# Automatic Large-Scale 3D Building Shape Refinement Using Conditional Generative Adversarial Networks

Ksenia Bittner
German Aerospace Center - DLR
Munich, Germany
ksenia.bittner@dlr.de

Marco Körner
Technical University of Munich
Munich, Germany
marco.koerner@tum.de

Three-dimensional realistic representations of buildings in urban environments have been increasingly applied as data sources in a growing number of remote sensing fields such as urban planning and city management, navigation, environmental simulation (*i.e*. flood, earthquake, air pollution), 3D change detection after events like natural disasters or conflicts, *etc*. With recent technological developments, it becomes possible to acquire high-quality 3D input data. There are two main ways to obtain elevation information: from active remote sensing systems, such as *light detection and ranging (LIDAR)*, and from passive remote sensing systems, such as optical images, which allow the acquisition of stereo images for automatic *digital surface models (DSMs)* generation. Although airborne laser scanning provides very accurate DSMs, it is a costly method. On the other hand, the DSMs from stereo satellite imagery show a large coverage and lower costs. However, they are not as accurate as LIDAR DSMs. With respect to automatic 3D information extraction, the availability of accurate and detailed DSMs is a crucial issue for automatic 3D building model reconstruction. We present a novel methodology for generating a better-quality stereo DSM with refined buildings shapes using a deep learning framework. To this end, a *conditional generative adversarial network (cGAN)* is trained to generate accurate LIDAR DSM-like height images from noisy stereo DSMs.

Over the past two decades, the need of accurate DSMs for 3D building modeling from remote sensing imagery increases research effort to develop the methodologies for automatic elevation model enhancement. For example, Maire [1], first, extract from high-resolution satellite imagery the user-defined semantic contents like sea, lakes, buildings or roads with a supervised classification. Then for each detected segment a plane is interpolated with geometric constraints given by the topological properties of each class and neighbor regions. Krauß *et al.* [2] segment and transfer one stereo image to the disparity map, then for each segment the original disparity map is filled with suitable interpolation of the disparities to recover the occlusion errors. Poli *et al.* [3] propose to use segmentation of a *very high-resolution (VHR)* satellite imagery to refine a given DSM at a coarser resolution. Mainly, the image scene is segmented with alpha-omega connectivity [4] and overlaid on the given DSM. Then the statistics like mean, median, standard deviation, maximum and minimum values of the heights of the points falling into each segment are calculated. The new surface model is determined afterward from the existing one by enforcing that the height values belonging to the same segment follow a certain mathematical function, *i.e*. constant value or planar surface. Although the previous methodologies show the promising results, automatic enhancement of buildings shapes in DSMs is still an open research problem.

The recent developments in artificial neural networks provide the best solutions to problems in various fields like computer vision, medicine, biology, and remote sensing. The introduction of *generative adversarial networks (GANs)* attracted a lot of attention in the field of machine learning as they offer a new possibility to generate high-quality images across a wide range of domains [5]. Recently, several works made an attempt to go from the 2D domain to 3D and generate 3D object shapes. Wu *et al.* [6] propose a generative model of 3D shapes from a probabilistic space by using volumetric convolutional networks and GANs. Yan *et al.* [7] generate 3D shape from a single 2D image using multiple projections from 3D shape from a known viewpoint. In opposition to the common GAN setup, we want to generate an artificial LIDAR DSM-like height image similar to some known input image (a stereo DSM in our case). For this purpose, we utilize the cGANs approach proposed by Isola *et al.* [8]. The cGANs consist of a generative model $G$ and a discriminative model $D$ which compete against each other. Training a cGANs is equivalent to a min-max game

$$G^{\star} = \arg\min_{G}\max_{D} \mathcal{L}_{cGAN}(G,D) + \lambda\mathcal{L}_{L1}(G) \qquad (1)$$

between the generator and the discriminator, where $G$ in-

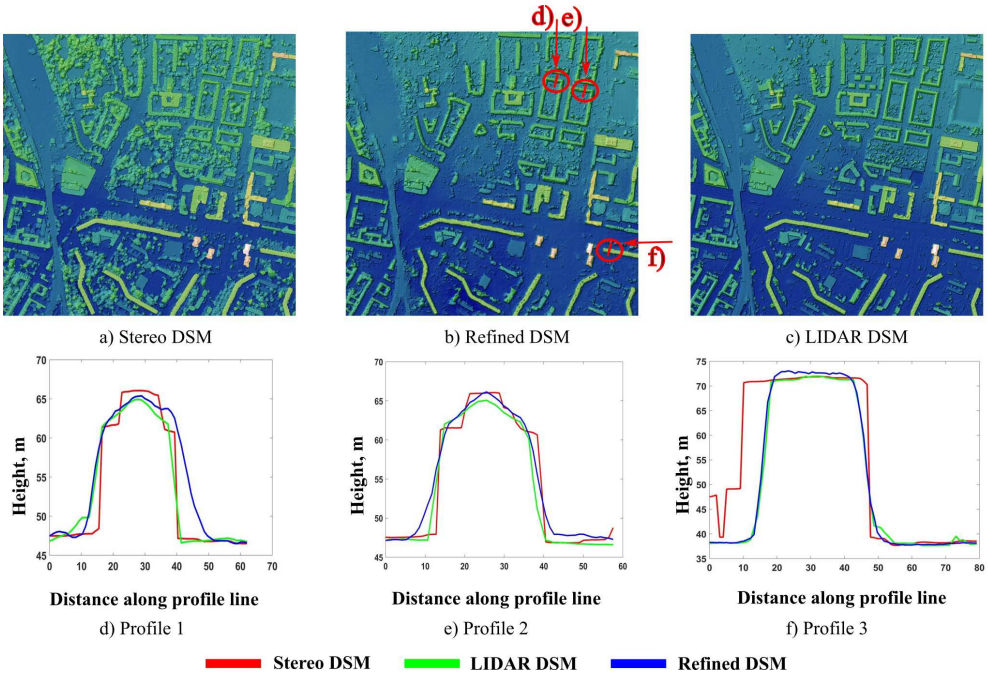
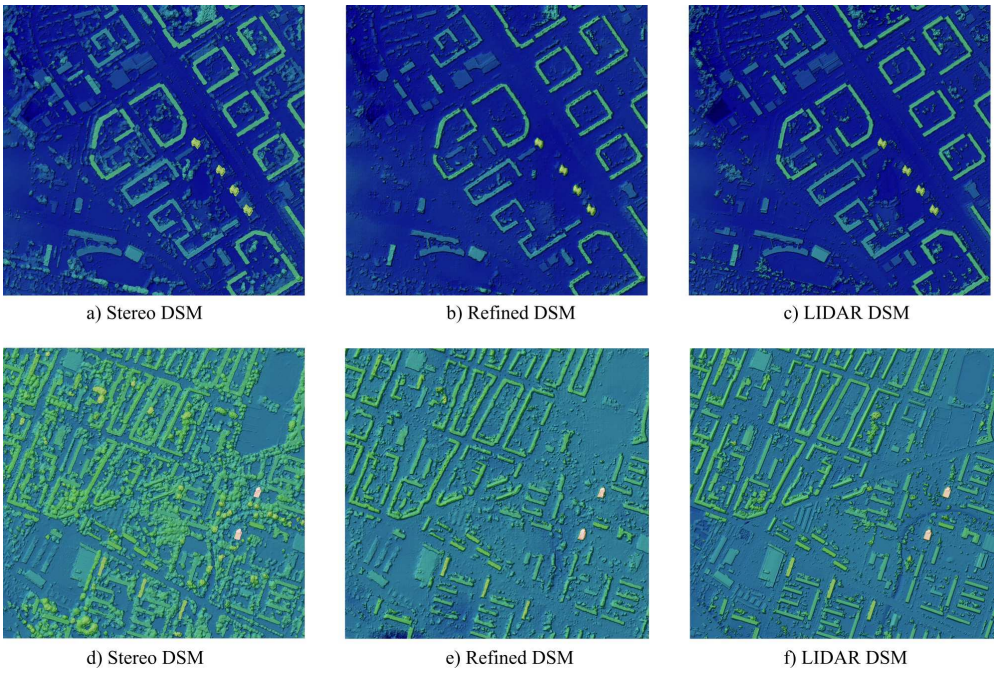



Figure 1. Example of the generated DSM with refined 3D buildings shapes and profiles of selected buildings.

tents to minimize the objective $\mathcal{L}_{cGAN}(G, D)$ against the $D$ that aims to maximize it. The second term in Equation 1 assures that the generator produces the output near the ground truth in a $L_1$ sense.

The architecture of the cGANs is organized as follows: The generator $G$ is represented by an U-Net architecture [9], an encoder-decoder type of network, which combines the encoder feature maps with up-sampled feature maps from the decoder at every stage by skip connections. The skip connections allow the decoder to learn back relevant features that are lost due to the pooling in the encoder. The discriminator $D$ is realized via several convolutional layers with a *sigmoid* activation function as the last layer. Since the remote sensing images are huge, for training and testing we tile the images into patches with a size of 256×256 pixels which fits into the available GPU memory. In order to avoid artifacts and object discontinuities at tile boundaries, the patches are generated with overlap. During optimization, the $G$ and $D$ networks are trained at the same time by alternating their trainings. The generator $G$ tries to synthesize realistic images to fool the discriminator $D$, and $D$ in its turn tries to tell which samples are real or generated. In the inference step, the predictions are performed for each patch separately from the test dataset to generate a predicted map of the same size as the patch. After that, the tiles are stitched together in order to generate an image with the same size as the original test area.

Experiments have been performed on data consisting of stereo DSM over Berlin city, Germany, derived from WorldView-2 very high-resolution stereo panchromatic imagery with a resolution of 0.5 m. A sample of the input image is illustrated in Figure 1(a). As ground truth, a LIDAR DSM from the Senate Department for Urban Development and Housing, Berlin, was used for learning the mapping function between the noisy DSM and the one with better quality. The test sample is illustrated in Figure 1(c). As the LIDAR DSM was generated from last pulse data, there is no or much less vegetation within a scene in comparison to the stereo DSM. Figure 1(b) shows the results generated by cGAN and depicts the elevation model of the same resolution as the input image. It can be clearly seen that geometric structures of buildings from stereo DSM are preserved in the generated sample and closer to the LIDAR DSM. Besides, the network has learned about the much smaller amount of vegetation from these data. More examples of generated images are illustrated in Figure 2. By investigating the profiles (see Figures 1(d)-(f)) of the selected buildings, highlighted by the red lines in Figure 1(b), we can confirm that the cGAN successfully learned the 3D buildings representations close to the LIDAR data representations. Regarding to the ridge lines of the buildings from the first two profiles we can see that they are much sharper in comparison to ridge lines from stereo DSM and are at the center of the roof which gives more realistic view and is geometrically correct. The profile in Figure 1(f) also shows very close resemblance of resulted building shape to the ground truth, especially regarding the width and borders of the building, although, the input 3D shape is much wider



a) Stereo DSM b) Refined DSM c) LIDAR DSM

d) Stereo DSM e) Refined DSM f) LIDAR DSM

Figure 2. Example of the generated DSM with refined 3D buildings shapes.

due to the trees in the neighborhood. Thus, it is evident that the trees have much less influence on the building shapes. More results on quantitative evaluation experiment can be found in original paper published in ISPRS Archives.

In the future, we will augment the generative part of cGAN network with additional branch which learns the information from spectral image: In our case it is 1 channel panchromatic image. We believe, the generative network will leverage mutual information of the spectral image and the height image to enhance the contours of building together with the ridge line for some roof types as this sort of information is more visible and accurate on spectral images. Moreover, we will experiment with total variation loss and 8-connected gradient loss used together with data loss to train generative networks for producing images even more closer to the real ones.

## References


[1] C Maire, "Image information extraction and modeling for the enhancement of digital elevation models," PhD thesis, Phd Thesis, Karlsruhe Institute of Technology (KIT), 09.02, 2010.

[2] T. Krauß and P. Reinartz, "Enhancment of dense urban digital surface models from vhr optical satellite stereo data by pre-segmentation and object detection," 2010.

[3] D. Poli and P. Soille, "Digital surface model extraction and re nement through image segmentation–application to the isprs benchmark stereo dataset," *Photogrammetrie-Fernerkundung-Geoinformation*, vol. 2012, no. 4, pp. 317–329, 2012.

[4] P. Soille, "Constrained connectivity for hierarchical image partitioning and simplification," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 7, pp. 1132–1145, 2008.

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[6] J. Wu, C. Zhang, T. Xue, B. Freeman, and J. Tenenbaum, "Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling," in *Advances in Neural Information Processing Systems*, 2016, pp. 82–90.

[7] X Yan, J. Yang, E Yumer, Y Guo, and H Lee, "Learning volumetric 3d object reconstruction from single-view with projective transformations," in *Neural Information Processing Systems*, 2016.

[8] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *arXiv preprint arXiv:1611.07004*, 2016.

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2015, pp. 234–241.