ENHANCING BUILDING SHAPE DETAILS THROUGH DEEP LEARNING IN SINGLE-IMAGE SAR-BASED DSM

Ksenia Bittner[1], Michael Recla[2], Stefan Auer[1], Michael Schmitt[2]

[1]Remote Sensing Technology Institute, German Aerospace Center (DLR),
82234 Weßling, Germany
[2]University of the Bundeswehr Munich, Department of Aerospace Engineering,
85579 Neubiberg, Germany

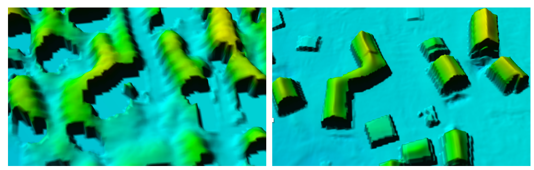
ABSTRACT

Due to the reliability of data acquisition, synthetic aperture radar (SAR) sensors are fundamental for remote sensing applications with the need for flexibility and fast response. For urban applications, besides the analysis of salient point signatures, extracted height information allows to evaluate the state of buildings. Recently developed deep learning approaches enable height estimates in situations where only one SAR image of an area of interest is available. However, building shapes still exhibit low quality in the resulting digital surface models (DSMs). This paper presents how derived surface models from the SAR image can be refined with knowledge about the shape of buildings. For that purpose, building representations are learned with a neural network from optical images and CityGML models. The results demonstrate that our model not only effectively transfers knowledge to process DSMs from various data sources but also showcases the ability to generalize across different regions.

Index Terms**—** Synthetic Aperture Radar, Digital Surface Models, Machine Learning, Artifical Intelligence, Data Fusion, Urban Applications

1. INTRODUCTION

The acquisition of accurate and detailed digital surface models (DSMs) is crucial for a wide variety of remote sensing applications, ranging from urban planning and disaster response to environmental simulation. While conventional methods such as light detection and ranging (LiDAR) measurements provide the highest-quality DSMs, their production is costly, time-consuming, and limited in coverage. In contrast, satellite-based systems offer many advantages, especially in terms of cost efficiency, coverage, and speed. With their all-weather capability, synthetic aperture radar (SAR) systems in particular offer a unique opportunity to obtain an initial overview of the situation in the event of disasters such as earthquakes, floods, or military conflicts. New developments in the field of deep learning-based single image height estimation methods now make it possible to derive height maps of urban areas from single SAR images [1]. This is a capability that used to require around 20 images of the same scene using methods such as TomoSAR. The new more data-efficient approaches offer enormous potential for rapid mapping applications. However, the quality of the resulting DSMs is



(a) Initial SAR DSM (b) Refined DSM

Fig. 1: A sample area demonstrating the inference results obtained from the proposed network for the building shape refinement task.

limited. This is due to multiple reflections and signal mixtures in complex urban areas, reprojection errors from the radar geometry into the map projection, and the complexity of the task itself, which remains an estimation and not a computation. To improve the quality of these surface models, we present a smart, learned filter that corrects for the noisy building shapes in SAR derived DSMs as illustrated in Fig. 1.

2. DEEP LEARNING FOR SINGLE-IMAGE DSM REFINEMENT

This paper addresses the deep-learning-based refinement of noisy digital surface models that were produced by single-image height estimation from SAR intensity imagery. In the following sections, we first describe the single-image height reconstruction procedure, then the DSM refinement method.

2.1. Infering height from SAR image

The digital surface models to be refined in section 2.2 are generated from individual SAR intensity images using the method originally described in [1] (SAR2Height model). This involves a previously trained model reconstructing a height value above ground to each pixel of an input image, for example, the building height in the case of a roof pixel. To solve this regression problem, a modified form of the popular U-Net [2] is used. The height estimation process is performed in the sensor-specific imaging geometry, which is slant range for SAR. This means that the training data, namely high-resolution LiDAR data of the training areas, have to be projected into the azimuth-range geometry of the corresponding SAR images. Knowing the exact orbits of the SAR data, this can be accomplished for each pixel by using the Zero-Doppler geometry. In the case of layover, it is always the height of the highest target that is mapped, as we are eventually aiming for a surface model. Details of this procedure can be found in [1]. In order to obtain a height map in an orthometric map projection, the relative model outputs are geocoded with the aid of a terrain model. The terrain model provides the

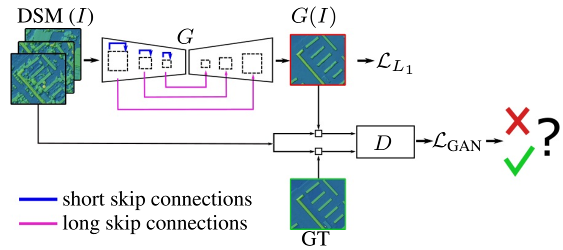


Fig. 2: Schematic overview of the proposed architecture.

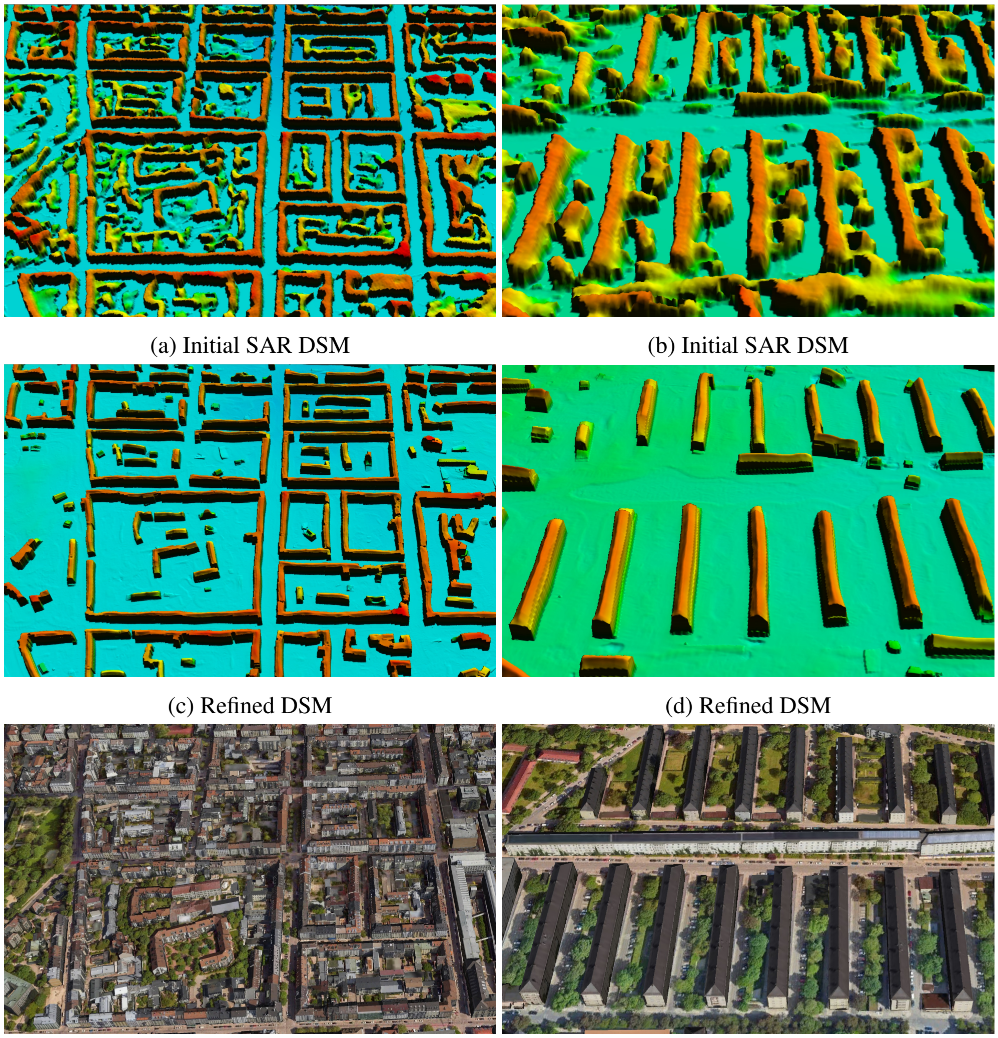
absolute height reference required to localize a pixel (row, col) in world coordinates (X, Y, Z). For this purpose, a system of non-linear equations is solved for each pixel and thus a point cloud is created, which is then rasterized to obtain the desired DSM. The terrain model should be globally available and free of all elevated objects like buildings and vegetation, such as the WorldDEM™ DTMlite from Airbus or the FABDEM [3]. Refer to [4] for further information on this projection algorithm.

In the projection step from slant range to ground range, however, reprojection errors are introduced into the resulting DSM. On the one hand, the used terrain model is subject to a certain degree of inaccuracy, which directly affects the localization of the pixels, and on the other hand, the relative heights themselves estimated by the model are not perfect, as we are dealing with an ill-posed problem and a difficult task. Especially blurry predicted edges of buildings lead to detached scattered points in the point cloud. These clearly erroneous points are filtered out of the point cloud using a rule-based method employing a density metric. However, unsightly edges still remain.

2.2. Building shape refinement

In this work, we follow the strategy introduced by Bittner *et al.* [5]. Mainly, we train a UResNet50 model on photogrammetric DSM as input and LoD2-like DSM as output within a conditional generative adversarial network (cGAN) framework, and later perform inference on DSMs generated from a single SAR images.

The reason we are maintaining our choice for UResNet50 is that in the work of Bittner *et al.* [6]



(a) Initial SAR DSM

(b) Initial SAR DSM

(c) Refined DSM

(d) Refined DSM

(e) Google Earth View

(f) Google Earth View

Fig. 3: Visual analysis of inferred DSMs processed by UResNet50 out of SAR-DSMs over selected urban areas. The DSM images are color-shaded for better visualization.

it has demonstrated superior results in comparison to other models. The explored hybrid architecture, termed ResNet-in-UNet, integrates a residual network (ResNet) as an encoder and utilizes five up-sampled decoder layers with five long skip connections, forming a U-shaped structure. In such a structure, the network integrates identity shortcuts, also referred to as residual connections, along with extensive skip connections within the generative part G of the network (see Fig. 2). Residual connections excel in mitigating the vanishing gradient problem, enabling the training of deeper networks and consequently enhancing the network's semantic comprehension of the scene. Long skip connections demonstrate their prowess by effectively transmitting intricate information from lower to upper layers, thereby enhancing the decoder's up-sampling accuracy that might diminish during hierarchical down-sampling in the encoder. However, unlike the previous work of Bittner *et al.* [6], we omitted the panchromatic image. This study focuses on assessing the network's generalization capabilities and aims to remain independent of additional data that might not be easily accessible.

To give a sharper appearance to the resulting building shapes, the adversarial manner of learning was employed. Mainly, a second network, referred to as a discriminator D with five convolutional layers and a sigmoid activation function placed on the top layer, is involved. Consequently, the discriminator's output indicates the likelihood that the input sample resembles either the authentic data distribution or the fabricated one.

3. EXPERIMENTAL SETUP

The SAR2Height model for generating DSMs from SAR images was trained on High Resolution and Staring SpotLight scenes acquired by the German TerraSAR-X satellite. For this purpose, acquisitions of eight different cities were collected from different orbits and with varying looking angles. The city on which the final methods are tested was excluded from the training.

The UResNet50 model for DSMs refinement was trained on DSM showing $353\,\mathrm{km}^2$ of Berlin city, Germany, which was generated from six PAN images acquired by the WorldView-1 satellite on two different days, following the workflow of d'Angelo *et al.* [7]. Our test image is a DSM from a single SAR image showing the region of Munich city, Germany. By performing a poor inference on SAR-DSM, we demonstrate the power of the developed methodology for the refinement task to generalize not only to a new scene but also to different, more challenging input data.

4. RESULTS AND DISCUSSION

Figure 3 illustrates two SAR-DSM samples over an area in Munich together with obtained refined DSMs. One can notice that despite the original blobby appearance of buildings on the initial SAR-DSM, our model was able to recognize the correct shape of rooftops (see Fig. 3e and Fig. 3f for visual comparison). The reason why our model was able to distinguish appropriate rooftops, even for roof types like gable or hip, is that the initial building appearances for those particular cases on SAR-DSM still feature prolonged shapes in the direction of rooftop ridgeline similar to building appearances in photogrammetric DSM. Additionally, as our model learned to identify the absence of vegetation in the ground truth LoD2-like DSM, it successfully removed this vegetation from the SAR-DSM as well. It's important to note that the model accurately distinguishes buildings from other objects, refining only the buildings themselves without generating any additional structures. The main limitation of our method is that it still struggles to reconstruct all buildings in a scene due to differences in photogrammetric DSMs and SAR-DSMs, as well as errors in SAR-DSM generation.

5. SUMMARY & CONCLUSION

With this paper, we introduce an approach to enhance building shape details in digital surface models derived from single synthetic aperture radar (SAR) images. The deep learning-based single image methods for height estimation offer unprecedented flexibility and data efficiency, but the resulting DSMs are not on par with the quality of light detection and ranging (LiDAR)-generated height maps. To improve the building shapes, a UResNet model, which was trained within a conditional generative adversarial network (cGAN) framework on photogrammetric digital surface models (DSMs) and level of detail (LoD)2 CityGML building models, was applied to the SAR-based surface models. Not only were the building outlines and shapes significantly improved, but vegetation was also reliably filtered out, as this does not occur in the LoD2 training data. This demonstrates an impressive transfer performance of the presented network, which has never been trained on SAR-based data. The introduced method showcases the great potential of using smart filters to improve noisy elevation data from urban environments.

References

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

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. of Medical Image Computing and Computer-Assisted Intervention 2015*, 2015, pp. 234–241.

[3] L. Hawker *et al.*, "A 30 m global map of elevation with forests and buildings removed," *Environmental Research Letters*, vol. 17, no. 2, 2022, Art. no. 024016.

[4] M. Recla and M. Schmitt, "From relative to absolute heights in SAR-based single-image height prediction," in *Proc. of the Joint Urban Remote Sensing Event (JURSE)*, 2023.

[5] K. Bittner, P. d'Angelo, M. Körner, and P. Reinartz, "Dsm-to-lod2: Spaceborne stereo digital surface model refinement," *Remote Sensing*, vol. 10, no. 12, p. 1926, 2018.

[6] K. Bittner, L. Liebel, M. Körner, and P. Reinartz, "Long-short skip connections in deep neural networks for dsm refinement," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 43, no. B2, pp. 383–390, 2020.

[7] P. d'Angelo and P. Reinartz, "Semiglobal matching results on the isprs stereo matching benchmark," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 38, pp. 79–84, 2012.