SEMI-AUTOMATIC CITYSCAPE 3D MODEL RESTORATION USING GENERATIVE ADVERSARIAL NETWORK

V.Gorbatsevich*, B. Kulgildin, M. Melnichenko,O. Vygolov ,Yu. Vizilter

FEDERAL STATE UNITARY ENTERPRISE «STATE RESEARCH INSTITUTE OF AVIATION SYSTEMS», Russian Federation (gvs, kulgildin, mmelnechenko, o.vygolov, viz)@gosniias.ru

Commission II, WG II/5

KEY WORDS: CNN, Cityscape 3D model, GAN, Monocular 3D Reconstruction, Heightmaps

ABSTRACT:

The paper addresses the problem of a city heightmap restoration using satellite view image and some manually created area with 3D data. We propose the approach based on generative adversarial networks. Our algorithm contains three steps: low quality 3D restoration, buildings segmentation using restored model, and high-quality 3D restoration. CNN architecture based on original ResDilation blocks and ResNet is used for steps one and three. Training and test datasets were retrieved from National Lidar Dataset (United States) and the algorithm achieved approximately MSE = 3.84 m on this data. In addition, we tested our model on the completely different ISPRS Potsdam dataset and obtained MSE = 5.1 m.

1. INTRODUCTION

Cityscape 3D models are widely used in practical applications, e.g. VR and computer games, data augmentation, etc. Since the creation of highly detailed 3D models requires a bunch of time-consuming handwork, the solutions that can automate this process are still in high demand.

Many modern techniques can automatically create such type of models using LIDAR data or an image flow, which is sufficient to implement structure from motion approaches. When this data is not fully available, the methods of 3D reconstruction from a single image are applied. In recent years, as in almost all other computer vision tasks, convolutional neural networks (CNN) are gaining well-deserved popularity as the core of these methods.

In this work, we address the problem of automatic 3D cityscape reconstruction using a satellite image and some manually reconstructed parts of 3D scene (e.g. several buildings). This approach is justified in fast semi-supervised 3D modeling of the real environment when no high precision needed (for example, for synthetic dataset generation).

The heightmap 3D models representation is considered, i.e. the 2D matrix that contains surface elevation data. This allows us to build the algorithm on classical CNNs instead of voxel or graph CNNs. Last years, generative adversarial networks (GANs) have made a great performance gain for such types of problems and are employed in this work as well.

Our GAN Generator-CNN receives satellite image and additional data (a building mask or the part of a heightmap) as the input and full heightmap as the output. For 3D reconstruction, two types of CNNs are used - MapNet and MaskNet. MapNet produces the heightmap from satellite image and the heightmap part or the buildings mask. MaskNet delivers intermediate buildings mask from the heightmap. Our CNNs has original architecture.

The training data includes LIDAR data from National Lidar Dataset (USA) for New-York City and corresponding satellite images from Google, Yandex, Nokian and Bing online services using QGis software. On test dataset MSE (mean square error) = 3.8 m is gained for restored heightmaps. The proposed NetMask outperforms popular CNN architectures such as ResNet36, DeepLabv3 and U-Net.

In addition, we have tested our algorithm on ISPRS Potsdam dataset and obtained MSE = 5.1 m. It should be noted that the Potsdam dataset is significantly different from our training dataset in sense of presented building types (the algorithm has seen a lot of skyscrapers and high buildings in the training dataset of New-York City).

2. RELATED WORKS

Monocular 3D reconstruction. In our work, for the 3D reconstruction of cityscapes we use two data sources – aerial(or satellite) images and some manually created area with 3D data. Very similar task of 3D reconstruction from single image is well-known in computer vision (El-Hakim, 2001), (Remondino, 2003),(Remondino, 2006).

As in other computer vision problems, the methods based on deep learning(Girdhar, 2016),(Choy, 2016),(Richter, 2018),(Shin, 2018),(Long, 2015),(Isola, 2015), (Wu, 2017), (Huang, 2015),(Zheng, 2013) are successfully used in this area. Some methods were developed for voxel 3D model restoration from a single depth map (Zheng, 2013), (Firman, 2016). In (Girdhar, 2016) CNN for image to voxel 3D model translation was introduced. The CNN architecture is an auto-encoder for direct voxel model prediction. Unfortunately, this approach can work only with small 3D models (up to 20×20×20 voxels). An approach that combines single-view and multi-view reconstruction modes was described in (Choy, 2016). In (Knyaz, 2018a) more accurate CNN that can generate voxel models of complex scenes with multiple objects was proposed.

Using heightmaps, landscape reconstruction problem can be easily transformed to classical image-to-image translation problem.

Image-to-image translation. Well-known grayscale colorization and style imitation methods (Zhang, 2016), (Gatys, 2015) are the examples of the early CNN based image-to-image translation methods. The next level of quality and the ability to

* Corresponding author

solve this problem in general are promised by the generative adversarial networks. The first one was Pix2Pix (Isola, 2017) model that can learn any type of high quality image-to-image translation using training datasets of corresponding image pairs. In (Zhu, 2017) a new generative adversarial network called CycleGAN was proposed, that have the ability to learn on unpaired datasets.

High quality 3D reconstruction. There are also some popular approaches for high quality 3D terrain reconstruction based on stereo matching (Knyaz, 2018b) and structure from motion (Knyaz, 2017). These approaches provide high quality 3D terrain models but require more input data - stereopairs or image sequences.

3. HEIGHTMAP

Normally, 3D cityscape models are represented as a set of points or triangles with texture. This type of representation is common for 3D modeling software and graphics cards. Unfortunately, this type of representation cannot be used with regular convolutional neural networks, since it can take only 2D fixed grids as an input. Of course, today there are, for example, some good realizations of graph-based neural networks, which can work directly on graph-like data structures. However, in fact, the theory of graph-based network are not as mature as regular CNNs.On the other hand for terrains, there is a 2D fixed grid data representation called heightmap. Heightmap or heightfield is a 2D matrix used mainly as Discrete Global Grid in secondary elevation modeling. Each element of this matrix corresponds to a point in 3D model and the value of the element represents the elevation in this point. The values of elevation are set relatively to some "zero" level. Such type of heightmap can be easily converted by triangulation into 3D mesh. On figure 1 an example of heightmap and corresponding 3D model for landscape are shown.

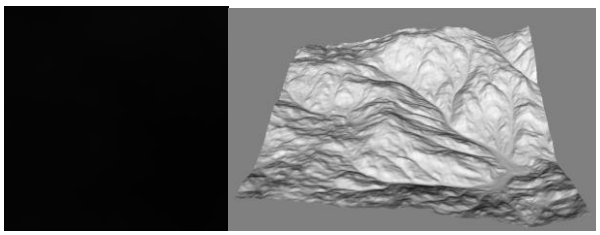

Figure 1. Heightmap (left) and corresponding 3D landscape

The heightmap is similar to an image in terms of data structuring. Therefore, classical CNNs can be used for 3D landscape processing. For example in (Vizilter, 2019) heightmaps are used for 3D landscape restoration using CNN.

In our work we also use heightmaps for cityscape representation. On Figure 2 an example of heightmap and corresponding 3D model for cityscape are shown.

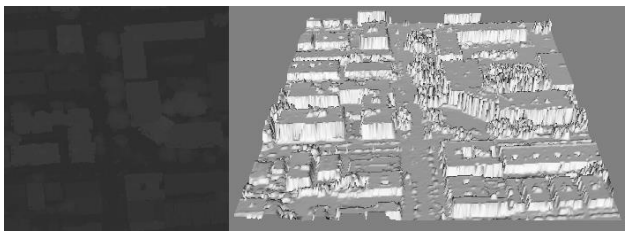

Figure 2. Heightmap (left) and corresponding cityscape 3D model

4. METHOD

Simple Generative adversarial networks generate some signal $\hat{B}$ based on random noise vector z, $G: z \rightarrow \hat{B}$. Conditional GAN transforms an input image A and vector z to an output $G: \{A, z\} \rightarrow \hat{B}$. The input A can be the image that is transformed by the generator network G. The discriminator network D is trained to distinguish "real" signals from the target domain B from the "fakes" B produced by the generator. Generator and discriminator are trained simultaneously. Discriminator provides the adversarial loss that enforces the generator to produce "fakes" $\hat{B}$ that cannot be distinguished from "real" signal B.

In our case, we have classical Conditional GAN problem, i.e. we have two inputs: aerial image and low quality heightmap (interpolation from a set of points), and get dense landscape model as an output(see Figure 3). Data fusion is made by a concatenation procedure.

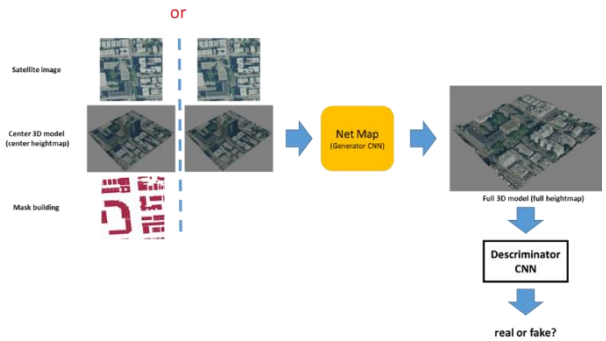

Figure 3. 3D reconstruction as GAN problem.

For 3D reconstruction, two types of CNNs are employed - MapNet and MaskNet. MapNet produces the heightmap from satellite image and the heightmap part or the buildings mask. Also the special intermediate CNN – MaskNet is used that produces intermediate buildings mask from the heightmap and input image to improve 3D reconstruction quality. The algorithm pipeline is shown on Figure 4.

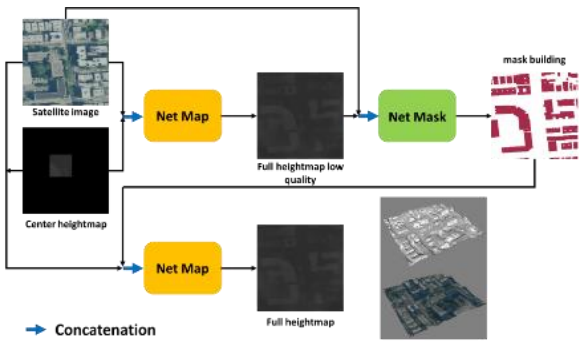

Figure 4. Proposed 3D cityscape reconstruction pipeline.

The reconstruction process can be divided in three stages:

1. Low quality 3D restoration from input image and heightmap part using MapNet CNN;
2. Building mask estimation based on low quality 3D model and input image using MaskNet CNN
3. High quality 3D restoration using MapNet CNN, which depends on data from previous stages.

5. IMPLEMENTATION DETAILS

To determine the height of a point, it is necessary to know what type of object it belongs to, what height the given object is, and how uniform it is. Moreover, most of the points, which have a height that differs from the ground level, belong to buildings whose scale can be diverse (small - private sector houses, medium and large - city buildings, hangars, etc.). The height of buildings is indirectly reflected in their shadows size and the deviation of the building roof position relative to the foundation. Moreover, depending on the height of the object and the angle of the sun at which the satellite image is acquired, its shadow on the image can take from just a few to hundreds of pixels. Also, the height of the object depends on the area in which it is located (private sector, residential quarter, skyscrapers, etc.). Thus, to determine the height of a point, it is necessary to take into account both closely situated and distant features. Since the construction of a 3D model is carried out using a satellite image (the resolution of which is a couple of meters or tens of centimeters), it is necessary to minimize the loss of spatial resolution, which can lead to a decrease in the accuracy of height maps restoration.

5.1 ResDilation block

For good 3D reconstruction we need multi-scale features. Such type of features can be created using convolution with different dilation (dilated convolutions (Fisher, 2016)). Following (Zhou, 2018) multi-scale features can be combined in one layer. In this paper, we propose a new layer called–ResDilation block that combines ideas of residual block from ResNet and multi-scale features.

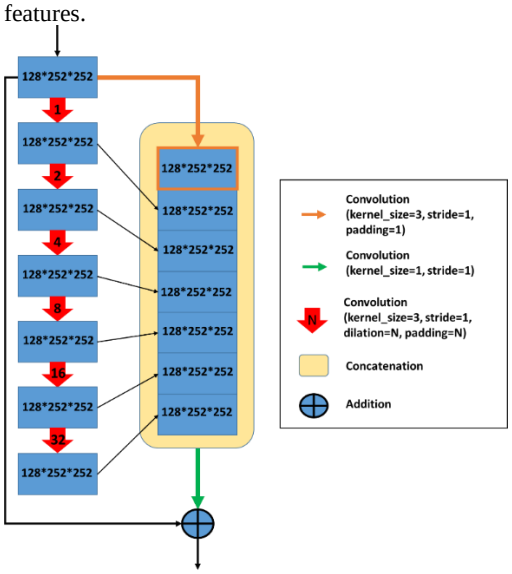

Figure 5. ResDilation block architecture.

ResDilation block (shown on Figure 5) contains a sequence of convolutional layers with different dilations. For ResDialtion block with convolutions dilations (1->2->4->8->16->32) the receptive field is 127x127. The block output is based on local information from first convolution layers and on global information from last layers. Concatenation is used to prevent any changes in global and local information.
ResDilation block is aimed at combining differently distant features and, depending on the position of the block in the network, determine what feature scale is important at this level.

5.2 MapNet architecture

The original network architecture based on ResDilation block is shown in Table 2.

CNN blocks	
Name	Layers
Conv_block(n)	Conv2d (num_filter=n, kernel_size=3, stride=1, padding=1) BatchNorm ReLU
ResDilation	Figure 3

Table 1. MapNet block.

Input	Layers
(Satellite image, center heightmap) or (Satellite image, center heightmap, mask build)	Conv_block_1(64)
Conv_block_1	Conv_block_2(128)
Conv_block_2	ResDilation(128) x 9
ResDilation	Conv_block_3(128)
Conv_block_3	Conv_block_4(128)
Conv_block_4	Conv2d (num_filter=1, kernel_size=3, stride=1)

Table 2 –MapNet architecture

5.3 Training process

During training we use generative adversarial network ideology with NetMap network as a generator CNN and the original network described in Table 3 as a discriminator CNN.

pixel-ResDilation
Layers
Conv2d (num_filter=64, kernel_size=3, stride=1) BatchNorm ReLU Conv2d (num_filter=128, kernel_size=3, stride=1) BatchNorm ReLU ResDilation(128) Conv2d (num_filter=1, kernel_size=1, stride=1)

Table 3 MapNet architecture

Training process and basic loss functions are similar to Pix2Pix(Isola, 2017). To prevent model from smoothing, the special border loss is added – L1 loss between "height difference maps", produced using Laplace operator from the heightmaps of ground truth and the generator CNN output. So the final loss is:

$$G^* = \arg\min_G \max_D L_{cGAN}(G,D) + \gamma(L_{L1}(G) + L_{dL1}(G)) \quad (2)$$

where

$$L_{GAN}(G,D) = E_y[\log D(y)] + E_x[\log(1 - D(G(x)))]$$
$$L_{L1}(G) = E_{x,y}[\|y - G(x)\|_1]$$
$$L_{dL1}(G) = E_{x,y}[\|\nabla y - \nabla G(x)\|_1]$$

G – generator CNN
D – discriminator CNN
x – input data
y – ground truth data
∇y – difference map
γ – equal to 100 in this work.

HRNet (Wang, 2019) was selected for MaskNet architecture. Cross entropy loss for semantic segmentation for two classes (building and background) with learning parameters from original paper are used in training process that contains several stages (see Figure 4):

Stage 1: MapNet#1 CNN pre training on low quality 3D model restoration.

Stage 2: MaskNet CNN pre training on semantic segmentation using low quality 3D model and aerial photo as input.

Stage 3: MapNet#2 CNN pre training on high quality 3D model restoration.

Stage 4: All three CNNs are trained simultaneously using full pipeline (Figure 1).On this stage we use Adam optimizer with $\beta 1 = 0.5$, $\beta 2 = 0.999$. Initial learning rate is 0.0001, learning rate decay is 0.1.

6. EXPERIMENTS

In our experiments we use PyTorch framework and 4 Nvidia Tesla P100 for training and testing.

6.1 Database

We use public LIDAR database from National Lidar Dataset (United States) for New-York city (http://gis.ny.gov/elevation/lidar-coverage.htm), 2017 and corresponding satellite images downloaded from Google, Yandex, Nokian and Bing map engines(using QGIS software), with 1 meter per pixel resolution. Training and testing datasets were created from this data.
Training dataset contains 18000 samples and 576000 unique pairs (3D model and 256x256 RGB image). Test dataset contains 2000 pairs.

6.2 Training results

Measurement quality is estimated by mean squared error (MSE) metric between ground truth and reconstructed heightmap, and building mask quality is evaluated by mean Intersection over Union (mIoU) metric.
The segmentation quality has been tested with different input data. Results are given in Table 4 and show that using low quality heightmaps leads to quality improvement.

Input	mIoU
Satellite image	85.2
Satellite image, heightmap low quality	**87.5**

Table 4. Segmentation quality on test dataset

Table 5 shows 3D reconstruction results for different processing pipelines using different input data (image, center, building mask).

Satellite image	Center heightmap (64×64)	Mask build	RMSE, м
✓	×	×	9.45
✓	✓	×	4.44
✓	×	✓	8.76
✓	✓	✓	**3.84**

Table 5. 3D reconstruction with different pipelines

Also we tried different popular architectures as MapNet like ResNet, U-Net,DeepLabv3.

Network	RMSE, м
Unet	6.1
Resnet 36 layers	5.54
DeepLabv3	7.01
Net Map	**3.84**

Table 6. NetMap architecture comparison using full training pipeline

Proposed architecture leads to significant better quality in comparison to competitors.

Also our approach was tested on completely different ISPRS Potsdam dataset from http://www2.isprs.org/commissions/comm3/wg4/tests.html and obtained RMSE = 5.1 without any pretraining. It should be noted that the Potsdam dataset is completely different from our training dataset in sense of presented building types (in New York there are a lot of skyscrapers and high buildings). On Figure 6 and 7 the qualitative example of 3D reconstruction on Potsdam dataset is shown.

Figure 6. Qualitative example of 3D reconstruction on Potsdam dataset. Ground truth (left) and restored model (right).

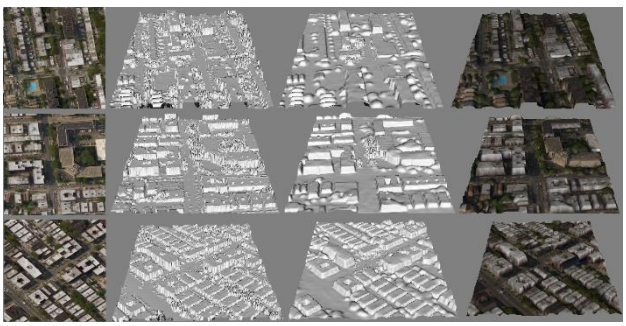

Figure 7. Qualitative example of 3D reconstruction on Potsdam dataset. From left to right: Aerial image, Ground truth 3D heightmap, Restored 3D heightmap, Restored 3D heightmap with textures

7. CONCLUSIONS

The paper addresses the problem of a city heightmap restoration using satellite view image and some manually created area with 3D data. This problem is kind of monocular 3D reconstruction problem. To solve this problem, we propose an approach that uses a set of convolution neural networks with proxy tasks. We use heightmap 3D models representation, i.e. the 2D matrix that contains surface elevation data. This allows us to use classical CNNs instead of voxel or graph CNNs.
Following very popular Pix2Pix technique, we use generative network approach to improve 3D restoration quality. Our Generator-CNN receives satellite image and additional data (a building mask or the part of a heightmap) as the input and full heightmap as the output. L1 and adversarial loss are used as a

loss function. To prevent 3D model smoothing the special border loss is added – L1 loss between "height difference maps", produced using Laplace operator from the heightmaps of ground truth and generator CNN output. For 3D reconstruction, two types of CNNs are used - MapNet and MaskNet. MapNet produces the heightmap from satellite image and the heightmap part or the buildings mask. MaskNet produces intermediate buildings mask from the heightmap. Both MapNet CNNs are trained using adversarial approach mentioned above. MaskCNN is trained in classical semantic segmentation way. After pretraining all three networks are trained simultaneously. For MapNet we propose architecture based on ResDilation blocks. Our test shows that proposed architecture significant better than popular architectures.

Lidar data from National Lidar Dataset (USA) for New-York City and corresponding satellite images from google, yandex, nokian and bing online services are used for training. QGis software is used to create satelliteview images with 1 meter on pixel resolution. On test dataset (2000 samples) MSE(mean square error) = 3.8 m is gained for restored heightmaps. The proposed NetMask outperforms popular CNN architectures such as ResNet36 (MSE = 5.54), DeepLabv3 (MSE=7.01) and U-Net (MSE=6.1). In addition, we tested our CNNs on ISPRS Potsdam dataset and obtained MSE = 5.1 m. It should be noted that the Potsdam dataset is completely different from our training dataset in sense of presented building types (in New York there are a lot of skyscrapers and high buildings).

The proposed algorithm is not supposed to be used for photogrammetric measurements due to the provided accuracy, but it can be effectively used for the automatic generation of surrounding 3D models.

ACKNOWLEDGEMENTS

This work was performed with the support of Grant No. 19-07-01140 of Russian Foundation for Basic Research (RFBR)

REFERENCES

Wang J., Sun K., Cheng T., Jiang B., Deng C., Zhao Y., Liu D., Mu Y., Tan M., Wang X., Liu W., Xiao B.,2019. Deep High-Resolution Representation Learning for Visual Recognition. CoRR, abs/1908.07919 (2019)

Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.,2017. Image-to-Image Translation with Conditional Adversarial Networks, Proc. CVPR2017. Papers 5967–5976(2017)

El-Hakim, S.,2001. A flexible approach to 3d reconstruction from single images, ACM SIGGRAPH. Volume 1. Papers 12-17 (2001)

Remondino, F., Roditakis, A.,2003. Human figure reconstruction and modeling from single image or monocular video sequence, Proc. 3DIM 2003. Papers 116–123(2003)

Remondino, F., El-Hakim, S., "Image-based 3D Modelling. A Review," Proc. The Photogrammetric Record, 269–291(2006)

Girdhar, R., Fouhey, D.F., 2016. Learning a predictable and generative vector representation for objects. Proc. M.R.E.C. 2016 Springer (chapter 34) 702–722 (2016)

Choy, C.B., Xu, D., Gwak, J., Chen, K., Savarese, S. 2016. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. Proc ECCV 2016 (2016)

Richter, S.R., Roth, S. 2018. Matryoshka Networks: Predicting 3D Geometry via Nested Shape Layers. Proc. arXiv.org (2018)

Shin, D., Fowlkes, C., Hoiem, D. 2018. Pixels, voxels, and views: A study of shape representations for single view 3d object shape prediction. Proc. CVPR 2018 (2018)

Long, J., Shelhamer, E., Darrell, T. 2015.Fully convolutional networks for semantic segmentation. Proc. CVPR2015. Papers 3431–3440(2015)

Wu, J., Wang, Y., Xue, T., Sun, X., Freeman, W.T., Tenenbaum, J.B. 2017. MarrNet: 3D Shape Reconstruction via 2.5D Sketches. arXiv.org (2017)

Huang, Q., Wang, H., Koltun, V. 2015. Single-view reconstruction via joint analysis of image and shape collections. ACM Transactions on Graphics. Papers 1–87 (2015)

Zheng, B., Zhao, Y., Yu, J.C., Ikeuchi, K., Z hu, S.C. 2013. Beyond point clouds: Scene understanding by reasoning geometry and physics. Proc. CVPR2013 (2013)

Firman, M., Mac Aodha, O., Julier, S., Brostow, G.J., 2016. Structured prediction of unobserved voxels from a single depth image. Proc. CVPR2016 (2016)

Girdhar, R., Fouhey, D.F., 2016. Learning a predictable and generative vector representation for objects. Proc M.R.E.C. 2016. Springer (chapter 34) 702–722 (2016)

Knyaz V.A., Kniaz V.V., Remondino F., 2018. Image-to-Voxel Model Translation with Conditional Adversarial Networks. Proc. ECCV 2018 Workshops (2018)

Zhang R., 2016. Colorful Image Colorization. ECCV2016. Pages 649–666 (2016)

Gatys L., Ecker A., Bethge M., 2015. A Neural Algorithm of Artistic Style. Proc. CoRR2015 (2015)

Zhu J., Park T., Isola P., Efros A., 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. Proc. ICCV2017 (2017)

Knyaz V., 2018. Deep learning performance for digital terrain model generation. Proc. SPIE 10789, Image and Signal Processing for Remote Sensing XXIV, 107890X (9 October 2018)

Knyaz V., and Zheltov S., 2017. Accuracy evaluation of structure from motion surface 3D reconstruction. Proc. SPIE 10332, Videometrics, Range Imaging, and Applications XIV, 103320P (2017)

Lichen Zhou., Chuang Zhang., Ming Wu., 2018.D-LinkNet: LinkNet with Pretrained Encoder and Dilated Convolution for High Resolution Satellite Imagery Road Extraction.Proc CVPR 2018(2018)

Fisher Yu., Vladlen Koltun., 2016. Multi-scale context aggregation dilated convolutions", Proc. ICLR 2016(2016)

Vizilter Yu., Gorbatsevich V., Melnichenko M., 2019. 3D Terrain Model Enhancing Using Generative Adversarial Network. Proc. Vol 11057, Modeling Aspects in Optical Metrology VII; 110571D (2019)

Wang J., Sun K., Cheng T., Jiang B., Deng C., Zhao Y., Liu D., Mu Y., Tan M., Wang X., Liu W., Xiao B., 2019. Deep High-Resolution Representation Learning for Visual Recognition. Proc. TPAMI 2019 (2019)