

AUTOMATIC LARGE-SCALE 3D BUILDING SHAPE REFINEMENT USING CONDITIONAL GENERATIVE ADVERSARIAL NETWORKS

K. Bittner^a, P. d'Angelo^a, M. Körner^b, P. Reinartz^a

^a Remote Sensing Technology Institute, German Aerospace Center (DLR), Wessling, Germany -
(ksenia.bittner, pablo.angelo, peter.reinartz)@dlr.de

^b Technical University of Munich, Munich, Germany - marco.koerner@tum.de

Commission II, WG II/6

KEY WORDS: Conditional generative adversarial networks (cGANs), Digital Surface Model, 3D scene refinement, 3D building shape.

ABSTRACT:

Three-dimensional building reconstruction from remote sensing imagery is one of the most difficult and important 3D modeling problems for complex urban environments. The main data sources provided the digital representation of the Earth's surface and related natural, cultural, and man-made objects of the urban areas in remote sensing are the *digital surface models (DSMs)*. The DSMs can be obtained either by *light detection and ranging (LIDAR)*, SAR interferometry or from stereo images. Our approach relies on automatic global 3D building shape refinement from stereo DSMs using deep learning techniques. This refinement is necessary as the DSMs, which are extracted from image matching point clouds, suffer from occlusions, outliers, and noise. Though most previous works have shown promising results for building modeling, this topic remains an open research area. We present a new methodology which not only generates images with continuous values representing the elevation models but, at the same time, enhances the 3D object shapes, buildings in our case. Mainly, we train a *conditional generative adversarial network (cGAN)* to generate accurate LIDAR-like DSM height images from the noisy stereo DSM input. The obtained results demonstrate the strong potential of creating large areas remote sensing depth images where the buildings exhibit better-quality shapes and roof forms.

1. INTRODUCTION

Three-dimensional realistic representations of buildings in an urban environment have been increasingly applied as a data source in a growing number of remote sensing applications, 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 mainly two main ways to obtain very high-resolution 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 the automatic generation of DSMs. 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, *i.e.*, the 3D building shapes do not feature steep walls and detailed rooftop representations. This is a major problem for stereo DSMs due to occlusions by dense and complex building structures or stereo matching errors during DSMs generation. Therefore, with respect to automatic 3D information extraction, the availability of accurate and detailed DSMs is a crucial issue for automatic 3D building model reconstruction.

1.1 Related work

Three-dimensional building modeling from remote sensing imagery has been an active research topic in the last decade. Various approaches developed till now can be mainly categorized into two strategies: data-driven and model-driven approaches. The model-driven approach selects beforehand a primitive building. Then, it

searches for the most appropriate model among basic building shapes in a model library and, by adjusting the parameter values of the selected model, fits it to the primitive (Tarsha-Kurdi et al., 2007). Several solutions based on this approach have been developed. (Brunn and Weidner, 1997), first, separate the buildings from vegetation using surface normals from DSMs. Then, they define the rooftop geometry based on detected surface normals and fit the polyhedral building descriptions to these. (Brenner and Haala, 1998) reconstruct 3D building rooftops by interpolating the basic building primitives obtained from the given ground plan to the LIDAR DSM segmented according to the direction of surface normals. (Arefi et al., 2010) introduce a new methodology to generate 3D models of buildings from a high-resolution DSM by, first, decomposing the buildings into several parts according to the number of existing ridge lines. Second, the pixels of building parts are projected into a plane and a 2D model is fitted to the projected points according to the roof type. Finally, the 2D models are translated back to 3D blocks which merged together to form a complete 3D model. (Lafarge et al., 2010) present the building reconstruction from DSM using the predefined primitive library. The buildings, considered as an ensemble of 3D parametric blocks, are placed on 2D approximate building footprints and optimized within a Bayesian decision framework. A similar approach is presented by (Huang et al., 2013). The predefined roof primitives are found to fit the building parts with given rules of combination and merging. For the sampling of the model parameters, a variant *markov chain monte carlo (MCMC)* sampler with specified jump mechanism is applied. The building footprints are not required for this methodology. Although a model-driven approach provides parametric and geometrical models without visual deformation, it encounters difficulties when processing complex roofs due to the limited building model libraries.

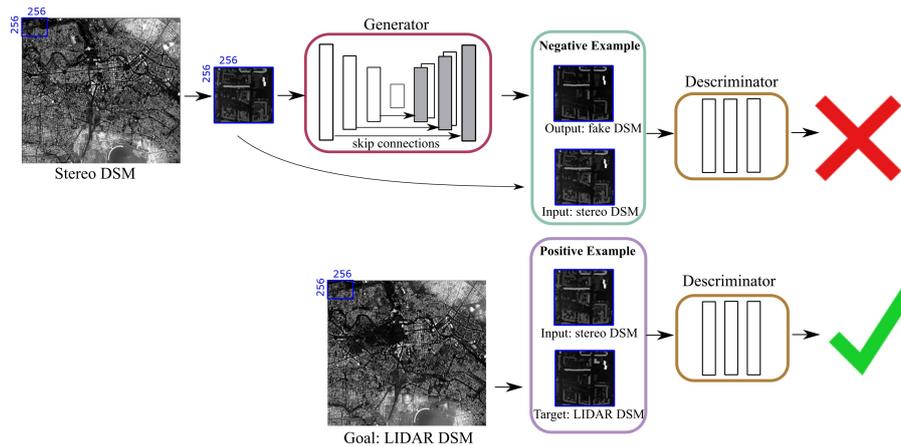


Figure 1. Schematic overview of the proposed method for the 3D building shape improvement in stereo *digital surface model (DSM)* by *conditional generative adversarial network (cGAN)*.

In contrast to the model-driven strategies, the data-driven methods model the buildings regardless of a specific model library. The roof planes are estimated by segmenting the point cloud into different parts employing methods like edge-based (Jiang and Bunke, 1994), region growing (Rottensteiner, 2006), *random sample consensus (RANSAC)* (Tarsha-Kurdi et al., 2008), clustering (Shan and Sampath, 2008), or 3D Hough transforms (Vosselman et al., 2001). (Sohn et al., 2009) proposed a method that reconstructs the buildings from both height and spectral images for creating prismatic building models based on a *binary space partitioning tree (BSP-Tree)*. (Zhou and Neumann, 2012) organize planar roof segments and roof boundary segments with “global regularities” considering orientation and placement similarities between planar elements in building structures. Despite the independence of data-driven strategies from the predefined model library, common limitations are the irregular or incomplete roof parts due to data noise. As a result, some works explore the possibilities to integrate both strategies for better building modeling. For example, (Verma et al., 2006) introduce the workflow for construction of a 3D geometric model for complex buildings incorporating *a)* the segmentation of roof and terrain points, *b)* roof topology inference by introducing a roof-topology graph and *c)* parametric roof composition. (Partovi et al., 2013) propose, first, to detect the ridgelines and analyze the height points of and perpendicular to ridge lines directions with the help of both orthorectified image and DSM. Then, the lines are fitted to ridge points using RANSAC and refined later by matching them or closing gaps in between. Finally, based on reconstructed ridge lines the roof model is selected.

In spite of efforts put into developing methodologies for automatic building shape enhancement in low-resolution urban DSM data, which are generated from satellite images, they are still not able to provide satisfactory results. In this work, we aim to implement a technique which will automatically, without any predefined library, refine the 3D building shape.

1.2 3D modeling with Generative Adversarial Networks

The recent developments in artificial neural networks provide the best solutions to the problems in various domains like computer vision, medicine, biology, and remote sensing. The revolutionary appearance of *generative adversarial networks (GANs)* (Goodfellow et al., 2014) attracted a lot of attention in the field of machine learning, as they offered a new possibility to generate

high-quality images. There have been some inspiring attempts in learning object representations in three-dimensional space based on GANs. (Wu et al., 2016) generate 3D objects from a probabilistic space by using a DSMs with volumetric convolutions. (Rezende et al., 2016) learn strong deep generative models and recover the 3D structures for simple shapes when 2D views are provided. Another approach proposed by (Yang et al., 2017) generates the complete 3D occupancy grid without any missing regions from a single depth view of the object. For this purpose, they combine the cGANs (Mirza and Osindero, 2014) together with autoencoders to infer accurate 3D structures of objects.

These new generation networks have been explored to work with shapes represented by a voxel volume. However, voxel representations are computationally very costly to handle in deep learning scenarios. This limits the output resolutions used in current implementations. In this paper, we explore the potential of a deep learning framework on complete depth image reconstruction, using data with continuous values, and present a novel methodology for generating a better-quality stereo DSMs with refined buildings shapes. To this end, a cGAN is trained to generate accurate LIDAR DSM-like height images from a noisy stereo DSMs. To our knowledge, it is the first time that cGANs accomplish a 3D object improvement for satellite depth-like images.

2. METHODOLOGY

2.1 Generative Adversarial Networks

GANs are a class of neural networks that were first introduced by (Goodfellow et al., 2014). The goal of GANs $G : z \rightarrow y$ is to output the realistic looking images y from some random input noise z . A typical GAN architecture consists of two modules: a generator G and a discriminator D , which compete against each other. The discriminator tries to distinguish generated images $G(z)$ from real ones, while the generator learns to create more and more realistic samples that are indistinguishable from real samples.

2.2 Conditional Generative Adversarial Networks

In this paper, we address a better-quality stereo DSM generation with refined buildings shapes. It means that we want to build a generative model capable of creating an artificial LIDAR DSM-like height image similar to a given input image, stereo DSM in

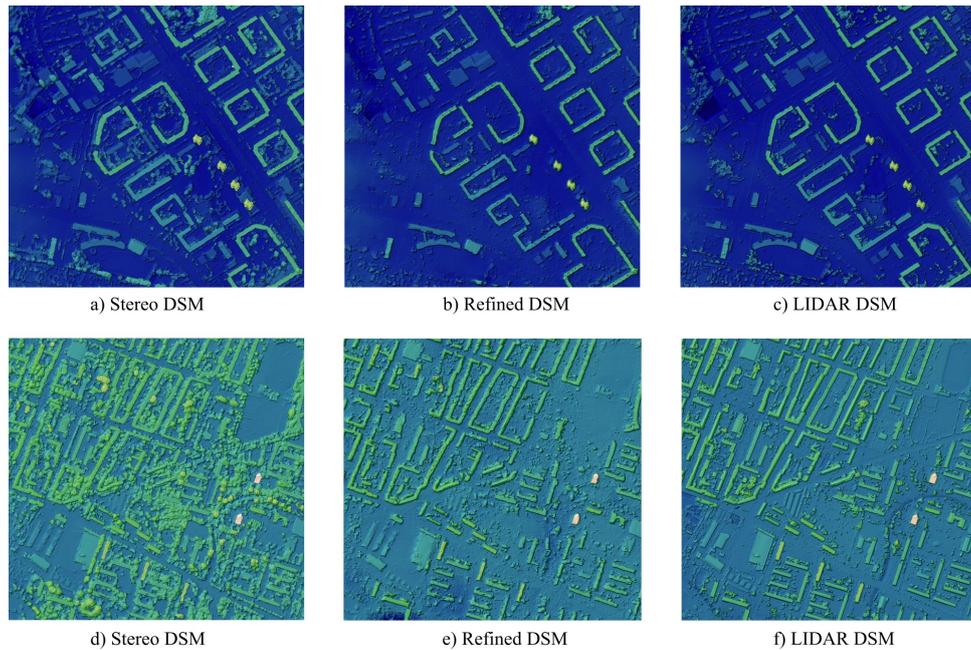


Figure 2. Examples of the generated DSM with refined 3D buildings shapes.

our case. Our method builds upon the adversarial system in a form of cGAN architecture proposed by (Isola et al., 2016). In opposition to the common GAN setup, cGANs $G : \{x, z\} \rightarrow y$ allow generating an artificial image similar to some known input image x . The problem can be defined as learning generative distribution of data through a two-player minimax game

$$\min_G \max_D \mathcal{L}_{cGAN}(G, D) = \mathbb{E}_{x, y \sim p_{real}(x, y)} [\log D(x, y)] + \mathbb{E}_{x, y \sim p_{real}(x, y), z \sim p_z(z)} [\log(1 - D(x, G(x, z)))] \quad (1)$$

where G intends to minimize this objective against the D that aims to maximize it. \mathbb{E} denotes the expected value, p_{real} the real data distribution and p_z the noise distribution. As the task of the generator is not only to fool the discriminator but also to create an artificial image close to the given output, the objective 1 is extended by a regularization term

$$\mathcal{L}_{L_1}(G) = \mathbb{E}_{x, y \sim p_{real}(x, y), z \sim p_z(z)} [\|y - G(x, z)\|_1] \quad (2)$$

which assures that the generator produces output images near the ground truth in a L_1 sense (Isola et al., 2016). Our final objective function can be expressed as

$$G^* = \arg \min_G \max_D \mathcal{L}_{cGAN}(G, D) + \lambda \mathcal{L}_{L_1}(G) \quad (3)$$

where λ is a balancing hyper-parameter. In our particular case, the generator G tries to synthesize realistic LIDAR DSM-like height images to fool the discriminator D and D , in its turn, tries to tell which samples are real or generated. The illustration of the training procedure is depicted in Figure 1.

2.3 Network Architecture

The architecture of cGANs is organized as follows: The generator G is represented by an U-Net architecture (Ronneberger et al., 2015) with tanh activation function at the top layer. The U-Net is an encoder-decoder type of network which combines the encoder layer i with up-sampled decoder layer $n - i$ at every stage by skip

connections, where n represents the total number of layers. Each skip connection is simply realized through the concatenation of all feature maps at layer i with those at layer $n - i$. The idea behind having skip connections is that an encoder-decoder type of network progressively down-samples the input through a series of layers until a bottleneck layer, at which point the process is coded back. The skip connections allow the decoder to recover important details that are lost due to the down-sampling in the encoder.

The discriminator D is realized via several convolutional layers with a sigmoid activation function as the last layer. It is a binary classification network which outputs the probability that the input image belongs either to class 1 ("real") or class 0 ("fake").

3. STUDY AREA AND EXPERIMENTS

Experiments have been performed on data over the Berlin city, Germany, covering 410 km². As input data, we use stereo DSMs derived from WorldView-1 very high-resolution stereo panchromatic imagery with a resolution of 0.5 m using *semi-global matching (SGM)* method (d'Angelo and Reinartz, 2011). As ground truth, a LIDAR DSM from *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. It was generated from airborne laser scanning data of 1 × 1 meter grid and up-sampled to the resolution of 0.5 m. As the LIDAR DSM was generated from the last pulse data, there is no or much less vegetation within a scene in comparison to the stereo DSM.

The experiment was performed using a slightly extended version of the cGAN architecture as implemented in (Isola et al., 2016). For the learning process, we prepared the training data consisting of 21480 pairs of patches covering an area of 353 km². For tuning the hyper-parameters the validation data covering 6 km² was used. The patch size is 256 × 256 pixels which fits into the available GPU. The upscale and random crop data augmentation from

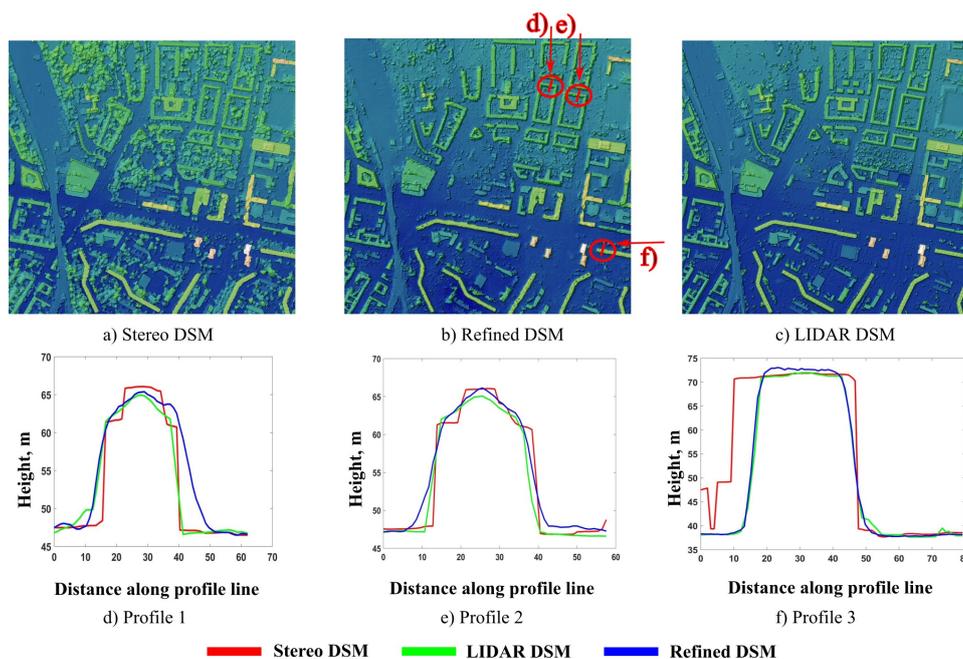


Figure 3. Examples of the generated DSM with refined 3D buildings shapes and roofs ridge lines.

original implementation was replaced by a random cropping of the tiles up to one tile size without up-scaling. It makes the network more robust and more flexible, as, at one moment, the whole building can be visible to the network but at another moment only part of it. Still, it is the same building with the same properties. The test data covers 50 km² and consist of 12122 patches generated with an overlap of 128 pixels in both directions.

The network is trained with minibatch *stochastic gradient descent* (SGD) using the ADAM optimizer (Kingma and Ba, 2014) with an initial learning rate of $\alpha = 0.0002$ and momentum parameters $\beta_1 = 0.5$ and $\beta_2 = 0.999$. We chose the weighting hyper-parameter $\lambda = 100$ after performing the experimental training and examining the resulting generated images and their profiles. Additionally, we follow the common practice and train G to maximize the $\log D(x, G(x, z))$ instead of minimizing $\log(1 - D(x, G(x, z)))$. The network is trained over 200 epochs with a batch of 5 training instances. A single NVIDIA GeForce GTX Titan X GPU was used for training and testing.

To investigate how similar the reconstructed elevation model is to the given LIDAR DSM, we performed the following experiment: Based on the idea from our previous work (Bittner et al., 2017) we adapted a *fully convolutional network* (FCN) (Long et al., 2015) and trained it on the LIDAR DSM for building footprints extraction task, and analyze its performance on the test area. As a reference, binary building mask from *open street map* (OSM) is taken. Next, we feed to the trained FCN-8s our generated LIDAR-like DSM covering the same area and see if and how well the network is able to predict the binary building mask for a generated data. This building mask is compared then against the ground truth label map using the standard FCN-score metrics. As we believe that DSMs generated by our method are close to the ground truth LIDAR DSM, we expect that the FCN applied to the generated image detects the buildings.

4. RESULTS AND DISCUSSION

In this section, we present and discuss the results obtained from the considered experiment for 3D building shape refinement. The network performance is analyzed on the test area which was never shown to the network during training. For the inference phase, only the trained generator network G is required. The predictions are obtained for each patch from the test dataset separately to generate a predicted map of the same size as the input patch. Then the patches are merged together in order to generate an image with the same size as the original test area.

4.1 Qualitative Evaluation

The examples of generated LIDAR DSM-like height images are illustrated in Figures 2(b),(e) and 3(b). It can be clearly seen that the cGAN architecture is able to reconstruct an elevation model of the same resolution as the input image, geometric structures of

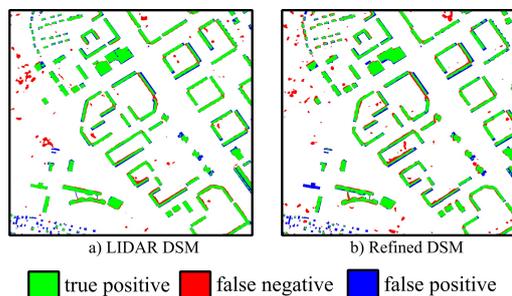


Figure 4. Pixel-based evaluation of footprint extraction from LIDAR and generated DSMs compared to OSM.

Table 1. FCN-score metrics evaluated on “real” DSM vs “fake” DSM compared to OSM.

	Mean acc.	Mean IoU	Overall acc.	IoU	$F_{meas.}$
“real” DSM	86.8	80.4	94	67.7	80.8
“fake” DSM	80.3	71.9	91	54	70

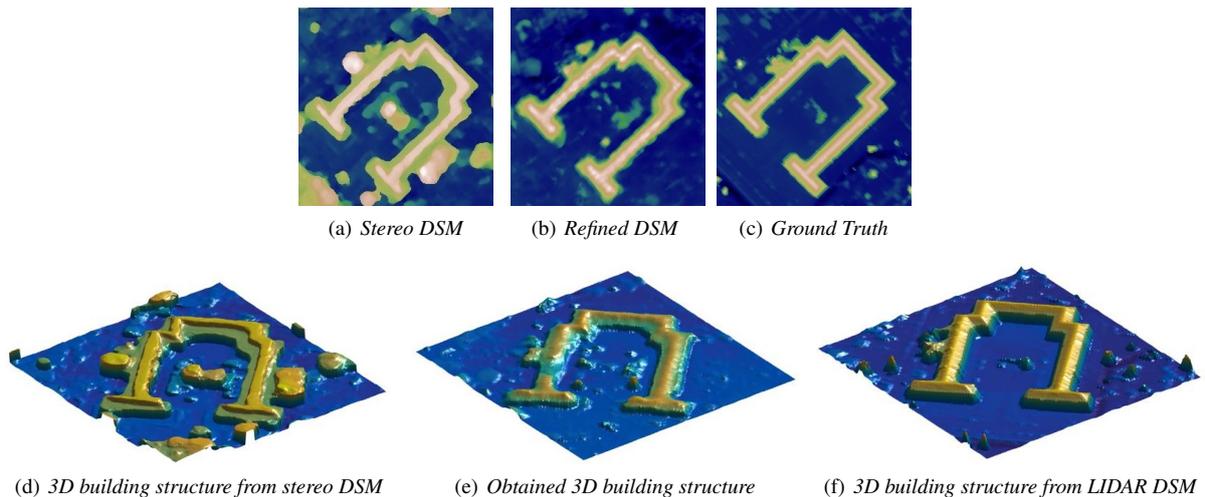


Figure 5. Example of the generated DSM with refined 3D buildings shapes.

buildings from stereo DSM are preserved in the generated sample and close to the LIDAR DSM. Besides, the network has learned about the much smaller amount of vegetation from ground truth data. However, it could look like some constructions are missing on the generated elevation model in comparison to LIDAR DSM. This is due to the time difference of data acquisition between available stereo DSM and given LIDAR DSM. The network is only able to refine the buildings presented on the input data. By investigating the profiles of three selected buildings, highlighted by the red line in Figure 3(b), we can confirm that the cGAN successfully learned the 3D building representation close to the LIDAR data representation (see Figures 3(d)-(f)). Regarding 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 a more realistic view and is geometrically more correct. The profile (f) also shows very close resemblance of resulting building shape to the ground truth, especially regarding the width and borders of the building, although the input 3D shape is much wider due to the trees in the neighborhood (see Figure 3(a)). Thus, it is evident that the trees have much less influence on the building shapes. As we are interested in the full geometry of the buildings, we illustrate one reconstructed building in 3D space (see Figure 5). From the illustration we can see that the structure of the building, especially at the corners and ridge line, is not only preserved but also refined and it is very similar to the ground truth. However, one can notice an inconsistency on the roof on the left side of the building. This can be explained by the loss of information into consistency on the input data (see Figure 5(d)) as at this part of the building the ridge line does not have a normal transition to the side border of the building but directly goes down to the ground. This confuses the network.

4.2 Quantitative Evaluation

To quantify the building mask predictions from a generated DSM, we evaluated metrics commonly used in deep learning for semantic segmentation problems. They are *mean accuracy*, *mean intersection over union (IoU)*, *overall accuracy* and *F_{measure}* taken from Long *et al.* (Long *et al.*, 2015), and the IoU adapted for the problem, where the number of pixels belonging to the objects are less compared to those belonging to the background (Maggiori *et al.*, 2017). Table 1 reports the performance of the FCN model

on two data. Here, one can see that the model is able to detect buildings from the generated DSM with an adequate accuracy that confirms the correctness of the generated elevation model's geometry. From the perspective of shape similarity metrics for the selected region in Figure 4, where footprints of the extracted buildings are close to the reference even for very curved and sophisticated shapes, one can conclude that the generated DSM is a valuable data source and can later be used for other remote sensing applications.

5. CONCLUSION

So far, *generative adversarial networks (GANs)* have been primarily used for modeling natural images. In this work, we presented a new approach for automatic better-quality *light detection and ranging (LIDAR)*-like *digital surface model (DSM)* generation with refined 3D building shapes from noisy stereo DSM applying a *conditional generative adversarial network (cGAN)*. The evaluation of the results show the potential of the proposed method to generalize over diverse city constructions even coping well with complex buildings. Additionally, we reveal that the network is capable not only to generate the objects in 3D space but also to improve their shapes and, especially, roof ridge lines to give the buildings more realistic appearance. Besides, the network was able to reconstruct the buildings without any influence of trees, though, stereo DSM is rich on vegetation. Overall, the introduced methodology opens up new horizons towards the goal of 3D city modeling. The future intention is to generate even more realistic artificial elevation models with improved quality of building structure.

References

- Arefi, H., Hahn, M. and Reinartz, P., 2010. Ridge based decomposition of complex buildings for 3d model generation from high resolution digital surface models. ISPRS.
- Bittner, K., Cui, S. and Reinartz, P., 2017. Building extraction from remote sensing data using fully convolutional networks. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences* 42, pp. 481.

- Brenner, C. and Haala, N., 1998. Rapid acquisition of virtual reality city models from multiple data sources. *International Archives of Photogrammetry and Remote Sensing* 32, pp. 323–330.
- Brunn, A. and Weidner, U., 1997. Extracting buildings from digital surface models. *International Archives of Photogrammetry and Remote Sensing* 32(3 SECT 4W2), pp. 27–34.
- d'Angelo, P. and Reinartz, P., 2011. Semiglobal matching results on the isprs stereo matching benchmark. *ISPRS Hannover Workshop* 38(4/W19), pp. 79–84.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y., 2014. Generative adversarial nets. In: *Advances In Neural Information Processing Systems*, pp. 2672–2680.
- Huang, H., Brenner, C. and Sester, M., 2013. A generative statistical approach to automatic 3d building roof reconstruction from laser scanning data. *ISPRS Journal of Photogrammetry and Remote Sensing* 79, pp. 29–43.
- Isola, P., Zhu, J.-Y., Zhou, T. and Efros, A. A., 2016. Image-to-image translation with conditional adversarial networks. *arXiv preprint arXiv:1611.07004*.
- Jiang, X. and Bunke, H., 1994. Fast segmentation of range images into planar regions by scan line grouping. *Machine vision and applications* 7(2), pp. 115–122.
- Kingma, D. P. and Ba, J., 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Lafarge, F., Descombes, X., Zerubia, J. and Pierrot-Deseilligny, M., 2010. Structural approach for building reconstruction from a single dsm. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32(1), pp. 135–147.
- Long, J., Shelhamer, E. and Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440.
- Maggiori, E., Tarabalka, Y., Charpiat, G. and Alliez, P., 2017. Convolutional neural networks for large-scale remote-sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing* 55(2), pp. 645–657.
- Mirza, M. and Osindero, S., 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.
- Partovi, T., Arefi, H., Krauß, T. and Reinartz, P., 2013. Automatic model selection for 3d reconstruction of buildings from satellite imagery. *ISPRS Archives* 40, pp. 315–320.
- Rezende, D. J., Eslami, S. A., Mohamed, S., Battaglia, P., Jaderberg, M. and Heess, N., 2016. Unsupervised learning of 3d structure from images. In: *Advances In Neural Information Processing Systems*, pp. 4996–5004.
- Ronneberger, O., Fischer, P. and Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, pp. 234–241.
- Rottensteiner, F., 2006. Consistent estimation of building parameters considering geometric regularities by soft constraints. *International Archives of Photogrammetry Remote Sensing and Spatial Information Sciences* 36(Part 3), pp. 13–18.
- Shan, J. and Sampath, A., 2008. Building extraction from lidar point clouds based on clustering techniques. *Topographic Laser Ranging and Scanning: Principles and Processing* pp. 423–446.
- Sohn, G., Huang, X. and Tao, V., 2009. A data-driven method for modeling 3d building objects using a binary space partitioning tree. *Topographic Laser Ranging and Scanning: Principles and Processing; CRC Press, Taylor Francis Group: New York, NY, USA* pp. 479–509.
- Tarsha-Kurdi, F., Landes, T. and Grussenmeyer, P., 2008. Extended ransac algorithm for automatic detection of building roof planes from lidar data. *The Photogrammetric Journal of Finland* 21(1), pp. 97–109.
- Tarsha-Kurdi, F., Landes, T., Grussenmeyer, P. and Koehl, M., 2007. Model-driven and data-driven approaches using lidar data: Analysis and comparison. In: *ISPRS Workshop, Photogrammetric Image Analysis (PIA07)*, pp. 87–92.
- Verma, V., Kumar, R. and Hsu, S., 2006. 3d building detection and modeling from aerial lidar data. In: *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, Vol. 2, IEEE, pp. 2213–2220.
- Vosselman, G., Dijkman, S. et al., 2001. 3d building model reconstruction from point clouds and ground plans. *International Archives of Photogrammetry Remote Sensing and Spatial Information Sciences* 34(3/W4), pp. 37–44.
- Wu, J., Zhang, C., Xue, T., Freeman, B. and Tenenbaum, J., 2016. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In: *Advances in Neural Information Processing Systems*, pp. 82–90.
- Yang, B., Wen, H., Wang, S., Clark, R., Markham, A. and Trigoni, N., 2017. 3d object reconstruction from a single depth view with adversarial learning. *arXiv preprint arXiv:1708.07969*.
- Zhou, Q.-Y. and Neumann, U., 2012. 2.5 d building modeling by discovering global regularities. In: *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, IEEE, pp. 326–333.