations of the method. The top row shows an area of Melbourne's inner city, consisting of densely-built skyscrapers. The layovers are too mixed to separate them reliably. The bottom row shows a skyscraper in Barcelona, the Diagonal ZeroZero. Due to its glass facade without any horizontal struts, it is basically invisible to the SAR sensor (the roof surface can be seen by an experienced observer).

CRediT authorship contribution statement

Michael Recla: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Michael Schmitt:** Conceptualization, Funding acquisition, Investigation, Project administration, Resources, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This project was supported by the German Research Foundation (DFG project SUSO, grant SCHM 3322/3–1) and Airbus Defence and Space (project SAR2Height). The SAR imagery shown in the paper was partially provided by the German Aerospace Center (DLR) in the frame of the proposal MTH3753, and partially by Airbus Defence and Space. The authors gratefully acknowledge the computing time granted by the Institute for Distributed Intelligent Systems and provided on the GPU cluster Monacum One at the University of the Bundeswehr Munich.

References

Akiki, R., Marí, R., De Franchis, C., Morel, J.-M., Facciolo, G., 2021. Robust rational polynomial camera modelling for SAR and pushbroom imaging. In: Proceedings of the IEEE International Geoscience and Remote Sensing Symposium. IGARSS, pp. 7908–7911.

Amirkolaee, H.A., Arefi, H., 2019. Height estimation from single aerial images using a deep convolutional encoder-decoder network. ISPRS J. Photogramm. Remote Sens. 149, 50–66.

Amirkolaee, H.A., Arefi, H., 2021. Generating a highly detailed DSM from a single high-resolution satellite image and an SRTM elevation model. Remote Sens. Lett. 12 (4), 335–344.

Bittner, K., D'Angelo, P., Körner, M., Reinartz, P., 2018. DSM-to-LoD2: Spaceborne Stereo digital surface model refinement. Remote Sens. 10 (12), 1926.

Cao, Y., Huang, X., 2021. A deep learning method for building height estimation using high-resolution multi-view imagery over urban areas: A case study of 42 Chinese cities. Remote Sens. Environ. 264, 112590.

Chen, S., Mou, L., Li, Q., Sun, Y., Zhu, X.X., 2021. Mask-height R-CNN: An end-to-end network for 3D building reconstruction from monocular remote sensing imagery. In: Proceedings of the IEEE International Geoscience and Remote Sensing Symposium. IGARSS, pp. 1202–1205.

Chen, L.-C., Papandreou, G., Schroff, F., Adam, H., 2017. Rethinking atrous convolution for semantic image segmentation. arXiv:1706.05587.

Curlander, J.C., 1982. Location of spaceborne SAR imagery. IEEE Trans. Geosci. Remote Sens. GE-20 (3), 359–364.

Ghamisi, P., Yokoya, N., 2018. IMG2DSM: Height simulation from single imagery using conditional generative adversarial net. IEEE Geosci. Remote Sens. Lett. 15 (5), 794–798.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Identity mappings in deep residual networks. In: Proceedings of European Conference on Computer Vision. ECCV, pp. 630–645.

Kang, S., Uchida, S., Iwana, B.K., 2021. Tunable U-Net: Controlling image-to-image outputs using a tunable scalar value. IEEE Access 9, 103279–103290.

Karatsiolis, S., Kamilaris, A., Cole, I., 2021. IMG2nDSM: Height estimation from single airborne RGB images with deep learning. Remote Sens. 13 (12), 2417.

Li, Q., Mou, L., Hua, Y., Shi, Y., Chen, S., Sun, Y., Zhu, X.X., 2023. 3DCentripetalNet: Building height retrieval from monocular remote sensing imagery. Int. J. Appl. Earth Obs. Geoinf. 120, 103311.

Li, X., Wang, M., Fang, Y., 2022. Height estimation from single aerial images using a deep ordinal regression network. IEEE Geosci. Remote Sens. Lett. 19, 6000205.

Liu, C.-J., Krylov, V.A., Kane, P., Kavanagh, G., Dahyot, R., 2020. IM2ELEVATION: Building height estimation from single-view aerial imagery. Remote Sens. 12 (17), 2719.

Mou, L., Zhu, X.X., 2018. IM2HEIGHT: Height estimation from single monocular imagery via fully residual convolutional-deconvolutional network. arXiv:1802.10249.

Müller, K., Leppich, R., Geiß, C., Borst, V., Pelizari, P.A., Kounev, S., Taubenböck, H., 2023. Deep neural network regression for normalized digital surface model generation with Sentinel-2 imagery. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 16, 8508–8519.

Pınar Örnek, E., Mudgal, S., Wald, J., Wang, Y., Navab, N., Tombari, F., 2022. From 2D to 3D: Re-thinking Benchmarking of Monocular Depth Prediction. arXiv:2203.08122.

Recla, M., Schmitt, M., 2022. Deep-learning-based single-image height reconstruction from very-high-resolution SAR intensity data. ISPRS J. Photogramm. Remote Sens. 183, 496–509.

Recla, M., Schmitt, M., 2023a. From relative to absolute heights in SAR-based single-image height prediction. In: Proceedings of the Joint Urban Remote Sensing Event. JURSE, 10144199.

Recla, M., Schmitt, M., 2023b. Improving deep learning-based height estimation from single SAR images by injecting sensor parameters. In: Proceedings of the IEEE International Geoscience and Remote Sensing Symposium. IGARSS, pp. 1806–1809.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional networks for biomedical image segmentation. In: Proceedings of the Medical Image Computing and Computer-Assisted Intervention. MICCAI, Springer International Publishing, pp. 234–241.

Stucker, C., Schindler, K., 2022. ResDepth: A deep residual prior for 3D reconstruction from high-resolution satellite images. ISPRS J. Photogramm. Remote Sens. 183, 560–580.

Sun, Y., Mou, L., Wang, Y., Montazeri, S., Zhu, X.X., 2022b. Large-scale building height retrieval from single SAR imagery based on bounding box regression networks. ISPRS J. Photogramm. Remote Sens. 184, 79–95.

Sun, W., Zhang, Y., Liao, Y., Yang, B., Lin, M., Zhai, R., Gao, Z., 2022a. Rethinking monocular height estimation from a classification task perspective leveraging the vision transformer. IEEE Geosci. Remote Sens. Lett. 19, 6518705.

Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E., 2004. Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. 13 (4), 600–612.

Xing, S., Dong, Q., Hu, Z., 2022. Gated feature aggregation for height estimation from single aerial images. IEEE Geosci. Remote Sens. Lett. 19, 6003705.

Xue, M., Li, J., Zhao, Z., Luo, Q., 2022. SAR2HEIGHT: Height estimation from a single SAR image in mountain areas via sparse height and proxyless depth-aware penalty neural architecture search for Unet. Remote Sens. 14 (21), 5392.

Zhang, Q., Ge, L., Hensley, S., Isabel Metternicht, G., Liu, C., Zhang, R., 2022b. PolGAN: A deep-learning-based unsupervised forest height estimation based on the synergy of PolInSAR and LiDAR data. ISPRS J. Photogramm. Remote Sens. 186, 123–139.

Zhang, F., Yan, M., Hu, C., Ni, J., Zhou, Y., 2022a. Integrating coordinate features in CNN-based remote sensing imagery classification. IEEE Geosci. Remote Sens. Lett. 19, 5502505.