ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume X-2-2024
ISPRS TC II Mid-term Symposium “The Role of Photogrammetry for a Sustainable World”, 11–14 June 2024, Las Vegas, Nevada, USA

Deep Learning-based DSM Generation from Dual-Aspect SAR Data

Michael Recla, Michael Schmitt

Department of Aerospace Engineering, University of the Bundeswehr Munich, Neubiberg, Germany
(michael.schmitt, michael.recla)@unibw.de

Commission II/WG 3

Keywords: Deep Learning, Synthetic Aperture Radar (SAR), 3D Reconstruction, Radargrammetry, DSM Generation

Abstract

Rapid mapping demands efficient methods for a fast extraction of information from satellite data while minimizing data requirements. This paper explores the potential of deep learning for the generation of high-resolution urban elevation data from Synthetic Aperture Radar (SAR) imagery. In order to mitigate occlusion effects caused by the side-looking nature of SAR remote sensing, two SAR images from opposing aspects are leveraged and processed in an end-to-end deep neural network. The presented approach is the first of its kind to implicitly handle the transition from the SAR-specific slant range geometry to a ground-based mapping geometry within the model architecture. Comparative experiments demonstrate the superiority of the dual-aspect fusion over single-image methods in terms of reconstruction quality and geolocation accuracy. Notably, the model exhibits robust performance across diverse acquisition modes and geometries, showcasing its generalizability and suitability for height mapping applications. The study’s findings underscore the potential of deep learning-driven SAR techniques in generating high-quality urban surface models efficiently and economically.

1. Introduction

In the context of rapid mapping, it is essential to employ methods that extract information from satellite data at an increasing speed with minimal data demands. A great variety of methods have been developed in this regard for the generation of high-resolution elevation data thanks to the rise of deep learning approaches, above all single-image methods. There, the advances are not limited to data from the optical spectrum. Meanwhile, even single SAR images can be used to generate digital surface models (DSMs), as demonstrated for coarse-resolution mountainous areas (Xue et al., 2022), and even for complex urban environments by using VHR SAR data (Recla and Schmitt, 2022, 2024). While providing surprisingly accurate results, the inherent oblique view of SAR results in extensive shadowing, depending on the (urban) topography of the scene under observation. Figure 1 illustrates the problem schematically: if only the image of the right-hand satellite were available, the shorter building would disappear in the shadow of the taller one. By adding another aspect, data gaps can be filled and redundant information can be combined into more reliable measurements.

The idea of using multiple SAR acquisitions of one scene from different aspects to reconstruct its surface is generally not new: For instance, StereoSAR, the absolute height determination of strong backscatterers, and Interferometric SAR (InSAR), the evaluation of phase differences from very similar viewing settings, can be combined to generate absolute DSMs (Eldhuset, 2017). In very complex urban scenes, however, classical InSAR reaches its limits. Overlapping layover effects and extended shadow areas prevent reliable phase unwrapping. More sophisticated methods using multiple aspects to reduce shadow areas and multiple baselines to mitigate the phase ambiguity problem are necessary to reliably reconstruct urban areas (Schmitt et al., 2014). However, these methods are very computationally intensive and data hungry, oftentimes not even feasible using spaceborne missions, where only ascending and descending passes are possible. Another approach to determining building heights from SAR images is to simulate the backscatter of buildings with different heights and then compare the simulation to the actual measurements (Brunner et al., 2009). Experiments with multiple aspects were also carried out using this procedure. Overall, however, this method is again very computationally expensive, not suitable for complex building shapes, and therefore not really applicable for large-scale campaigns.

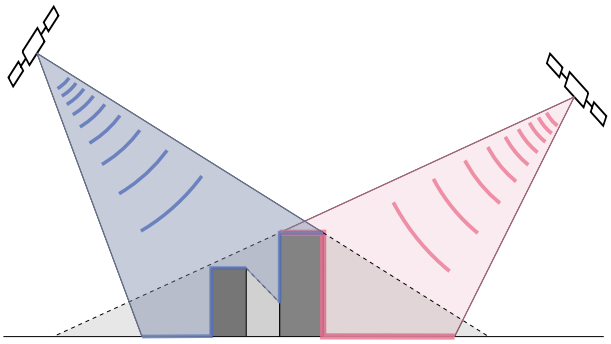
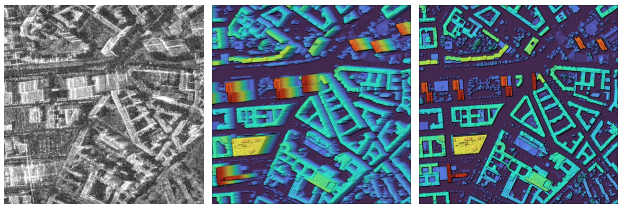


Figure 1. High buildings lead to extensive radar shadows in urban areas. If the buildings are located close to each other, small buildings might even be invisible to a SAR sensor observing the scene from only one aspect angle.

Hence, there is a lack of data-efficient and more generic methods for height extraction from multi-aspect SAR, considering that some work has already been published for optical data in this regard. Deep learning-driven multi-view stereo matching was demonstrated for aerial imagery by (Liu and Ji, 2020) or (Yu et al., 2021). But even multi-aspect satellite data is processed with deep learning techniques, as to find in (Cao and Huang, 2021) or in (Jabbar and Taj, 2024). However, to the best of the authors’ knowledge, no deep learning-based end-to-

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-2-2024-193-2024



(a) SAR intensity image (b) Pixel-wise height map in radar geometry (c) Heights in map projection

Figure 2. Comparison of a SAR intensity image with its pixel-wise height map (middle) and the corresponding height map in an orthonormal projection (right). The geometric distortion effects of layover are corrected in the map projection, and the building facades are no longer visible.

end method for dual- or multi-aspect 3D reconstruction from SAR data has ever been described. And yet SAR systems, due to their all-weather capability, are particularly well-suited for rapid mapping applications such as reconnaissance or disaster management.

In this paper, we present such an approach, using two space-borne SAR images from opposing orbit directions to reconstruct urban sceneries. The model will estimate the height values directly in ground range, i.e. an orthometric projected map system, which is why we refer to the approach as end-to-end. This differs from the procedure in (Recla and Schmitt, 2024), where the heights per pixel are estimated in the sensor-specific imaging geometry, slant range, and then transferred to world space in a post-processing step by a geocoding procedure. Refer to Figure 2: Without the correction for elevated objects, the building facades appear as if they were lying on the ground (2b) due to the oblique view and imaging geometry of SAR. In an orthonormal ground range representation (2c), the view appears from directly above with geometric distortions rectified. This offers the great advantage that the loss during training can be applied directly to the final height map and not to an intermediate result. This means that there is no need for any reprojection, which would again introduce gaps and errors in the localization of the objects due to incorrectly predicted heights into the result. Furthermore, it speeds up the processing.

2. Height Estimation from Ascending and Descending SAR Images

The overall workflow for the generation of digital elevation models from two SAR images as proposed in this paper is shown schematically in Figure 3 in the form of a flow chart. The entire block up to the georeferenced image data is dealt with in Section 2.1, the component relating to the deep learning model in Section 2.2.

2.1 Data Preparation

In order to be able to relate two images from different orbits of the same scene to each other, they have to be mapped to a common reference surface, as the Level 1B SLC (single look complex) data used here are provided in their sensor-specific Zero-Doppler slant range geometry and thus are not georeferenced out of the box. A digital terrain model (DTM) is used for this purpose, which contains no elevated objects such as buildings and vegetation. The projection of the image data onto the DTM is done in a pixel-wise manner using rational polynomial coefficients (RPCs). Setting up an RPC sensor model leads to a more efficient and robust projection procedure and mitigates the use of numerical solvers. To determine the coefficients of the RPC model, a set of (virtual) ground control points (vGCPs) with their corresponding image pixel coordinates is required, which can be generated in any number via the known rigorous sensor model (Range-Doppler): A numerical solver is used to find the position on the DTM (with added random heights) for each vGCP that satisfies the following two equations to do the localization of a pixel (r, c) to a world coordinate system (X, Y, Z). The range equation

$$R = |\boldsymbol{P_S} - \boldsymbol{P_t}| \tag{1}$$

where $\boldsymbol{P_S}$ stands for the sensor's and $\boldsymbol{P_t}$ for the target's position in an earth-centered frame defines a sphere around the sensor with the range R as radius, corresponding to the image's column (range gate). Second, the Doppler equation allows us to find the current azimuth position (i.e. the row in the image raster) and is defined as

$$f_{Dc} = \frac{2}{\lambda R} (\boldsymbol{V_S} - \boldsymbol{V_t}) \cdot (\boldsymbol{P_S} - \boldsymbol{P_t}) , \tag{2}$$

with f_{Dc} as the Doppler centroid frequency, λ the signal wavelength, $\boldsymbol{V_S}$ 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 $\boldsymbol{\omega_e}$ as the Earth's rotational velocity vector. Solving this non-linear set of equations is computationally intensive, it relies on reasonable approximation values and can end up in local minima, which is why it is more advisable to determine the robust polynomial model as a

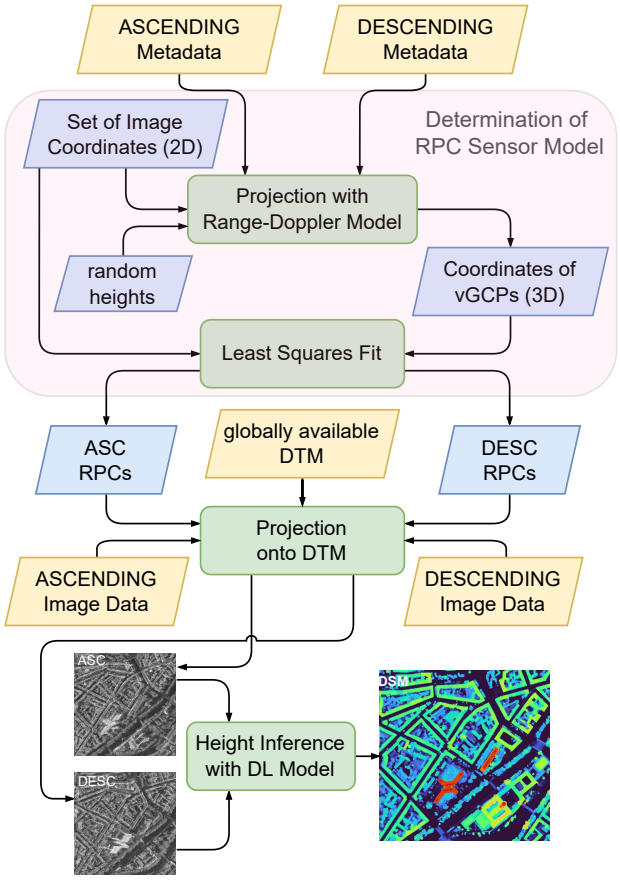


Figure 3. Flowchart of the proposed process for estimating an nDSM from an ascending and a descending SAR scene.

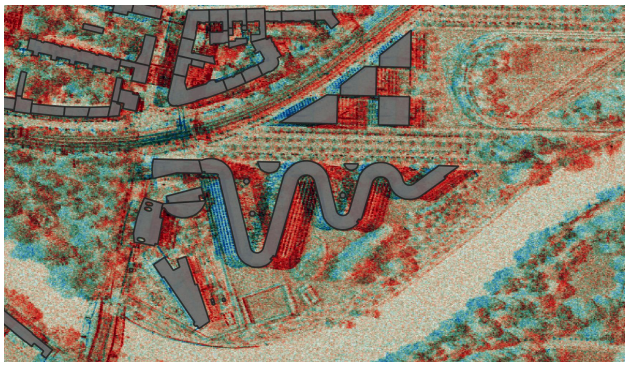


Figure 4. An ascending (white → blue) and a descending (white → red) intensity image projected on a DTM surface are shown for the same location. The building outlines are superimposed in gray. It can be seen how the different imaging geometries affect the resulting layover and shadowing effects in opposing directions.

(very good) approximator.

For a row r and column c, the RPC sensor model is defined as

$$r = \frac{a\,(X,Y,Z)}{b\,(X,Y,Z)} \quad c = \frac{e\,(X,Y,Z)}{f\,(X,Y,Z)}, \tag{3}$$

with X, Y and Z the world-coordinates of a 3D point, and a, b, e and f the cubic polynomials each of the shape

$$\begin{aligned} p(X,Y,Z) = {} & p_0 + p_1 Z + p_2 Y + p_3 X + p_4 ZY + p_5 ZX \\ & + p_6 YX + p_7 X^2 + p_8 Y^2 + p_9 Z^2 + p_{10} ZYX \\ & + p_{11} Z^2 Y + p_{12} Z^2 X + p_{13} Y^2 Z + p_{14} Y^2 X \\ & + p_{15} ZX^2 + p_{16} YX^2 + p_{17} Z^3 + p_{18} Y^3 + p_{19} X^3, \end{aligned} \tag{4}$$

incorporating 20 coefficients p_i. With $p_0 = 1$ for both of the denominator polynomials, a total of 78 coefficients have to be determined by least square fitting for the complete sensor model (Akiki et al., 2021).

Using this sensor model with the previously determined polynomial coefficients, the position of a point on the earth's surface (X, Y, Z) can be robustly converted into sub-pixel-accurate pixel coordinates (r, c) of the corresponding SAR image. To georeference the SAR data, for each point in a regular grid on the earth's surface the pixel in the SAR image is determined and its calibrated intensity value is mapped accordingly to that grid cell. If very-high-resolution slant range images shall be downsampled during the production of the ground-range image, the intensity values of several pixels in a window are averaged, which also results in reduced speckle as a side effect. Objects not included in the DTM, such as buildings and vegetation, are still subject to the geometric distortions of the range-based SAR acquisition principle. Mapping two acquisitions from opposing orbit directions onto a DTM surface in this fashion causes the layover areas (on the facades and roofs) from both images to spread out in the exact opposite directions, with only the base of the building overlapping. Figure 4 is an attempt to illustrate this effect. The ascending and descending images take up the red and blue channels (white → red/blue), while the building footprints are superimposed in black. The length of the layover of a building is proportional to its height. The model should learn these relations from the provided training data by itself. In this sense, the distortions become our measurement signal.

In order to produce annotated training data, the georeferenced SAR images are paired with normalized digital surface models (nDSMs) of the corresponding areas, which are projected to the same reference system and resolution. An nDSM only contains relative heights above ground as a result of the difference between DSM and DTM. The information about the absolute height above sea level is not contained in the raw pixel values of a SAR image but can be re-added after the inference step via the DTM, as this is indispensable for the projection of the SAR images anyway. For model training, the available data is sliced in a 50% overlapping grid to patches of 512×512 pixels in size, and ascending/descending pairs are formed together with the corresponding height data.

2.2 Deep Learning Component

The model architecture employed is a modified form of the very common and widely used U-Net (Ronneberger et al., 2015) with its encoder-decoder structure, as illustrated in Figure 5. The two input images from ascending and descending orbits are fed into the network via two separate heads instead of simply two channels. The resulting feature maps are then concatenated and sent to the decoder. The entire network is comprised of residual blocks (ResBlocks), whose structure is dissected in Figure 6. The blocks in the two input heads differ from the remaining ResBlocks in two respects: On the one hand, the so-called Multi-Scale ResBlocks consist of several successive convolutional layers with different dilation rates, similar to the structure known from *DeepLabv3* (Chen et al., 2017). This is intended to enable the model to obtain a more global view of the input features already in this early stage of the network than conventional 3×3 convolutions would allow. In addition, known sensor parameters of the input images are fed into the two heads, very similar to how it is described in (Recla and Schmitt, 2024). The tangent of the looking angle as well as the cosine and sine of the azimuth angle are used for this purpose. The looking angle has a direct influence on the length of the recorded layover effect of raised objects such as buildings, while the azimuth angle determines the direction of the layover (always towards the sensor). For ascending and descending orbits, the azimuth fluctuates around $90°$ and $270°$ respectively, while the looking angle remains within an interval of approx. $20° - 55°$. Knowing these parameters beforehand should make it easier for the model to understand and reliably interpret the representation of the scenery.

The SAR intensity input data of the size 512×512 pixels is clipped below $-30\,\mathrm{dB}$ and over $10\,\mathrm{dB}$ and then min-max normalized to a range between 0 and 1. The target data, i.e. the nDSM, is scaled through division by $50\,\mathrm{m}$, ensuring that the values to be predicted remain numerically modest. The model is trained in the conventional supervised manner, using the Adam optimizer with a learning rate of 10^{-4} and the $\mathcal{L}_1$ norm as a loss function. 7200 input pairs are randomly drawn from the dataset in every epoch and fed to the model without a certain order of ascending and descending images since the model can draw this information about the acquisition setting already from the azimuth angles as additional parameters. With a batch size of 24, the model is trained for 100 epochs. No data augmentation methods are applied, since there are already approximately 15 different images per city in the dataset with various imaging modes and acquisition geometries, resulting in very different-appearing views of the same sceneries.

Since the model already provides georeferenced ground range detected outputs, the postprocessing effort is very low. First, the

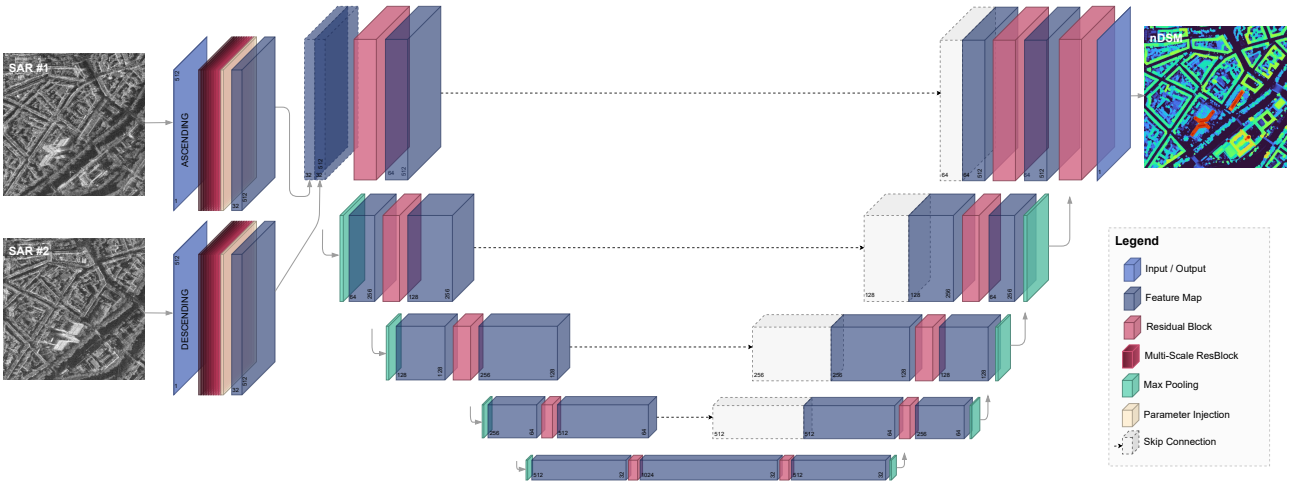


Figure 5. Network architecture: Two SAR intensity images from opposing orbit directions are fed into two separate heads, whose feature maps are then concatenated and sent through a U-Net-like encoder/decoder structure. The output is an nDSM of the same dimensions as the input data, already in a conventional map reference system like UTM (ground range detected).

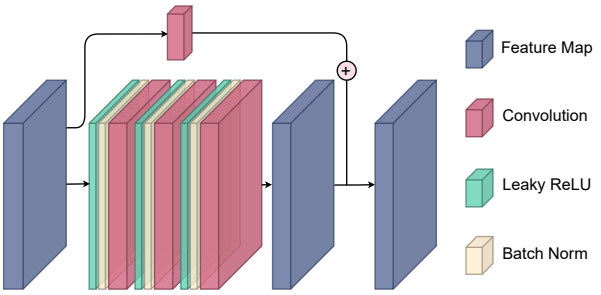


Figure 6. The basic building block of the network architecture shown in Figure 5: the Residual Block (ResBlock). It consists of three consecutive activated batchnorm and convolution layers. The input features are added back to the result.

heights must be denormalized to correspond to a metric measure and, second, the individual image snippets must be reassembled into a coherent mosaic. To avoid unsightly artifacts at their boundaries, the individual images are extracted from the SAR images with an overlapping area and their predictions are then linearly weighted against each other.

3. Experimental Setup and Results

The SAR data used here all originate from the German *TerraSAR-X* satellite. A dataset consisting of 85 individual SpotLight images was compiled and paired with normalized digital elevation models. The data is available for eight different cities, namely Munich, Berlin, Frankfurt am Main, London, Vienna, Barcelona, Melbourne, and St. Louis, and was acquired from different orbits and looking angles using different imaging modes, specifically “normal” SpotLight (SL), High Resolution SpotLight (HS), and Staring SpotLight (ST). StripMap (SM) images were only used for testing the final models. The corresponding elevation data comes exclusively from freely available sources of local authorities. For the training dataset, the image data was projected onto high-resolution terrain models derived from LiDAR campaigns. For the inference phase of the model, these are replaced by globally available, lower-resolution alternatives, such as the *FABDEM* (Hawker et al., 2022), in order to mimic a realistic test case in which LiDAR data is not necessarily available. For all of the following experiments, any data taken over Berlin was excluded from the training. Similar performance scores of the method can thus be expected for similarly developed cities. The images used for the following experiments are listed together with their properties in Table 1.

To evaluate the performance of the proposed methodology numerically, a set of metrics is introduced: The mean absolute error (MAE), the mean value and the median of the discrepancy between prediction and target serve as absolute error metrics for this study. More qualitative metrics are the Pearson coefficient and the Structural Similarity Index Measure (SSIM): Quantifying the magnitude and direction of a linear relation between two variables, the Pearson Correlation Coefficient, often denoted as Pearson’s r, spans from -1 to 1. A value of 1 denotes a perfect positive linear relationship, -1 reflects a perfect negative linear relationship, and 0 signifies no linear correlation. It is given by

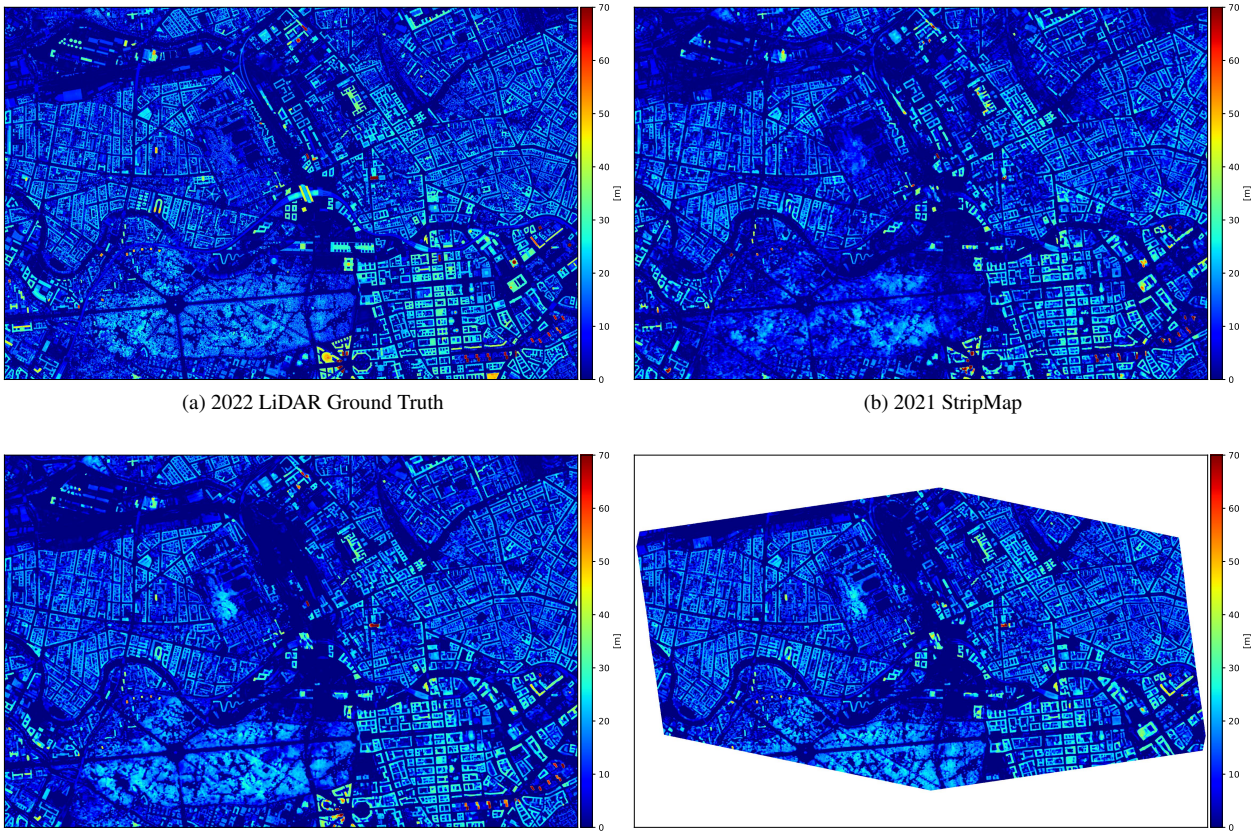
$$r = \frac{\sum_{i=1}^{n}(y_i - \mu_y)(\hat{y}_i - \mu_{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \mu_y)^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \mu_{\hat{y}})^2}}, \tag{5}$$

with y_i and $\hat{y}_i$ being the target and predicted value for pixel i and μ_y and $\mu_{\hat{y}}$ as their corresponding mean values.
In the fields of image processing and computer vision, the SSIM is a metric to quantify the perceptual likeness of an image pair. It factors in luminance, contrast, and structure to mirror human perception. Using Gaussian kernels in local windows, the final measure is the average of these local results. It is defined as

$$\mathrm{SSIM} = \frac{(2\mu_y\mu_{\hat{y}} + C_1)(2\sigma_{y\hat{y}} + C_2)}{(\mu_y^2 + \mu_{\hat{y}}^2 + C_1)(\sigma_y^2 + \sigma_{y\hat{y}}^2 + C_2)}, \tag{6}$$

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 numerically stabilize the division (Wang et al., 2004).

Figure 7 displays a comparison of the results using different acquisition modes with the ground truth from LiDAR. Despite not encountering a single StripMap image during training, the resulting DSM is comparable to those from SpotLight, although



(a) 2022 LiDAR Ground Truth

(b) 2021 StripMap

(c) 2014/2015 HighRes SpotLight

(d) 2016/2018 Staring SpotLight

Figure 7. Comparison between the generated DSMs from different imaging modes and the LiDAR ground truth. A central part of the city of Berlin can be seen. The higher level of detail of the DSMs from the higher-resolution SpotLight images is noticeable. In the central north, several newly constructed buildings draw the eye, which are still absent in the image pair from 2015, but are clearly visible in the one from 2021. Figure 8 shows a further comparison of this area.

Image	Imaging Mode	Orbit Direction	Incidence Angle	Acquisition Date
HS#1	HS	ASC	42°	28.03.2014
HS#2	HS	DESC	36°	12.03.2015
ST#1	ST	ASC	42°	23.12.2016
ST#2	ST	ASC	30°	17.09.2018
ST#3	ST	DESC	36°	17.08.2018
SM#1	SM	ASC	29°	06.07.2021
SM#2	SM	DESC	35°	15.02.2021

Table 1. Listing of the image pairs used for the experiments.

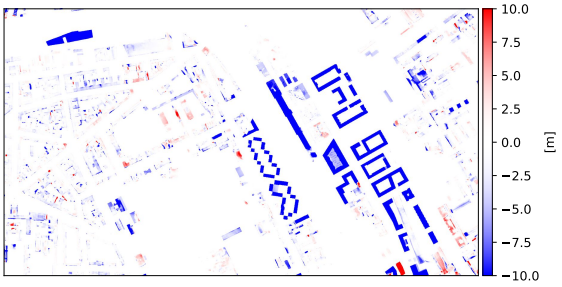
the level of detail is reduced slightly. However, the urban structure is still easily recognizable, and also unusually tall buildings are reliably mapped. The zone north of Berlin Central Station is particularly interesting. A number of new buildings have apparently been built in this area, which are included in the ground truth data from 2022. When comparing the HS data from 2015 and the 2021 DSMs from SM, this development is clearly recognizable and directly comparable across different resolutions, recording modes, and geometries. In Figure 8, error maps of the zoomed-in view of the new construction zone can be seen. These error maps were masked with building footprints from *OpenStreetMap* (OSM) to illustrate the effect of the missing buildings. While a large number of buildings were still missing in 2015, the error map from the 2018 image pair shows precisely which buildings had already been built by that time. The numerical evaluations of the various DSMs generated can be found in the lower part of Table 2. In order to guarantee a fair comparison, the metrics were only collected for the overlapping areas of all settings and the previously discussed construction site was masked out. All metrics improve with higher-resolution input data. However, the recording geometry has an impact as well. The 42° Staring SpotLight image (ST#1) has a positive effect on the quality of the resulting DSM. Higher incidence angles seem to be favorable in urban areas due to the reduced layover effect.

In order to examine whether the end-to-end model presented here for the fusion of ascending and descending images offers significant advantages over the single image case, another model is introduced with only one input head, taking a single image as its only input. It is trained with identical data and hyperparameters for the same number of steps in order to ensure a fair comparison. Two HS scenes depicting the center of Berlin

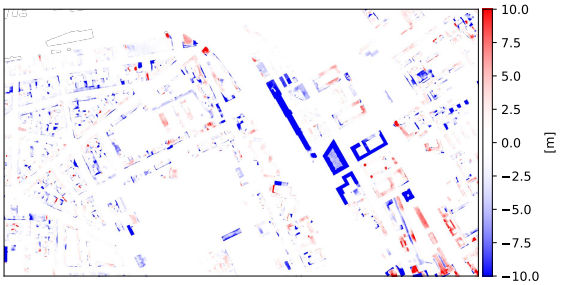
Setting	**MAE ↓ [m]**	**Mean ↓ [m]**	**Median ↓ [m]**	**Pearson ↑**	**SSIM ↑**
Single HS ASC	4.42	−1.09	−0.03	0.68	0.85
Single HS DESC	4.76	−0.86	−0.04	0.64	0.83
Late Fusion HS	4.53	1.21	0.18	0.70	0.84
Dual HS	3.54	−0.42	0.03	0.77	0.87
Dual SM	4.09	−1.05	0.03	0.73	0.85
Dual ST (#2 & #3)	3.72	0.30	0.05	0.77	0.88
Dual ST (#1 & #3)	3.34	0.03	0.04	0.80	0.89

Table 2. Numerically obtained error metrics of the different experiments. The upper part compares the DSMs generated from single images and their late fusion. In the lower part, the results from different image pairs, originating from different acquisition modes and geometries, are being compared with each other.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-2-2024-193-2024 | © Author(s) 2024. CC BY 4.0 License.



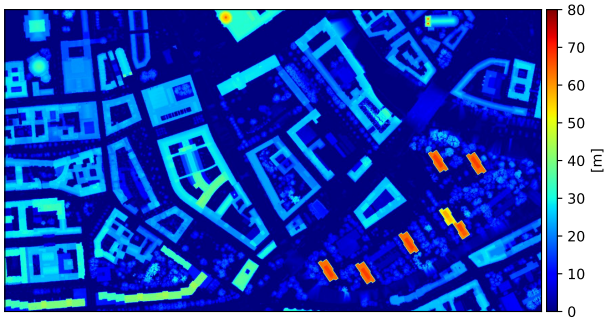
(a) 2015 HighRes SL Error Map



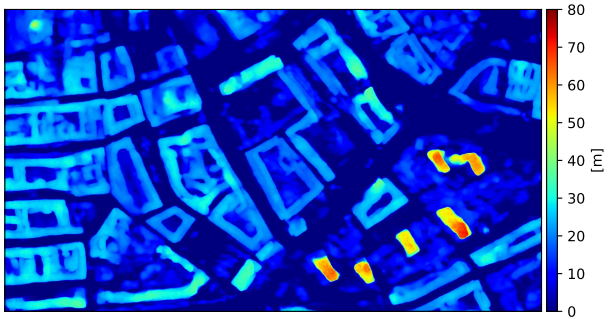
(b) 2018 Staring SL Error Map

Figure 8. Detail of a new development area. The newly constructed buildings contained in the (current) ground truth data lead to large differences in the error maps. These were masked with OSM building footprints. The pair of images from 2018 clearly shows which buildings had already been built by then and which were constructed later.

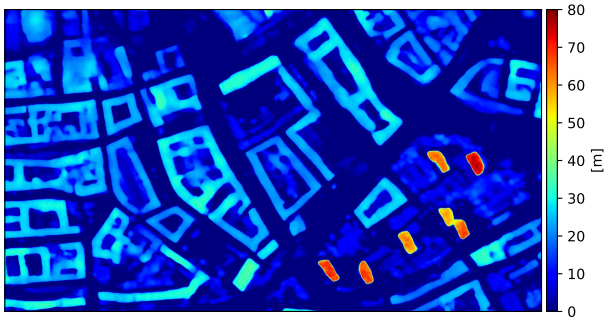
are used as test images. The two DSMs generated from the individual images are then fused with the following strategy: The single-image height maps are coregistered and the maximum height value occurring for each pixel is selected. For instance, if a building is obscured in one image due to radar shadow, it should theoretically be identified in the opposite view where the estimated height would be higher. This is called a “late fusion”, as the results are only merged at decision level, in contrast to “early fusion” of the raw sensor images as implemented in the method proposed by this paper. One would assume a similar result for both methods. However, in practice, the additional image already within the model not only helps to fill gaps, but also has a positive effect on the reconstruction quality of the entire scene. Figure 9 shows a comparison of the results from late fusion (9b) and early fusion (9c). Due to the less precise positioning of the individual buildings in the single image case, the fusion results in blurred edges and a reduction in the quality of the building shapes. The height values are also less accurate and the model tends to underestimate building heights. The upper part of Table 2 reflects this in the numerical results. Compared to the first row of the lower part (DUAL HS), the results are significantly worse, both the DSMs from the individual images and their late fusion. Figure 11 emphasizes these observations. An unusually shaped building was reconstructed in three different ways: Once from only one HS ascending image (11b), once from one HS descending image (11c), and once using early fusion of the ascending and descending acquisitions together (11d). The fusion result is much closer to the ground truth than the DSMs from the individual images. The additional information available to the model through the simultaneous evaluation of both views helps to better capture the structure of the buildings and more reliably depict the transition from slant range to ground range.



(a) LiDAR Ground Truth



(b) Late Fusion: The single-image derived DSMs are merged by taking the individual maxima per pixel during post-processing.



(c) Early Fusion (*this paper*): The two input images are processed jointly within the model.

Figure 9. Comparison between the LiDAR ground truth (top), the late fusion result of the single-image-derived DSMs (middle), and the DSM resulting from the early fusion of ascending/descending images presented in this paper (bottom). Not only the heights but also the buildings’ positionings and their edges are significantly superior in the case of early fusion.

4. Discussion

As the results described in Section 3 show, the model seems to have learned to interpret the SAR data’s underlying imaging geometry. This is likely due to the large number of different imaging scenarios of the same scenes during training. A high-rise building primarily leads to high intensity values along its layover. However, the individual bright pixels should not be assigned a large height in the resulting DSM, but only those pixels within the building footprint. While this task seems far from straightforward, the model has learned to perform it with surprising accuracy. It performs the transition from slant-range geometry to ground-range geometry implicitly. The error metrics support the qualitative impression of the generated DSMs. With a mean absolute deviation of less than $4\,\mathrm{m}$ and an SSIM of greater than 0.85, analysis-ready elevation models are produced

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-2-2024-193-2024 | © Author(s) 2024. CC BY 4.0 License.

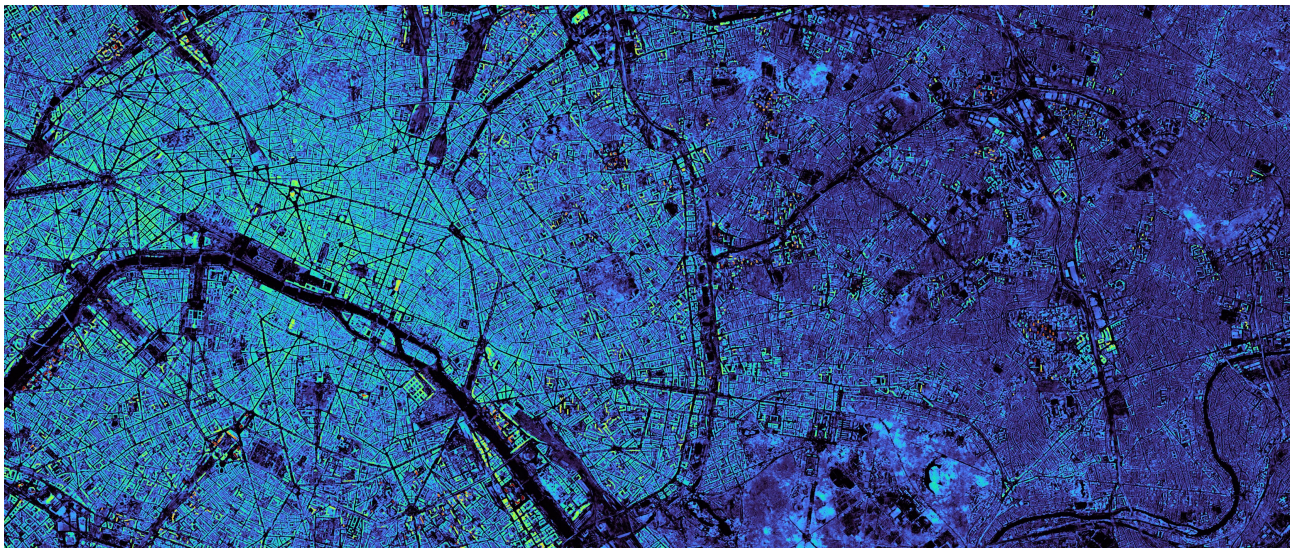
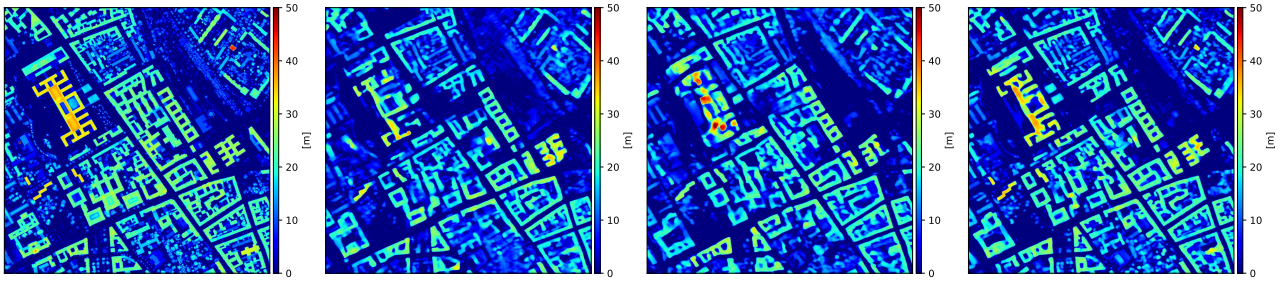


Figure 10. nDSM of Paris, generated from two StripMap scenes. Although the model was trained exclusively on SpotLight images, it shows impressive transfer capabilities on StripMap data. This example is a testimony to how this model can already be used to generate very large-scale DSMs in a computationally inexpensive and data-lean manner.



(a) LiDAR Ground Truth (b) Only Ascending (c) Only Descending (d) Asc/Desc Fusion

Figure 11. Visual comparison between the LiDAR DSM, the DSMs generated from single images, and the fusion result. The example of this unusually shaped building clearly shows that the fusion is more than the sum of the two individual results. The additional information simultaneously available to the model for interpretation not only leads to more reliable height values but above all to significantly better building outlines.

with only two SAR images. The example of the newly constructed buildings in Figure 8 shows how such a method can be used to detect changes in the urban topography independent of the recording geometry and mode. This can be highly useful for rapid mapping applications such as disaster management after natural catastrophes. However, when using two acquisitions with different timestamps, it remains difficult to understand with this method what information the generated DSMs reflect. This is where the black-box nature of end-to-end deep learning methods becomes apparent. Notably, with missions like TerraSAR-X, two acquisitions from opposing orbit directions can be made only half a day apart, if the method, like the one presented here, does not require the exact same looking angles for both images.

Although the dual-aspect variant loses some flexibility and transparency compared to the single-image method, the results from Section 3 leave no doubt that for this type of end-to-end model, the use of two images from opposite views leads to a significant performance improvement in the reconstruction quality of the observed area. The results of the experiments suggest that the type of data fusion presented here provides better results than if the incoming data were analyzed individually with their results being merged. The fusion is thus more than just the sum of its parts (Schmitt and Zhu, 2016); refer to Figure 11 for a comparison. However, it is not only the building shapes that benefit from the joint evaluation of the two views, their geolocalization is also improved as a result. If the terrain model used for the projection of the input data contains errors, which can often occur with the coarse-resolution globally available DTMs, these affect the projection of the two images. If the terrain is too high, a pixel is shifted away from the sensor, if it is too low, it is shifted toward the sensor. Using opposing orbit directions (and identical incidence angles), these errors have exactly the opposite effect for both images. The true position of the pixel would therefore be in the middle. This effect can be seen in the result of the late fusion in Figure 9b. Some buildings appear as if they have been reconstructed twice, with a certain offset in between. These are the localization distortions due to the imperfect DTM of the two individual images. However, when the images are jointly processed, the model corrects this influence and the buildings are overall closer to the high-precision ground truth (compare Figure 9c).

The scores in Table 2 still show a fairly significant gap between the results from SpotLight data and those from StripMap. Nevertheless, the transfer performance of the model to this un-

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-2-2024-193-2024 | © Author(s) 2024. CC BY 4.0 License.

known data type is remarkable, as no SM data was used at all during training. If the model were fine-tuned to this data, a further improvement of the DSMs would be very likely. Compared to SpotLight data, SM images are not only much more cost-efficient, their coverage is also many times larger. With archive data from TerraSAR-X, a global DSM for urban areas would theoretically be conceivable. Figure 10 shows a section of the DSM of Paris generated with two SM images. Conventional InSAR methods are not capable of such a level of detail in complex regions with as many layover and shadow areas as urban areas like these.

5. Summary & Conclusion

In conclusion, this paper presents a comprehensive methodology for height estimation in urban areas using two SAR acquisitions taken from opposing orbit directions. The SAR images are mapped to a common reference surface, namely a globally available digital terrain model. The proposed deep learning model, based on a modified U-Net architecture, effectively fuses information from ascending and descending images to generate accurate 3D reconstructions in a common map reference system, like UTM. The conversion from the sensor-specific imaging geometry, slant range, to the orthometric projection, ground range, happens implicitly within the model. Comparisons between single-image and dual-aspect fusion approaches highlight the superiority of the latter in terms of reconstruction quality and geolocalization accuracy. The paper provides a numerical evaluation of the achieved accuracies for different settings, i.e. for different acquisition modes and geometries. As expected, the method performs better with high-resolution SpotLight data than with the StripMap images. However, these are not as far apart as one might expect. The model has a high degree of generalizability to unseen data types and locations. The method presented here proves to be less data-hungry and lightweight compared to conventional technologies such as TomoSAR and thus underscores the potential to generate fast and inexpensive comparably high-quality urban surface models.

Acknowledgement

This project was supported by the German Research Foundation (DFG project SUSO, grant SCHM 3322/3–1). The SAR imagery shown in the paper was provided by the German Aerospace Center (DLR) in the frame of the proposal MTH3753.

References

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

Brunner, D., Lemoine, G., Bruzzone, L., 2009. Estimation of building heights from detected dual-aspect VHR SAR imagery using an iterative simulation and matching procedure in combination with functional analysis. *IEEE Radar Conference*.

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 Sensing of Environment*, 264. Art. no. 112590.

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

Eldhuset, K., 2017. Combination of stereo SAR and InSAR for DEM generation using TanDEM-X spotlight data. *International Journal of Remote Sensing*, 38(15), 4362-4378.

Hawker, L., Uhe, P., Paulo, L., Sosa, J., Savage, J., Sampson, C., Neal, J., 2022. A 30 m global map of elevation with forests and buildings removed. *Environmental Research Letters*, 17(2). Art. no. 024016.

Jabbar, S., Taj, M., 2024. Stereoential Net: Deep network for learning building height using stereo imagery. B. Luo, L. Cheng, Z.-G. Wu, H. Li, C. Li (eds), *Neural Information Processing*, Springer Nature Singapore, 478–489.

Liu, J., Ji, S., 2020. A novel recurrent encoder-decoder structure for large-scale multi-view stereo reconstruction from an open aerial dataset. *Proc. of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 6050–6059.

Recla, M., Schmitt, M., 2022. Deep-learning-based single-image height reconstruction from very-high-resolution SAR intensity data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 183, 496-509.

Recla, M., Schmitt, M., 2024. The SAR2Height framework for urban height map reconstruction from single SAR intensity images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 211, 104-120.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional networks for biomedical image segmentation. *Proc. of the MICCAI*, Springer International Publishing, 234–241.

Schmitt, M., Schönberger, J. L., Stilla, U., 2014. Benefit of Using Multiple Baselines and Multiple Aspects for SAR Interferometry of Urban Areas. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(10), 4107-4118.

Schmitt, M., Zhu, X. X., 2016. Data Fusion and Remote Sensing: An ever-growing relationship. *IEEE Geoscience and Remote Sensing Magazine*, 4(4), 6-23.

Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E., 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4), 600-612.

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 Sensing*, 14(21). Art. no. 5392.

Yu, D., Ji, S., Liu, J., Wei, S., 2021. Automatic 3D building reconstruction from multi-view aerial images with deep learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 171, 155-170.