GENERATIVE ADVERSARIAL NETWORKS FOR SINGLE PHOTO 3D RECONSTRUCTION

V. V. Kniaz [1,2], F. Remondino [3], V.A. Knyaz [1,2]

[1] State Res. Institute of Aviation Systems (GosNIIAS), 125319, 7, Victorenko str., Moscow, Russia (vl.kniaz, knyaz)@gosniias.ru
[2] Moscow Institute of Physics and Technology (MIPT), Russia
[3] 3D Optical Metrology (3DOM) unit, Bruno Kessler Foundation (FBK), Trento, Italy - remondino@fbk.eu, http://3dom.fbk.eu

Commission II

KEYWORDS: generative adversarial networks, deep convolutional neural networks, cultural heritage, single image

ABSTRACT:

Fast but precise 3D reconstructions of cultural heritage scenes are becoming very requested in the archaeology and architecture. While modern multi-image 3D reconstruction approaches provide impressive results in terms of textured surface models, it is often the need to create a 3D model for which only a single photo (or few sparse) is available. This paper focuses on the single photo 3D reconstruction problem for lost cultural objects for which only a few images are remaining. We use image-to-voxel translation network (Z-GAN) as a starting point. Z-GAN network utilizes the skip connections in the generator network to transfer 2D features to a 3D voxel model effectively (Figure 1). Therefore, the network can generate voxel models of previously unseen objects using object silhouettes present on the input image and the knowledge obtained during a training stage. In order to train our Z-GAN network, we created a large dataset that includes aligned sets of images and corresponding voxel models of an ancient Greek temple. We evaluated the Z-GAN network for single photo reconstruction on complex structures like temples as well as on lost heritage still available in crowdsourced images. Comparison of the reconstruction results with state-of-the-art methods are also presented and commented.

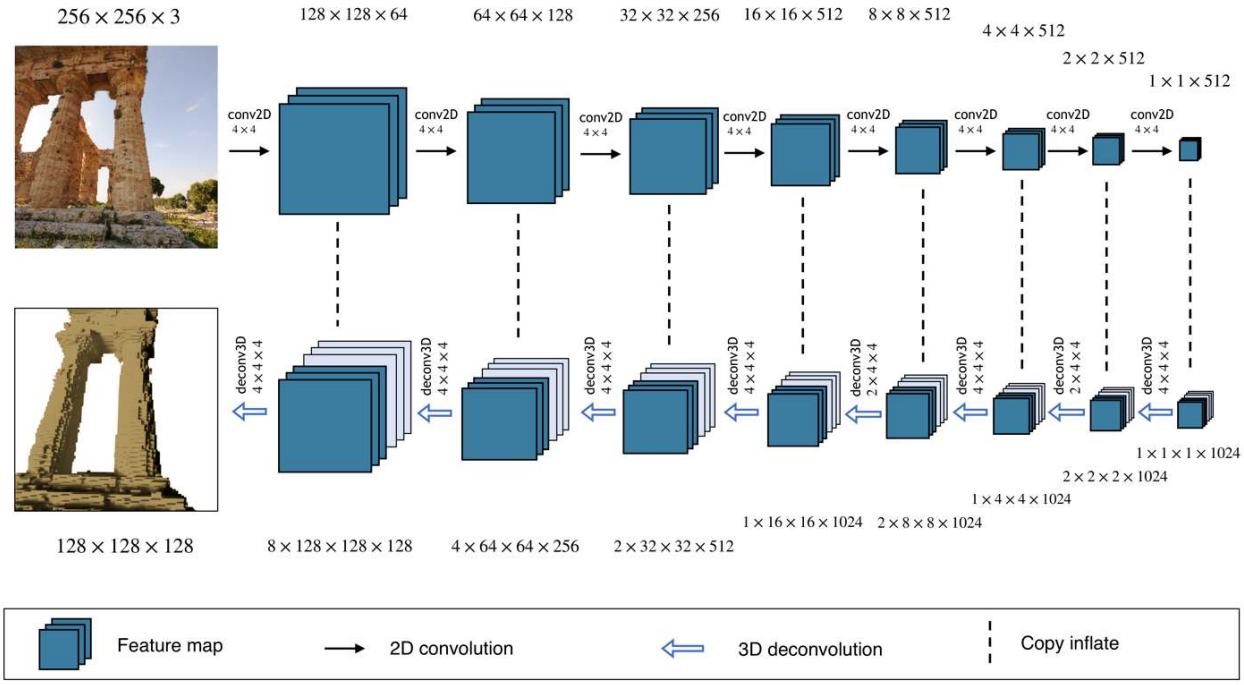

Figure 1: Overview of the Z-GAN generator network employed for 3D reconstruction from a single image.

1. INTRODUCTION

3D documentation and digital reconstruction of cultural heritage is an important application of photogrammetric methods. While modern multi-image algorithms provide reliable, fast and affordable solutions for the 3D reconstruction of an object seen in multiple images acquired from various viewpoints, they are not applicable if only a few or a single image are available. Recently various deep learning-based algorithms were proposed for single photo 3D object reconstruction (Huang et al., 2015; Choy et al., 2016; Wu et al., 2017; Richter and Roth, 2018). Such methods take a single image as an input and translate it to a low-resolution voxel model. However, to achieve high-quality results, a large training dataset is required. Moreover, most of the modern single image 3D reconstruction methods are trained using a fully supervised approach. Therefore, the resulting voxel models tend to be similar to the models present in the training dataset. To overcome the drawbacks of the fully supervised training, Generative Adversarial Networks (GANs - Goodfellow et al., 2014) were recently proposed.

This paper is focused on providing a reliable and efficient method, based on deep learning, for the 3D reconstruction of cultural heritage scenes from a single image. We use image-to-voxel translation network (Z-GAN - Kniaz et al., 2018a) as a

starting point for our method. Z-GAN network utilizes the skip connections in the generator network to transfer 2D features to a 3D voxel model effectively (Figure 1). Therefore, the network can generate voxel models of previously unseen objects using object silhouettes present on the input image and the knowledge obtained during the training stage. In order to train Z-GAN network (Section 4.3), we use a large dataset that includes oriented images and corresponding voxel models. We combine synthetic and real images to evaluate the network's ability to reconstruct previously unseen objects, considering also lost heritage objects seen in few and sparse crowdsourced images. We performed a qualitative and quantitative evaluation of the method (Section 4.4-4.5) using ground truth voxel models and the Intersection-over-Union (IoU) metric.

1.1 Contribution

Starting from the image-to-voxel translation network (Z-GAN) presented in Kniaz et al. (2018a), we present three key technical contributions: (1) an evaluation of the Generative Adversarial Networks (GAN) for single photo voxel model generation and 3D reconstruction purposes, (2) some heritage datasets with aligned color images, frustum voxel models and code available for research purposes and (3) the use of crowdsourced heritage images for the 3D reconstruction of lost monuments.

2. RELATED WORKS

2.1 Multi-view photogrammetric reconstruction

The multi-view image-based 3D reconstruction pipeline is nowadays based on the integration of photogrammetric and computer vision algorithms. It is composed of tie points extraction and image orientation for sparse point cloud generation (often called Structure from Motion – SfM) (Changchang et al., 2011; Heinly et al., 2015; Schoenberger and Frahm, 2016), dense image matching for dense point cloud generation (Remondino et al., 2014) and generation of final products like surface models, orthoimages, etc. It has become a powerful, automated, low-cost and valuable method for 3D scene reconstruction, documentation and modeling (Remondino et al., 2017), simply using a camera or even a smartphone (Nocerino et al., 2017). Online Cloud-based processing (Tefera et al., 2018) are also available to decouple the user from a powerful hardware that carries out the 3D processing. Multi-view 3D reconstructions could be coupled to deep learning methods to facilitate reconstructions in case of low-texture areas (Kniaz et al., 2018b) or for thermal images (Knyaz at al., 2017).

2.2 Single photo 3D reconstruction

Accurate 3D reconstruction is challenging if only a single image is available. This problem was always of great interest for the photogrammetric community (El-Hakim, 2001; Remondino and Roditakis, 2003; Remondino and El-Hakim, 2006). In the last years many new approaches for single image 3D reconstruction based on deep learning were proposed (Huang et al., 2015; Tatarchenko et al., 2015; Girdhar et al., 2016; Choy et al., 2016; Yan et al., 2016; Wu et al., 2016; Richter and Roth, 2018; Shin et al., 2018; Wu et al., 2017). While a number of methods were proposed for prediction of unobserved voxels from a single depth map (Zheng et al., 2013; Firman et al., 2016; Song et al., 2017; Yang et al., 2017; Yang et al., 2018), prediction of the voxel model of a complex scene from a single (color) image is more ambiguous and challenging. Prior knowledge of 3D shape is required for the robust performance of a single image method. Hence, most of the methods split the problem into two steps: (i) object recognition and (ii) 3D shape reconstruction. In Girdhar et al. (2016), a deep learning method for a single image voxel model reconstruction was proposed. The method leverages an auto-encoder architecture for a voxel model prediction. While the model has demonstrated promising results, the resolution of the voxel model was limited to 20x20x20 elements. An approach that combines single-view and multi-view reconstruction modes was proposed in Choy et al. (2016). In Richter and Roth (2018) a new voxel decoder architecture was proposed that leverages voxel tube and shape layers to increase the resulting voxel model resolution. A comparison of surface-based and volumetric 3D model prediction is reported in Shin et al. (2018).
Recently, 3D shape synthesis from a latent space has received a lot of attention (Brock et al., 2016; Girdhar et al., 2016; Wu et al., 2016). Wu et al. (2016) have proposed a GAN model for a voxel model generation (3D-GAN). The model was capable of predicting voxel models with resolution 64x64x64 from a randomly sampled noise vector. 3D-GAN was used for single image 3D reconstruction using an approach proposed in Girdhar et al. (2016). While 3D models produced by the 3D-GAN model provided more details compared to Girdhar et al., (2016), the generalization ability of the approach was insufficient to predict voxel models of previously unseen 3D shapes.

2.3 Generative adversarial networks

Generative Adversarial Networks (GANs - Goodfellow et al., 2014) provide a mapping from a random noise vector to a domain of the desired outputs (e.g., images, voxel models, etc.). GANs are gaining increasing attention in recent years. Indeed they provide encouraging results in tasks like image-to-image translation (Isola et al., 2017) and voxel model generation (Wu et al., 2016).

3. DEVELOPED METHODOLOGY

We use pix2pix (Isola et al., 2017) framework as a starting point to develop our Z-GAN model. We keep the encoder part of the generator unchanged. The 2D convolution kernels is changed into a 3D deconvolution kernels to encode a correlation between neighbor slices along the Z-axis. As proposed in the U-net model (Ronneberger et al., 2015), we keep the skip connections between the layers of the same depth. Indeed we believe that skip connections help to transfer high-frequency components of the input image to the high-frequency components of the 3D shape. The resulting architecture of our Z-GAN model is presented in Figure 2.

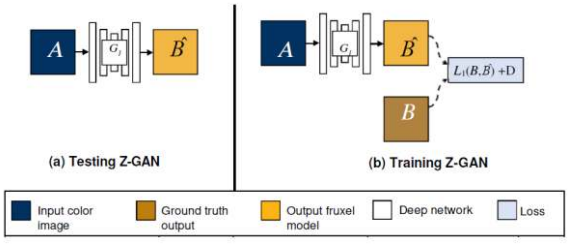

Figure 2: Z-GAN framework.

3.1 Fruxel model

A meaningful representation of data is required for the effective application of deep learning methods. For example, Lab color space provides a significant improvement in colorization performance (Zhang et al., 2016) compared to the RGB color model. Voxel models can be generated using two kinds of coordinate systems: view-centered and object-centered (Shin et al., 2018). For voxel models with object-centered coordinates, the

reference frame is aligned to some feature on the object. When a net is trained using voxel models with the object-centered coordinate system, it produces the same voxel output for any input image. On the other hand, the view-centered coordinate system is aligned to the camera optical axis. While the 3D model of the object remains the same, its view-centered voxel model changes with respect to the camera pose relative to the object. A special kind of view-centered voxel model was proposed recently (Kniaz et al., 2018a). The proposed voxel model was called "frustum voxel model" (fruxel model). It provides the following benefits:

- It offers 2D to 3D contour correspondences that are crucial for effective training of a generator network with skip connections (Ronneberger et al., 2015).
- It can be automatically transformed into three common 3D vision data formats. The frontmost non-empty elements of a fruxel models provide a depth map of an object. The sum of non-empty elements along the optical axis is equal to pixel level annotation of an object.
- Interpolation of a fruxel model to a new rectangular mesh provides for a common voxel model. We divide the camera frustum pyramid into equal slices to generate a fruxel model. While the real size of a slice changes with respect to the distance to the camera optical center, all slices have the same resolution.

A fruxel model is characterized by a following set of parameters:

$$\{z_n, z_f, d, \alpha\}$$

where

- z_n is a distance to a near clipping plane,
- z_f is a distance to a far clipping plane,
- d is the number of frustum slices,
- α is a field of view of a camera.

3.2 Z-net generator

The main idea of our volumetric generator G is to use the correspondence between silhouettes in a color image and slices of a fruxel model. We used the U-Net generator (Ronneberger et al., 2015) as a starting point to develop the model. The original U-Net generator leverages skip connections between convolutional and deconvolutional layers of the same depth to transfer fine details from the source to the target domain effectively.

Two modifications to the original U-Net model were realized: firstly, we replaced the 2D deconvolutional filters with 3D deconvolutional filters; secondly, we modified the skip connections to provide the correspondence between shapes of 2D and 3D features. The outputs of 2D convolutional filters in the left (encoder) side of Z-Net generator are $F_{2D} \in \mathbb{R}^{w \times h \times c}$ tensors, where w, h are the width and the height of a feature map and c is the number of channels. The output of 3D deconvolutional filters in the right (decoder) side are $F_{3D} \in \mathbb{R}^{w \times h \times d \times c}$ tensors. We use d copies of each channel of F_{2D} to fill the third dimension of F_{3D}. We term this operation as "copy inflate."

The architecture of the generator is presented in Figure 1.

4. EXPERIMENTS AND EVALUATION

4.1 Data collection

The methodology was tested using heritage datasets, in particular images of the following monuments and areas:

- Neptune temple (Paestum, Italy): some 680 terrestrial images, acquired with a Nikon D3X camera coupled with a 14mm focal length. The temple, approx. 24.5 x 60 m, consists of 6 frontal and 14 lateral Doric columns while the interior area has two rows of double ordered columns that divide the naos in three parts.
- Cerere temple (Paestum, Italy): some 212 terrestrial images acquired with a Nikon D3X camera coupled with a 14mm focal length. The temple, approx. 14.5 x 33 m, has a series of 6 x 13 Doric columns.
- Bosra archaeological area (Syria): it is a crowdsourced dataset of the UNESCO site in Syria, heavily damaged by the local war. It was collected within the Reckrei project activities https://projectmosul.org/). The disposal of such images, couples with the proposed Z-GAN method, could facilitate the digital reconstruction of lost heritage in case a single or few historical images are available.

For our tests, the Neptune temple was processed in order to retrieve the camera poses and generate a dense point cloud. These data are the base for our training. Cerere and Bosra images are used to prove the 3D reconstruction potential of Z-GAN.

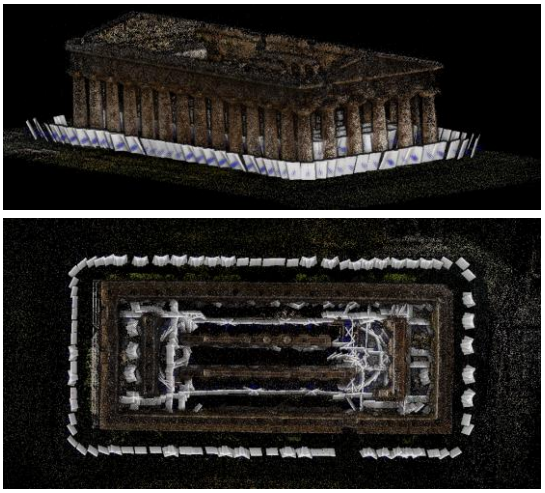

Figure 3: Camera poses and sparse point cloud for the Neptune image dataset.

4.2 Dataset generation

The fruxel model dataset generation uses Blender 3D Creation Suite and a dedicated Python script (Kniaz et al., 2018a). The dataset generation method consists of four steps:

- Generation of the undistorted images using the bundle adjustment results.
- Import of the available 3D models and camera positions into Blender.
- Generation of a slicing plane object, that moves normal to the camera's optical axis. A Boolean intersection operation to the slicing plane and the target object is applied to generate d keyframes for each camera position. For each keyframe i, we place the slicing plane on the distance $l = z_n + i \cdot t$, where $t = \frac{z_f - z_n}{d}$ is the thickness of the slicing plane.
- Rendering of all keyframes and combine them to arrays $F_{3D} \in \mathbb{R}^{d \times d \times d}$. We assign a black color to the background and white emissive material to the object. Therefore, in each rendered frame we have a pixel level labelling of the object slice at the distance l from the camera.

4.3 Network training

Our Z-GAN framework was trained on the Neptune dataset using PyTorch library (Paszke et al., 2017). We used 600 images (out of 680) and the corresponding fruxel models with parameters $\{z_n = 2, z_f = 12, d = 128, \alpha = 103^\circ\}$. Our Z-GAN model predicts fruxel

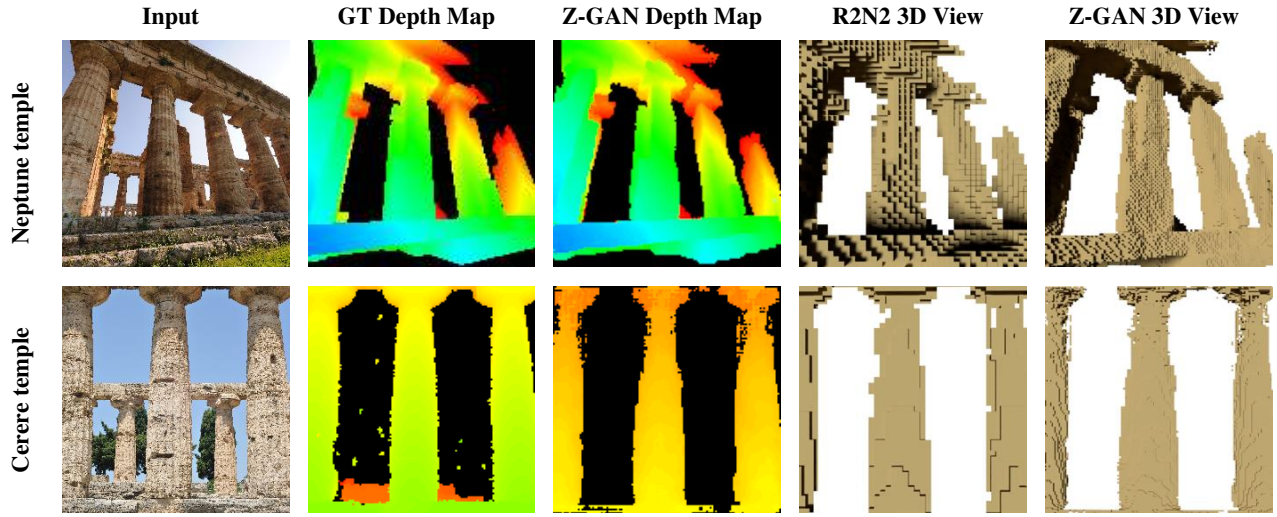

Figure 4: Results of 3D reconstruction using Z-GAN on the Neptune and Cerere temple datasets with respect to Ground Truth (GT) depth map and the state-of-the-art method R2N2.

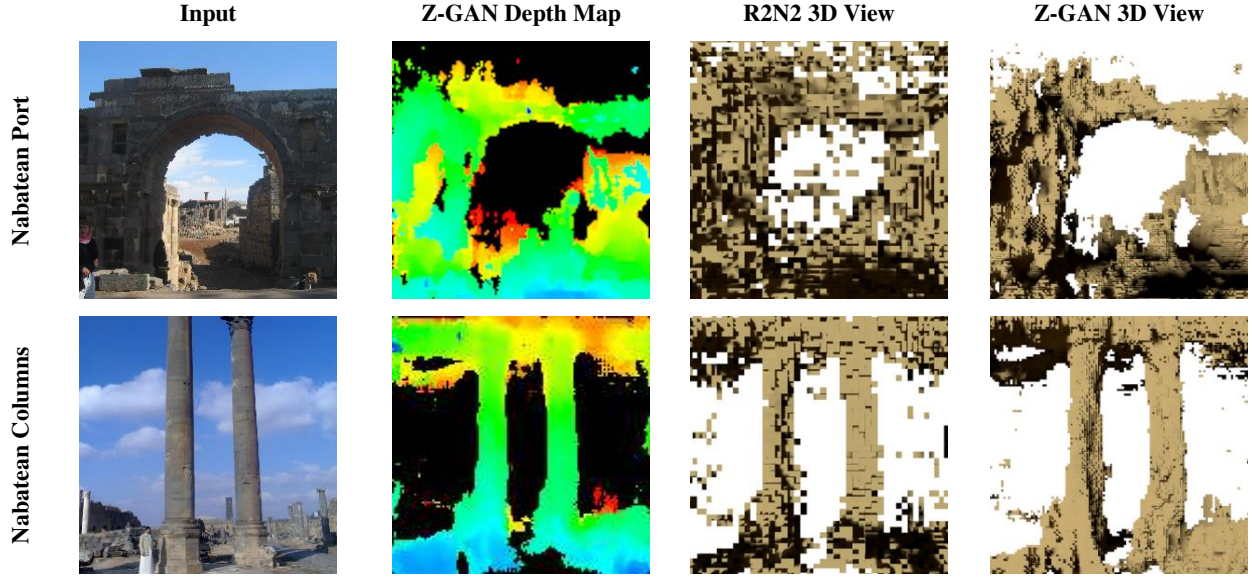

Figure 5: Results of 3D reconstruction using Z-GAN on the crowdsourcing dataset (Borsa, Syria - courtesy of Rekrei project) with respect to a state-of-the-art method R2N2.

models with a resolution 128×128×128 elements (due to GPU card limitations). We used the remaining 80 images for evaluation. The training was performed using the NVIDIA 1080 Ti GPU and took 6 hours for the generator and the discriminator networks. For network optimization, we use minibatch SGD with an Adam solver. We set the learning rate to 0.0002 with momentum parameters $\beta1 = 0.5$, $\beta2 = 0.999$, similar to Isola et al. (2017).

4.4 Qualitative evaluation

We evaluate the trained network qualitatively using two datasets: Cerere temple and crowdsourced images of the Bosra roman amphitheater. The Cerere dataset includes 212 images and fruxel models with parameters $\{z_n = 2, z_f = 12, d = 128, \alpha = 38^\circ\}$. For the Cerere dataset we compared the Ground Truth (GT) voxel models with the network reconstruction (Figure 4).

The Bosra dataset was collected using crowdsourced images with unknown camera internal orientation parameters. Hence, for the Bosra dataset, we did not have the ground truth fruxel models. To perform the evaluation, we presented the reconstruction results to an expert, who verified the structure of the reconstructed monument. The overall evaluation results in terms of expert score are given in Table 1. Expert evaluation proves that the Z-GAN is capable of reconstruction of complex buildings with features that were not present in the training dataset. For example, there were no arcs in the training dataset. Nevertheless, the Z-GAN model was able to reconstruct buildings with arcs (Figure 5).

Score / *Object*	R2N2	Z-GAN
Cerere temple	0.54	**0.65**
Nabatean port	0.18	**0.51**
Nabatean pillars	0.43	**0.60**

Table 1: Expert score for the Cerere temple and two objects from the crowdsourcing dataset.

4.5 Quantitative evaluation

We use the Cerere temple dataset for the quantitative evaluation of the presented Z-GAN method. Results are given in Table 2 in

terms of Intersection over Union and Surface Distance metrics (Rock et al., 2015).
Comparison of results for 3D-R2N2 and our Z-GAN models proves that our model outperforms the 3D-R2N2 in IoU accuracy and Surface Distance. While 3D-R2N2 was quite suitable to reconstruct the Cerere temple, its performance became very unstable for the crowdsourced dataset with images very different from the training dataset.

Network / *Object*	R2N2		Z-GAN	
	IoU	**Surface Distance**	**IoU**	**Surface Distance**
Columns	0.59	0.151	0.69	0.102
Basement	0.65	0.231	0.87	0.145
Average	0.62	0.191	**0.78**	**0.124**

Table 2: The IoU metric and surface distance for R2N2 and Z-GAN networks on two objects from the Cerere temple dataset.

4.6 Application

In case of a building or monument seen in some sparse images (typical case of crowdsourced images of a lost heritage), the Z-GAN model can be use to reconstruct the object as seen in all available images. Indeed, if the images are adjacent (but not necessarily overlapping), a large part of the object could be digitally reconstructed by merging the single fruxel models.
We tested the idea on the Cerere temple with a twofold approach: firstly, we reconstruct the separate pieces of the temple seen in the single images using the trained network; secondly, we join the resulting fruxel models using an ICP algorithm registration (in case the images very partly overlapping, we could also use the external orientation parameters estimated within a bundle adjustment). To perform the ICP registration, we consider each nonempty fruxel element as a 3D point. The final 3D model of the temple from three images (with very few overlap) is shown in Figure 6.

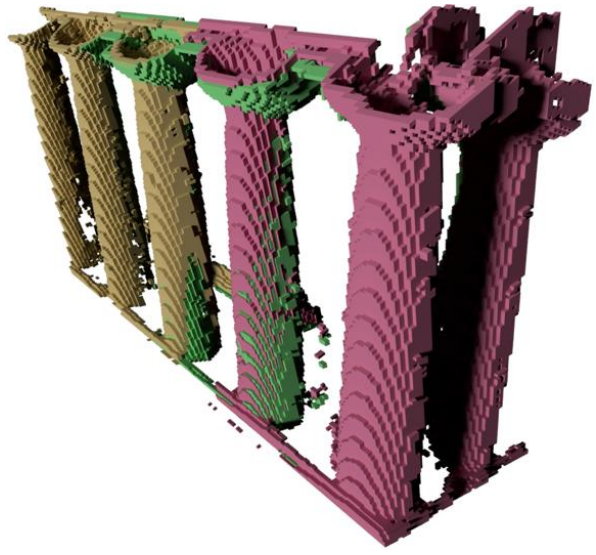

Figure 6: Reconstruction of a part of the Cerere temple using three adjacent images.

5. CONCLUSIONS

The paper presented a single image 3D reconstruction approach based on a deep learning method: the Z-GAN model. Results show that generative adversarial networks for the generation of 3D information from a single image can be used efficiently in case of historical buildings and archaeological monuments. The employed network can generalize to previously unseen data even if the monument in the training dataset does not share much features with the test images. We presented the predicted voxel models and the input images to a human expert for the qualitative evaluation. The expert had compared the network reconstruction with a manual reconstruction created using his experience in the field of archaeology. The overall qualitative score assigned by the expert proved that the Z-GAN network is capable of a single photo realistic reconstruction of lost cultural objects.
The fruxel model 3D object representation proves to be efficient in deep learning applications. Fruxel models like voxel models provide a discrete representation and 2D to 3D contour alignment and this facilitates 3D model reconstruction.
While the generalization ability of Z-GAN network is encouraging, changes in the camera focal length reduce the reconstruction accuracy. The network fails to estimate the distance to the object if the focal length differs for images in the training data and the test image. We estimated the dependency between the difference of the focal lengths for training and test images and depth reconstruction accuracy. The dependency can be linearized for correction of depth estimates if the focal length for the test image is known. We hypothesize that camera focal length can be used as a network input to improve the depth estimation accuracy.
As future works, we will employ higher resolution fruxel model (e.g. 256x256x256) exploiting better GPU card in order to have more smoothed 3D reconstructions. We will also enlarge the network in order to include other types of structures and being able to process other single images from crowdsourcing datasets of lost heritage. The dataset used so far is available at goo.gl/U5C2Wh, whereas the pretrained model and code of Z-GAN is available at: https://github.com/vlkniaz/Z_GAN.

ACKNOWLEDGEMENTS

The work was performed with the support by Grants №17-29-04509 and №17-29-04410 of Russian Foundation for Basic Research (RFBR). Authors are thankful to the Reckrei project (https://projectmosul.org/) for providing the crowdsourced heritage images used in the presented tests.

REFERENCES

Brock, A., Lim, T., Ritchie, J. and Weston, N., 2016. Generative and discriminative voxel modeling with convolutional neural networks. Proc. Neural Information Processing Conference: 3D Deep Learning- NIPS, pp. 1-9.

Changchang, W., Agarwal, S., Curless, B., Seitz, S.M., 2011. Multicore bundle adjustment. Proc. CVPR.

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

El-Hakim, S., 2001. A flexible approach to 3D reconstruction from single images. Proc. ACM SIGGRAPH, Vol. 1, pp. 12-17.

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

Girdhar, R., Fouhey, D. F., Rodriguez, M., Gupta, A., 2016. Learning a predictable and generative vector representation for objects. Proc. ECCV pp. 702-722.

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

Heinly, J., Schonberger, J.L., Dunn, E., Frahm, J.-M., 2015. Reconstructing the World* in six days *(as captured by the Yahoo 100 million image dataset). Proc. CVPR.

Huang, Q., Wang, H. and Koltun, V., 2015. Single-view reconstruction via joint analysis of image and shape collections. ACM Transactions on Graphics, Vol. 34(4), pp. 87:1-87:10.

Isola, P., Zhu, J.-Y., Zhou, T. and Efros, A. A., 2017. Image-to-Image translation with conditional adversarial networks. Proc. CVPR, pp. 5967-5976.

Knyaz, V. A., Vygolov, O. V., Kniaz, V. V., Vizilter, Y. V., Gorbatsevich, V. S., Luhmann, T. and Conen, N., 2017. Deep learning of convolutional auto-encoder for image matching and 3D object reconstruction in the nfrared range. Proc. CVPR, pp. 2155-2164.

Kniaz, V. V., Knyaz, V. A., Remondino, F., 2018a. Image-to-voxel model translation with conditional adversarial networks. Proc. ECCV.

Kniaz, V. V., Fedorenko, V. V., and Fomin, N. A., 2018b. Deep Learning for low-textured image matching. ISPRS Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., Vol. XLII-2, pp. 513-518

Nocerino, E., Poiesi, F., Locher, A., Tefera, Y., Remondino, F., Chippendale, P., Van Gool, L., 2017. 3D reconstruction with a collaborative approach based on smarthphones and a cloud-based server. ISPRS Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., Vol. XLII-2/W8, pp. 187-194.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L. and Lerer, A., 2017. Automatic differentiation in pytorch. Proc. NIPS

Remondino, F. and Roditakis, A., 2003. Human figure reconstruction and modeling from single image or monocular video sequence. Proc. 4th IEEE International Conference on 3-D Digital Imaging and Modeling (3DIM), pp. 116-123.

Remondino, F. and El-Hakim, S., 2006. Image-based 3D Modelling: A Review. The Photogrammetric Record, Vol. 21(115), pp. 269-291.

Remondino, F., Spera, M.G., Nocerino, E., Menna, F., Nex, F., 2014: State of the art in high density image matching. The Photogrammetric Record, Vol. 29(146), pp. 144-166.

Remondino, F., Nocerino, E., Toschi, I., Menna, F., 2017. A critical review of automated photogrammetric processing of large datasets. ISPRS Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., Vol. XLII-2/W5, pp. 591-599.

Richter, S. R. and Roth, S., 2018. Matryoshka Networks: Predicting 3D geometry via nested shape layers. Proc. CVPR.

Ronneberger, O., Fischer, P. and Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. Proc. Inter. Conference on Medical Image Computing and Computer-assisted Intervention, Springer, pp. 234-241.

Schoenberger, J.-L., Frahm, J.-M., 2016. Structure-from-motion revisited. Proc. CVPR.

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

Song, S., Yu, F., Zeng, A., Chang, A. X., Savva, M. and Funkhouser, T., 2017. Semantic scene completion from a single depth image. Proc. CVPR.

Tatarchenko, M., Dosovitskiy, A. and Brox, T., 2015. Multi-view 3D models from single images with a Convolutional Network. arXiv.org.

Tefera, Y., Poiesi, F., Morabito, D., Remondino, F., Nocerino, E., Chippendale, P., 2018. 3DNOW: image-based 3D reconstruction and modeling via web. ISPRS Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., Vol. XLII-2, pp. 1097-1103.

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. Proc. NIPS, pp. 82-90.

Wu, J., Wang, Y., Xue, T., Sun, X., Freeman, W. T. and Tenenbaum, J. B., 2017. MarrNet: 3D shape reconstruction via 2.5D sketches. arXiv.org.

Yan, X., Yang, J., Yumer, E., Guo, Y., Lee, H., 2016. Perspective transformer nets: Learning single-view 3D object reconstruction without 3D supervision. Proc. NIPS.

Yang, B., Wen, H., Wang, S., Clark, R., Markham, A., Trigoni, N., 2017. 3D object reconstruction from a single depth view with adversarial learning. Proc. ICCV.

Yang, B., Rosa, S., Markham, A., Trigoni, N., Wen, H., 2018. 3D object dense reconstruction from a single depth view. TPAMI, DOI: 10.1109/TPAMI.2018.2868195.

Zheng, B., Zhao, Y., Yu, J. C., Ikeuchi, K., Zhu, S.-C., 2013. Beyond point clouds: scene understanding by reasoning geometry and physics. Proc. CVPR.