# COMPARISON OF SINGLE-IMAGE URBAN HEIGHT RECONSTRUCTION FROM OPTICAL AND SAR DATA

Michael Schmitt*, Michael Recla

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

**KEY WORDS:** Remote Sensing, Photogrammetry, Radargrammetry, Single-Image Depth, 3D Reconstruction, Urban Areas

**ABSTRACT:**

Deep learning-based depth estimation has become an important topic in recent years, not only in the field of computer vision. Also in the context of remote sensing, scientists started a few years ago to adapt or develop suitable approaches to realize a reconstruction of the Earth's surface without requiring several images. There are many reasons for this: First, of course, the aspect of general economization, since especially high-resolution satellite images are often accompanied by high data acquisition costs. In addition, there is also the desire to be able to acquire high-quality geoinformation as quickly as possible in time-critical cases – for example, the provision of up-to-date maps for emergency forces in disaster scenarios. Finally, a reconstruction of topography based only on single images can also provide important approximate values for the classic multi-image methods. For example, various processing steps in a classical InSAR process chain require a rough knowledge of the Earth's surface in order to achieve the most accurate and reliable results. In this paper, we review the developments documented in the remote sensing literature so far. Using an established neural network architecture, we produce example results for both very-high-resolution SAR and optical imagery. The comparison shows that SAR-based single-image-height reconstruction seems to bear an even greater potential than single-image height reconstruction from optical data.

## 1. INTRODUCTION

Since it was understood that it is possible to reconstruct the three-dimensional environment with the help of images, there have been efforts to reduce the number of images required for this to a minimum – and this minimum is, of course, a single image. Already in the 1970s, a first approach to this end was developed by computer scientists for the 3D reconstruction of objects at close range. This approach is based on an evaluation of illumination directions and their corresponding shadows, and was therefore called *Shape from Shading*. For an overview of important developments and techniques in this field, the interested reader is referred to the detailed review by (Zhang et al., 1999). In remote sensing, *Shape from Shading* has been adapted primarily for the analysis of SAR images, since these have many well-defined and geometrically easy-to-model shadow regions due to the oblique-viewing image geometry inherent in the SAR technique (Di Martino et al., 2014). However, shape-from-shading approaches have also been implemented for optical data, e.g., to evaluate images from the Mars Express mission (O'Hara and Barnes, 2012), or to reconstruct ice-covered terrain (Cooper, 1994). However, because the results of these methods never reached the quality of classical stereo results, 3D reconstruction from single images then enjoyed a rather shadowy existence for a long time and was mainly used to support classical stereo methods - for example, in the matching of optical multi-view images (Heipke, 1992), or the coregistration of SAR images in classical interferometric SAR processing pipelines (Natsuaki and Hirose, 2012).

With the rise of deep learning, researchers and engineers, especially from the field of robotics and autonomous driving, returned to single-image-based 3D reconstruction. The primary motivation was (and still is) the desire to economize spatial perception for autonomous, AI-based navigation tasks – on the one hand by reducing the number of sensors required for data acquisition, and on the other hand by accelerating automated data evaluation. Since this usually involves the derivation of distances between the navigation subject and the environment (i.e., spatial depth), this research field is referred to in the literature as *single image depth estimation (SIDE)*. For a summary of the developments in the general SIDE field, which mainly refers to the analysis of classical optical images with horizontal viewing direction, interested readers are referred to the overview article by (Mertan et al., 2021).

Recently, remote sensing researchers also started to adapt Deep Learning-based SIDE approaches to the reconstruction of elevation maps from single satellite images. With their model `IM2HEIGHT`, (Mou and Zhu, 2018) have developed a network architecture in which deconvolution layers follow convolution layers to first encode the spatial information present in the input image into an abstract, compressed form, and then decode this abstract form back into an elevation image. In contrast, the `IMG2DSM` approach presented by (Ghamisi and Yokoya, 2018) uses the principle of Generative Adversarial Networks (GANs), in which a generator network is trained to generate realistic artificial height images from satellite data, while a discriminator network is trained to efficiently distinguish artificial height images from real height images. By alternating between the generator and discriminator, this approach produces very high quality result images, although the comparatively difficult training is a disadvantage compared to classical CNN architectures. Since these pioneering prototypes, other research groups have also attempted the development of CNN approaches for reconstructing elevation data from individual aerial and satellite images, e.g. (Amirkolaee and Arefi, 2019, Pellegrin and Martinez-Carranza, 2020). While most of the literature available so far focuses on high-resolution aerial photographs and semi-urban or small-town terrain, the work of (Recla and Schmitt, 2022) is the first to investigate single-image height reconstruction for

* Corresponding author

(a) (b)

Figure 1. Illustration of data used for training and testing in this study. (a) Optical imagery, (b) SAR imagery. The large parts constitute the areas used for training, the smaller rectangles the common test area. For display purposes, the images are not shown to scale.

the second important remote sensing modality – SAR remote sensing. Their results, achieved for densely built-up inner city areas, are found to provide a reconstruction quality similar to that of approaches based on classical optical imagery. With this paper, we intend to provide a comparison of single-image height reconstruction results for very-high-resolution SAR and optical remote sensing images of urban areas. Our intention is to provide a better understanding of the potential of such modern height reconstruction approaches given the different sensor modalities provided by remote sensing.

## 2. MATERIALS AND METHODS

In this section, we describe both the remote sensing data sources and the backbone deep learning approach used in this study. All experiments are carried out on real data.

.

Table 1. Properties of the remote sensing image data used in this work

| **Property** | **Optical Imagery** | **SAR Imagery** |
|---|---|---|
| Type | aerial color orthoimages | TerraSAR-X staring spotlight mode intensity images |
| Source | Geoportal Berlin | German Aerospace Center (DLR) |
| Acquisition date | 2021 | 2018 |
| Resolution | 0.2 m × 0.2 m | 0.24 m × 0.6 m |
| Sampling | 0.7 m × 0.7 m | 1 m × 1 m |

### 2.1 SAR and Optical Remote Sensing Data

The experiments in this paper are based on very-high-resolution SAR and optical remote sensing data depicting the city of Berlin, Germany. The images are shown in Fig. 1, and their specifications are described in Tab. 1.

The height reference for the study area is an image-derived digital surface model provided by the Geoportal Berlin.

### 2.2 Deep Learning-based Single-Image Height Reconstruction

For the experiments shown in this paper, we use a slightly adapted version of the `IM2HEIGHT` architecture proposed by (Mou and Zhu, 2018), whose structure is displayed in Fig. 2. Immediately, a so-called encoder-decoder structure can be recognized. The left half of the network forms the encoder – a series of successively connected processing blocks, which essentially consist of a normalization, a convolution with $3 \times 3$ pixel filters and a nonlinear activation function. After each processing block, the resolution of the resulting images is halved by so-called max-pooling. In this process, only the maximum pixel value is retained in each case in a $2 \times 2$ pixel window slid over the image. These values are then used to compose the input image for the next processing block. In this way, the original image is transformed step by step into increasingly complex representations. In the center of the network, a compact abstraction of the original image is thus created. On the right-hand side of the network, there are processing blocks of almost identical construction, whose contribution is to use so-called deconvolution kernels to transform the compact abstraction back into an image-like representation with the same size as the input image – in the case of the height reconstruction, a 2.5-dimensional height image, i.e. an image in which each pixel contains an individual height value. For this process to work, training must be performed in which the network is shown a large number of sample image pairs – i.e., aerial or satellite image with an exactly matching elevation image. In the case of SAR imagery, which follows a slant-range projection geometry with significant difference between the nadir point and the scene center, we use the projection approach described by (Recla and Schmitt, 2022) to connect height values and SAR image pixels. In the case of optical imagery, which follows a close-to-nadir viewing geometry, we use a simple spatial co-registration of 2.5D height reference map and satellite image for that purpose.

## 3. EXPERIMENTS & RESULTS

This section describes the experiments and results achieved with the materials and methods described in the preceding section. These experiments are aimed at a comparison of single-image height reconstruction capabilities with either SAR or optical remote sensing imagery used as input to the convolutional neural network architecture `IM2HEIGHT` described above.

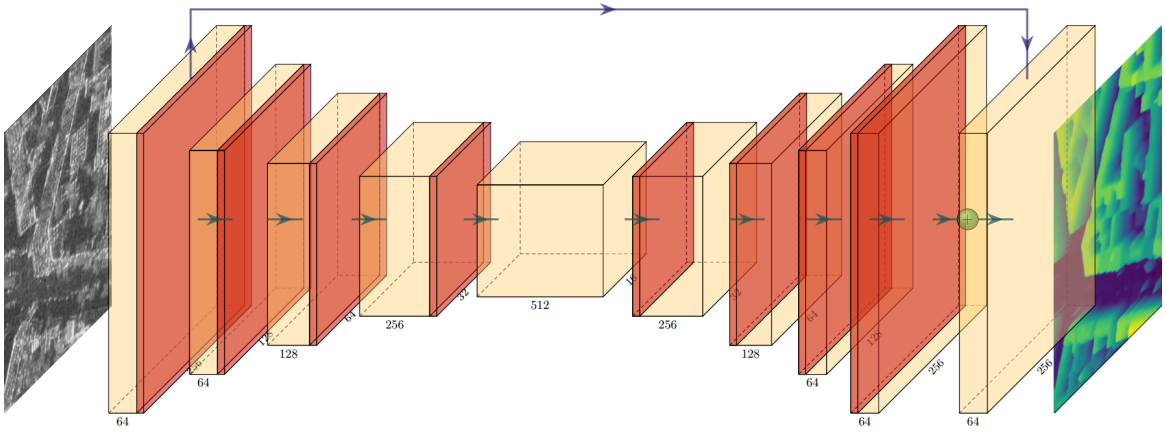


Figure 2. CNN architecture based on the IM2HEIGHT model (Mou and Zhu, 2018). In the left half of the network, the incoming image data is successively transformed into increasingly abstract representations using learned convolution kernels. In the right half, the center-resulting "representation cube" is again unfolded into an image-based representation, with the desired result in this case being a height image. The connecting arrow shown above between the first and the last layer of the network represents a so-called skip connection, which allows the spatial structures still present in the image at full resolution to be transferred to the resulting image.

Table 2. Quantitative evaluation results calculated for predictions on the hold-out test sets.

| **Data set** | **RMSE [m]** | **RMSE$_{\log}$** | **Rel** | **Rel$_{\log}$** | $\delta_1$ [%] | $\delta_2$ [%] | $\delta_3$ [%] | **SSIM** |
|---|---|---|---|---|---|---|---|---|
| | lower is better | | lower is better | | | higher is better | | high. bett. |
| SAR | **5.30** $\pm$ 1.23 | **0.23** $\pm$ 0.04 | **0.42** $\pm$ 0.13 | **0.16** $\pm$ 0.03 | **44.97** $\pm$ 10.21 | **66.38** $\pm$ 8.09 | **77.81** $\pm$ 6.41 | 0.78 $\pm$ 0.07 |
| Optical | 5.50 $\pm$ 1.68 | 0.28 $\pm$ 0.06 | 0.50 $\pm$ 0.14 | 0.20 $\pm$ 0.06 | 35.96 $\pm$ 10.94 | 53.59 $\pm$ 10.98 | 64.87 $\pm$ 11.09 | **0.82** $\pm$ 0.11 |

### 3.1 Experimental Setup

In both cases (SAR and optical), the model was trained on a large number of specially prepared training data. For that, an ADAM optimizer with the default learning rate of 0.001 was used. The mean square error (MSE), also referred to as $\mathcal{L}_2$ loss, was chosen as loss function. Just as described in the original publications, each mini-batch consists of only a single image. Because of the use of ReLUs as activation functions, all weights are initialized with a Kaiming uniform distribution. To mitigate overfitting effects, 15% of the training data is randomly declared as a validation set for an early stopping mechanism.

To make the network more robust to different viewing angles, random flipping was included in the data loader process. Thus, for each draw from the data pool, the image is randomly mirrored at one of the main axes before being fed to the network.

All in all, the training set consisted of 1,671 annotated SAR images and 2,775 annotated optical images, respectively, each with a size of $256 \times 256$ pixels and a ground resolution of 1 m and 0.7 m per pixel, respectively. During training, the underlying neural network was provided with these sample images in multiple runs to learn how to convert the respective input image data into an elevation image by mathematically optimizing the model parameters.

### 3.2 Results

The two final models were then applied to the images from the test subsets marked by the white rectangles shown in Fig. 1. Since those subsets were unseen to the model during the training, a fair assessment is ensured. Results of the quantitative evaluation are summarized in Tab. 2, relying on metrics commonly used in the SIDE literature (Eigen et al., 2014, Amirkolaee and Arefi, 2019). Whereas the root mean square error (RMSE) is a well-established standard for providing an idea about average error in a test set, its logarithmic variant, RMSE$_{log}$ is more robust against outliers and independent of error scale. Both the relative error Rel and its logarithmic variant Rel$_{log}$ measure the average error relative to the size of the measured quantity. Finally, the delta measure $\delta_i$ calculates a ratio for each pair of pixels and then counts the percentage of how many of the pixel values are below a certain threshold. While all those metrics are intended to measure the quantitative error of reconstructed heights per pixel, it is also interesting to investigate the visual reconstruction quality of the resulting height images. For this, we use the structural similarity index measure (SSIM), which is a metric to evaluate the structural similarity of a predicted image with respect to a reference image (Wang et al., 2004).

From the numbers in Tab. 2 it can be seen that overall the SAR-based reconstruction performs better in all metrics regarding the quality of reconstructed heights, whereas the reconstruction achieved from optical data performs better in terms of visual structure.

A more detailed view on the results is provided in Fig. 3. It shows two randomly selected example patches represented by both SAR and optical test images. For each test image additionally the target height maps and the predicted height maps are shown. Finally, the pixel-wise error distributions are shown in the form of histograms. These examples illustrate two things: First, the structurally somewhat clearer reconstruction in the optical case becomes visible. Second, the difference in error distributions becomes apparent, with mean differences between predicted and reference heights of 1.97 m and -4.25 m in the optical case, and 0.38 m and -1.32 m in the SAR case, respectively. This seems to confirm the overall results (cf. Tab. 2) indicating that single-image height reconstruction works slightly better for SAR input data than for optical input data.

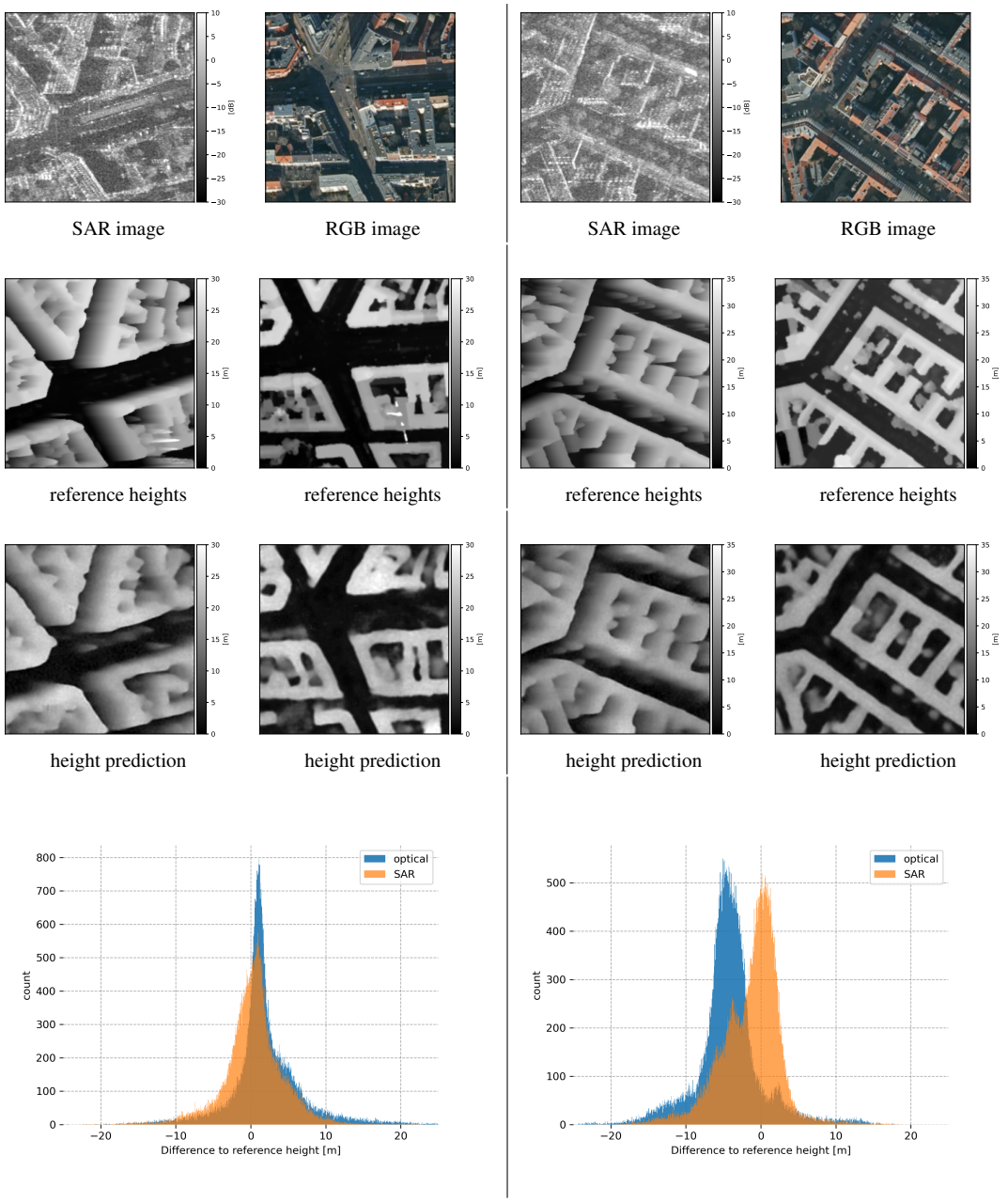


Figure 3. Two randomly selected example images from the test sets, depicted for both the SAR case (columns 1 and 3), and the optical case (columns 2 and 4). The bottom subfigure displays the corresponding error distributions.

## 4. DISCUSSION

### 4.1 Accuracies Reported in the Literature

In order to embed the experimental results presented in this paper into a bigger picture, quantitative results achieved in other works on single-image height reconstruction from remote sensing imagery are summarized in Tab. 3[1]. While they are, of course, not directly comparable due to significant differences in training and test sets, they can still provide a rule-of-thumb understanding of the order of magnitude of single-image height reconstruction qualities. From this summary, several insights can be drawn:

- All single-image height reconstruction results reported in the literature so far achieve RMSE values in the range of a few meters. This indicates a certain saturation in terms of achieveable accuracy.
- Besides only a few exceptions, all results are achieved with models trained and tested on VHR aerial photographs – mostly provided in the form of the ISPRS Postdam/Vaihingen datasets (ground sampling distance: 0.05 m for Potsdam and 0.09 m for Vaihingen).
- Generally, results on VHR aerial imagery (RMSEs from 1.40 m to 3.89 m) are better than results achieved on satellite imagery (RMSEs from 5.02 m to 6.45 m). This can

[1] Please note that we do not show the results reported in (Mou and Zhu, 2018) and (Pellegrin and Martinez-Carranza, 2020), as they differ from all other results by at least one order of magnitude, which seems to be caused by an erroneous calculation of (R)MSE values.

Table 3. Reconstruction quality metrics achieved in related work. For sake of comparability, only the best RMSE value achieved by training and testing on the same scene is reported for each work. Results of transferability were not considered where available at all.

| Paper | Data basis | RMSE |
|---|---|---|
| (Ghamisi and Yokoya, 2018) | Optical data (ISPRS Potsdam dataset) | 3.89 |
| | Optical data (ISPRS Vaihingen dataset) | 2.58 |
| (Amirkolaee and Arefi, 2019) | Optical data (ISPRS Vaihingen dataset) | 2.87 |
| | Optical data (ISPRS Potsdam dataset) | 3.47 |
| | Optical data (own satellite-based dataset) | 6.45 |
| (Mahmud et al., 2020) | Optical data (DFC 2019 dataset) | 5.02 |
| | Optical data (Urban 3D dataset) | 6.15 |
| | Optical data (ISPRS Potsdam dataset) | 3.73 |
| | Optical data (ISPRS Vaihingen dataset) | 1.93 |
| (Li et al., 2020) | Optical data (ISPRS Vaihingen dataset) | 1.70 |
| | Optical data (ISPRS Potsdam dataset) | 1.40 |
| (Recla and Schmitt, 2022) | SAR data (own satellite-based dataset) | 5.25 |

at least partially be explained with the lower spatial resolution of the satellite imagery (ground sampling distance around 1 m).

This literature review confirms the validity of the results reported in this paper, which fall in the same accuracy category as the results reported for optical satellite imagery.

### 4.2 Reasons for Better Performance of SAR-based Height Reconstruction

As the results presented in Tab. 2 and Fig. 3 indicate, SAR-based single-image height reconstruction seems to perform slightly better than single-image height reconstruction using optical imagery as input. Also, the literature review collected in Tab. 3 does not seem to contradict that perception. In spite of the fact that the results are not directly comparable, it can be seen that the reported RMSE values for experiments carried out on spaceborne optical imagery are 5.02 m (DFC 2019 dataset), 6.15 m (Urban 3D dataset), and 6.45 m (unpublished dataset), whereas for spaceborne SAR data 5.25 m are reported (Recla and Schmitt, 2022).

There are two hypotheses that can possibly explain this behaviour:

1. In case of optical data, the annotation is usually carried out be providing orthorectified imagery with a co-registered, georeferenced height map, which can be considered a valid approximation given the nadir-like acquisition geometry of most optical remote sensing imagery. However, as has been well-known to the photogrammetry community, this is not fully correct from a geometric point of view (Amhar et al., 1998), especially if VHR images with non-nadir viewing angles are used. Thus, in a recent study from the field of computer vision, single-image height estimation was paired with a variant of optical flow for static images in order to predict more accurate height maps (Christie et al., 2020). The only available study on single-image height prediction from SAR imagery, however, made use of a sophisticated annotation workflow that projected every height value to its correct pixel counterpart in the original SAR image geometry (Recla and Schmitt, 2022). This might lead to slightly larger height prediction errors for the optical imagery compared to the SAR imagery – especially for facade regions.
2. Due to the system-inherent side-looking viewing geometry of SAR sensors, SAR images contain more information about building facades than about roof areas. Together with the complementary phenomenon of radar shadowing, this leads to very strong height cues, which can be exploited beneficially by the convolutional neural network for height prediction. In the end, there have already been works on building height estimation from single SAR images using domain expertise about the imaging geometry in the pre-deep learning era (Wegner and Soergel, 2008). In contrast, optical images in nadir geometry more or less only contain ground and roof areas and thus lack significant height cues. Height predictions will therefore mainly be supported by context information and can more easily suffer from ambiguities. A conceptual comparison of SAR vs. optical imaging is sketched in Fig. 4.

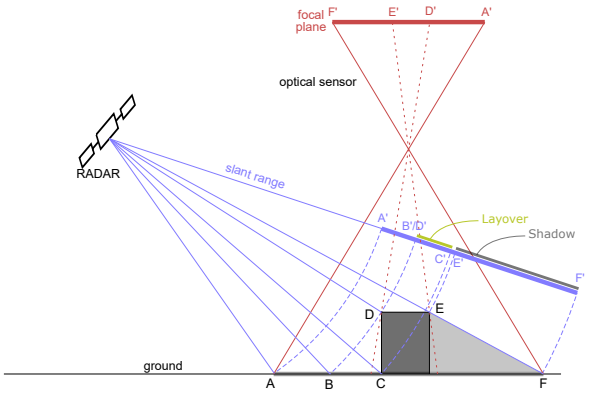


Figure 4. Schematic comparison between the SAR and optical imaging principles. While an optical system measures angles, a SAR converts signal travel times to distances and maps them into its resolution cells – and is inherently side-looking.

In summary, there are some hints towards the notion that SAR is the favorable sensing modality when it comes to single-image height reconstruction in remote sensing. However, future work will have to consider the following points:

- A more sophisticated data annotation needs to be used in the optical case, i.e. similar to the SAR case, a direct 3D to 2D relationship has to be employed rather than a mere overlay of ortho-imagery and 2.5D height data.

- Besides (near-)nadir optical imagery, also oblique-view optical imagery should be used in the experiments to have a really fair comparison.
- Evaluations should not be carried out anymore in the patch-relative height space, but in actual 3D world coordinates.

Only if those issues are properly addressed will we be able to get a really fair comparison of single-image height prediction from SAR and optical sensors. This paper can be seen as a step in that direction.

## 5. SUMMARY & CONCLUSION

In this paper, we have used a simple state-of-the-art convolutional neural network architecture to train models for the prediction of urban height maps from single optical and SAR images. Using training and test data from the city of Berlin, Germany, we compared the reconstruction results regarding the achieved accuracies. In conjunction with an analysis of evaluation metrics reported in related literature, we conclude that SAR-based single-image height reconstruction might perform slightly better than single-image height reconstruction from optical imagery. We hypothesize that besides pre-processing issues, the difference between side-looking and nadir-looking imaging geometries is the main reason for this difference: Since SAR systems observe facade and shadow information, which provides a lot of valuable height cues, optical systems are more prone to ambiguous height predictions, because they observe mostly ground and roof information.

## ACKNOWLEDGMENTS

This work is supported by the German Research Foundation (DFG) as grant SCHM 3322/3-1. The SAR imagery was provided by the German Aerospace Center (DLR) in the frame of the proposal MTH3753.

## REFERENCES

Amhar, F., Jansa, J., Ries, C. et al., 1998. The generation of true orthophotos using a 3D building model in conjunction with a conventional DTM. *International Archives of Photogrammetry and Remote Sensing*, 32/4, 16-22.

Amirkolaee, H. A., Arefi, H., 2019. Height estimation from single aerial images using a deep convolutional encoder-decoder network. *ISPRS Journal of Photogrammetry and Remote Sensing*, 149, 50-66.

Christie, G., Abujder, R. R. R. M., Foster, K., Hagstrom, S., Hager, G. D., Brown, M. Z., 2020. Learning geocentric object pose in oblique monocular images. *Proc. CVPR*, 14512–14520.

Cooper, A. P. R., 1994. A simple shape-from-shading algorithm applied to images of ice-covered terrain. *IEEE Transactions on Geoscience and Remote Sensing*, 32(6), 1196-1198.

Di Martino, G., Di Simone, A., Iodice, A., Riccio, D., Ruello, G., 2014. On shape from shading and SAR images: An overview and a new perspective. *Proc. IGARSS*, 1333–1336.

Eigen, D., Puhrsch, C., Fergus, R., 2014. Depth map prediction from a single image using a multi-scale deep network. *Proc. NeurIPS*, MIT Press, Cambridge, MA, USA, 2366–2374.

Ghamisi, P., Yokoya, N., 2018. IMG2DSM: Height Simulation From Single Imagery Using Conditional Generative Adversarial Net. *IEEE Geoscience and Remote Sensing Letters*, 15(5), 794-798.

Heipke, C., 1992. Integration of digital image matching and multi image shape from shading. *Mustererkennung*, 186–198.

Li, X., Wang, M., Fang, Y., 2020. Height Estimation From Single Aerial Images Using a Deep Ordinal Regression Network. *IEEE Geoscience and Remote Sensing Letters*. in press.

Mahmud, J., Price, T., Bapat, A., Frahm, J.-M., 2020. Boundary-aware 3d building reconstruction from a single overhead image. *Proc. CVPR*, 441–451.

Mertan, A., Duff, D. J., Unal, G., 2021. Single image depth estimation: An overview. arXiv:2104.06456.

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

Natsuaki, R., Hirose, A., 2012. InSAR local co-registration method assisted by shape-from-shading. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 6(2), 953-959.

O'Hara, R., Barnes, D., 2012. A new shape from shading technique with application to Mars Express HRSC images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 67, 27-34.

Pellegrin, L., Martinez-Carranza, J., 2020. Towards depth estimation in a single aerial image. *International Journal of Remote Sensing*, 41(5), 1970-1985.

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.

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.

Wegner, J. D., Soergel, U., 2008. Bridge height estimation from combined high-resolution optical and SAR imagery. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 37(B7-3), 1071–1076.

Zhang, R., Tsai, P.-S., Cryer, J. E., Shah, M., 1999. Shape-from-shading: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(8), 690-706.