THERMAL TEXTURE GENERATION AND 3D MODEL RECONSTRUCTION USING SFM AND GAN

V. V. Kniaz[a,b], V. A. Mizginov[a*]

[a] State Res. Institute of Aviation Systems (GosNIIAS), 125319, 7, Victorenko str., Moscow, Russia
(vl.kniaz, vl.mizginov)@gosniias.ru
[b] Moscow Institute of Physics and Technology (MIPT), Russia

Commission II, WG II/5

KEY WORDS: infrared images, structure from motion, generative adversarial networks, object recognition, deep convolutional neural networks

ABSTRACT:

Realistic 3D models with textures representing thermal emission of the object are widely used in such fields as dynamic scene analysis, autonomous driving, and video surveillance. Structure from Motion (SfM) methods provide a robust approach for the generation of textured 3D models in the visible range. Still, automatic generation of 3D models from the infrared imagery is challenging due to an absence of the feature points and low sensor resolution. Recent advances in Generative Adversarial Networks (GAN) have proved that they can perform complex image-to-image transformations such as a transformation of day to night and generation of imagery in a different spectral range. In this paper, we propose a novel method for generation of realistic 3D models with thermal textures using the SfM pipeline and GAN. The proposed method uses visible range images as an input. The images are processed in two ways. Firstly, they are used for point matching and dense point cloud generation. Secondly, the images are fed into a GAN that performs the transformation from the visible range to the thermal range. We evaluate the proposed method using real infrared imagery captured with a FLIR ONE PRO camera. We generated a dataset with 2000 pairs of real images captured in thermal and visible range. The dataset is used to train the GAN network and to generate 3D models using SfM. The evaluation of the generated 3D models and infrared textures proved that they are similar to the ground truth model in both thermal emissivity and geometrical shape.

1. INTRODUCTION

Thermal emission of objects captured by an infrared camera provides a whole new way for scene analysis. A large training dataset is required to develop and train effective algorithms for processing thermal images. In contrast with the visible range, where a great number of datasets are publicly available (Geiger et al., 2013, Nex et al., 2015, Menze and Geiger, 2015), only a limited number of datasets with infrared imagery can be found to date. Recently, 3D object models with realistic textures have become one of the main instruments for the creation of extensive image datasets with accurate ground truth annotations. Many state-of-the-art datasets have been created using 3D modeling in such fields of photogrammetry and computer vision as optical flow estimation (Wulff et al., 2012), autonomous driving (Hosseinyalamdary and Yilmaz, 2015, Menze and Geiger, 2015) and camera external orientation estimation (Kehl et al., 2016, Kluger et al., 2017). Structure from Motion (SfM) algorithms provide a fast and robust approach for a textured 3D model generation. Still, for most objects, a direct 3D reconstruction using SfM and infrared range images is challenging (Hajebi and Zelek, 2008, Yamaguchi et al., 2017).

Recently Generative Adversarial Networks (GAN) have shown significant success in the arbitrary image-to-image transform problems such as season change (Zhu et al., 2017), object transfigurations (Isola et al., 2017) and image colorization (Zhang et al., 2016). Most of the issues listed above have a unimodal nature. In other words, for a given image in the source domain, there is the only single possible correct solution in the target domain.

*Corresponding author

The problem of transformation of visible image to the thermal image has a highly multimodal nature. For example, for a given picture of a car in the visible range, there can be infinite possible infrared images conditioned by a sequence of events that have occurred with the car. If it were left in a parking place for a long time, there would be no significant contrast between the car and the background. On the other hand, the same car after a long ride will have hot wheels and a hot bonnet due to brake friction and thermal energy radiated by the engine.

It is impossible to guess the real temperature of a car from a visible range image. However, the random correct rendering can be guessed. Therefore, it is only required to avoid impossible thermal renderings of a visible range image (e.g., a car with a cold bonnet and a hot roof). In such setting, the visible range-to-thermal range transformation can be considered as a general multimodal image-to-image transformation. Such problem statement fulfills requirements of an important application of such transformation: augmentation of large existing image datasets with synthetic thermal imagery. The new generation of GAN had overcome difficulties of the multimodal image-to-image transformations and had shown impressive results for such problems as sketch colorization, and map-to-satellite image transformation. In spite of this, realistic thermal imagery synthesis remains a challenging problem as a trained GAN tends to transfer features that exist only in the visible range (color patterns, reflections on affront glass) to a thermal range. Moreover, most of GANs fail to reconstruct the location of the source of thermal radiation (i.e., an engine) correctly.

In this paper, we propose a novel method for generation of realistic 3D models with thermal textures using the SfM pipeline and

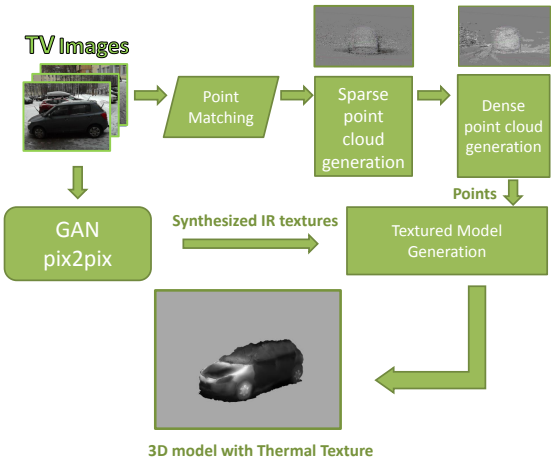


Figure 1. The proposed pipeline for generation of a 3D model with realistic infrared texture.

GAN. The proposed pipeline is shown in figure 1. We present a novel iterative training method for a multimodal GAN that overcomes most of the problems of thermal image synthesis. In contrast with a traditional GAN training pipeline, where the discriminator network receiver only positive ground truth samples, we propose to use both positive and negative ground truth samples. Negative ground truth samples for the training dataset are obtained by manual selection of unrealistic thermal images that were created on the previous iteration. We collected a large dataset of thermal images with 1000 images of five object classes. The dataset contains paired images in the visible and the infrared range. We train various GANs for image transformation on the dataset using the classical and the proposed pipelines, to show that the proposed method outperforms other approaches for thermal image synthesis using GAN.

The rest of the paper is organized as follows. In the second section, we give an outline of GANs for image synthesis and present other approaches for the thermal image synthesis. In the third section, we present the pipeline of the proposed training method. The architectures of the evaluated GANs are briefly described. We discuss selected objects and image resolution as well, as the structure of the collected dataset. In the fourth section, we show the results of GAN training. We compare the traditional and the proposed pipeline using various metrics. The fifth section concludes the paper with the summary of the achieved results. We briefly discuss the proposed approach and the prospects for future work.

2. RELATED WORK

It is always challenging to visualize the scene that is not perceptible by human eyes. The thermal image synthesis has been long studied by the computer vision society. The approaches developed to date can be broadly divided into three large groups. The first group is based on a 3D modeling. The thermal image is generated using the direct numerical calculation of temperatures of objects present in the scene. After the calculation, the scene is visualized using the computer graphics pipeline. Numerical 3D modeling had received a lot of scholar attention in the 1980s after the invention of thermal cameras. The main disadvantage of this approach is a high numerical complexity. Also, most of the numerical thermal image models do not provide a realistic noise and thermal reflections.

Another large group of approaches is based on a 3D modeling with real infrared textures. Base 3D models are either generated manually or reconstructed using the SfM technique. Real infrared textures provide a significant boost in the quality of the generated image and, hence, the precision of the algorithms trained using the generated images. The main drawback of the 3D modeling with real textures is the small number of infrared textures available in the public domain. The generation of thermal textures for all required objects can be costly. The projection of thermal textures back to a 3D model can also pose some problems as it is often hard to find the corresponding points in the visible range texture and the infrared texture. The last advances in the direct SfM methods for the thermal range imagery (Knyaz et al., 2017) provide a promising approach for the generation of textured infrared 3D models.

The recently invented GANs have shown impressive results in arbitrary image-to-image transforms (Goodfellow et al., 2014, Zhu et al., 2017, Isola et al., 2017). GAN consists of two deep convolutional neural networks: a generator network tries to synthesize an image that is visually indistinguishable from a given sample of images in a target domain. A discriminator network tries to distinguish the “fake” images $\hat{\mathbf{Y}}$ generated by the generator network from the real image in the target domain $\mathbf{Y}$. Both Generator and Discriminator networks are trained simultaneously. Such approach can be considered as an antagonistic game of two players.

3. METHODS

3.1 GAN architecture

The proposed ThermalGAN network is based on the `pix2pix` framework (Isola et al., 2017). The `pix2pix` framework was designed to perform an arbitrary image-to-image transformation. The framework consists of two deep convolutional networks: a generator network is a modified version of the U-Net (Ronneberger et al., 2015); a discriminator network is based on PatchGAN classifier (Li and Wand, 2016). The generator consists of 12 convolutional layers connected in two ways. Firstly, the output of each layer is coupled with the input of the next layer. Secondly, the output of the first layer is concatenated with the input of the last layer (the output of layer 11). Such feed-forward connections increase the generator’s performance for restoration of small details and increase learning convergence. The resulting generator network architecture is presented in figure 2.

In other words, the U-Net is similar to a convolutional auto-encoder with feedforward connections between the convolutional layers of the same depth. We have made two contributions to the original `pix2pix` framework: (1) the modified version of the U-Net architecture that takes a three channel color image of 256×256 pixels as an input and produces a single channel infrared image of the same size, (2) a new loss function to convert the GAN learning from antagonistic game of 2 players to a game of 3 players.

3.2 Objective function

Given an input color image $\mathbf{X} \in \mathbb{R}^{H \times W \times 3}$, our objective is to learn a mapping $\hat{\mathbf{Y}} = \mathcal{G}(\mathbf{X})$ to thermal emission $\mathbf{Y} \in \mathbb{R}^{H \times W}$, where H, W are image dimensions (Kniaz et al., 2017).

The traditional GAN loss function is designed to provide an antagonistic game of 2 players: a “forger” (generator) and a “policeman” (discriminator). The “forger” is trained to produce a

realistic output $\hat{\mathbf{Y}}$ similar to the given input domain $\mathbf{Y}$. The “policeman” is trained to distinguish fake images $\hat{\mathbf{Y}}$ from the real ones of $\mathbf{Y}$. We have tried to train the `pix2pix` framework in this setting and found out that the discriminator fails to distinguish the fake infrared image from the real one even in such cases when the fake is obvious for a human.

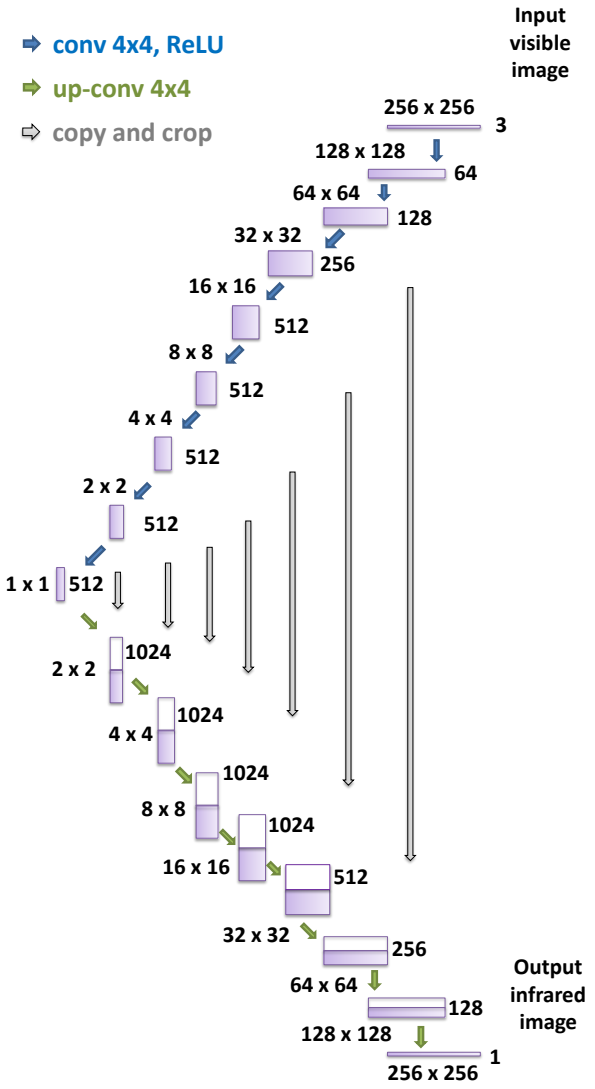


Figure 2. Generator network architecture.

We have modified the loss function, to transfer knowledge that is required to distinguish the unrealistic fake infrared image from the real one effectively. We have added one extra “player” to the game. The player represents the “expert” that provides the policeman with true fake examples. The true fake examples are collected from the previous iteration of GAN training. Hence, the discriminator network has to operate with two kinds of fake images: static fake images $\mathbf{S}$ that were produced by the generator during the previous iteration, and dynamic fake images $\hat{\mathbf{Y}}$ produced by the generator. In such setting the discriminator trains to distinguish fake images much faster. Hence, the generator has to increase the quality of fake images $\hat{\mathbf{Y}}$ to avoid errors that have happened during the previous iteration. The proposed training process is given in the algorithm 1. The modified loss function L_{GAN3} is based on the conditional GAN loss $L_{cGAN}(\mathcal{G},\mathcal{D})$ (Isola et al., 2017) and defined as

$$\begin{aligned} L_{GAN3}(\mathcal{G},\mathcal{D}) =& \mathbb{E}_{\mathbf{X},\mathbf{Y}}[\log \mathcal{D}(\mathbf{X},\mathbf{Y})]+ \\ & \mathbb{E}_{\mathbf{X},\mathbf{Z}}[\log(1-\mathcal{D}(\mathbf{X},\mathcal{G}(\mathbf{X},\mathbf{Z}))]+ \\ & \mathbb{E}_{\mathbf{X},\hat{\mathbf{Y}}}[\log(1-\mathcal{D}(\mathbf{X},\hat{\mathbf{Y}})], \end{aligned} \tag{1}$$

where $\mathbf{Z}$ – is a random noise vector, that is used to avoid the deterministic output of the generator.

Algorithm 1: Proposed training process

1. $i = 0$.
2. Train the generator $\mathcal{G}_i$ and the discriminator $\mathcal{D}_i$ using no true negative images $\mathbf{S}_i = \hat{\mathbf{Y}}$.
3. **for** $i \neq N_{max}$: **do**
4. Using $\mathcal{G}_i$ transfer the dataset $\mathbf{X}$ to infrared images $\hat{\mathbf{Y}}$.
5. Using $\mathcal{D}_i$ find such images in $\mathbf{Y}$ that are classified as real images. The set of such images are assigned to $\mathbf{S}_{i+1} = \mathbf{S}_i + \mathbf{Y}$
6. $i = i + 1$.
7. Reset the weights of generator and discriminator and train them again using true positives from $\mathbf{X}$ and false positive from $\mathbf{S}_i$.
8. **end**

3.3 Dataset

We evaluate the proposed training method using a specially designed dataset. The dataset includes pairs of geometrically aligned images of the visible and the infrared range. Images provide samples of five classes: person, cat, dog, car, building. All data was collected using the FLIR ONE PRO thermal camera. The detailed technical specifications of the camera are presented in Table 1.

Parameter	Value
Visible range resolution	1440×1080
Infrared resolution	160×120
Field of view	43°×55°
Temperature range	-20…400 °C
Spectral range	8 – 14 μm
Pixel size	12 μm

Table 1. FLIR ONE PRO camera specification

The images were scaled and cropped to square pictures 256×256 pixels to match the resolution of the generator’s input. We intentionally captured all classes in similar conditions to provide a uniform thermal contrast between the background and the object. Such approach provides semi unimodal distribution of $\mathbf{X}$ and $\mathbf{Y}$.

The collected data was divided into independent training and test datasets. The size of the training dataset is 200 images per class (1000 image). The test dataset provides 20 images per class.

The FLIR ONE PRO camera provides the calibrated 16-bit thermal image with real temperature values. In this paper, we projected 16-bit images to 8-bit and discarded the absolute temperature value. Hence, the generator is trained to reconstruct only the relative thermal contrast between the object and the background. Examples from the dataset are presented in figure 3.

4. EVALUATION

We used the generated dataset to evaluate the three players GAN training. We evaluated the networks in two ways. Firstly, we used the test part of the dataset with the ground truth thermal images. We used root mean square (RMS) error between the real thermal images $\mathbf{Y}$ and the synthesized output $\hat{\mathbf{Y}}$. Secondly, we evaluate the generator's ability to generalize from the training dataset using the PASCAL VOC 2012 dataset. The following section presents the details of the training as well as the evaluation of the synthesized thermal image quality.

Figure 3. Examples from the training dataset

4.1 3D model generation

The first step of the proposed method is the generation of a sparse point cloud using the images captured in the visible range. We evaluate 3D model generation using two test objects of the class car: Škoda Fabia and Citroën C3. The ground truth 3D models of real objects were generated using fringe projection scanner (Knyaz, 2010). The source images of the visible range were captured using the FLIR ONE camera in all-around configuration. The images were processed using the AgiSoft Photoscan software to generate point pairs, a sparse point cloud, and a dense point cloud. The resulting point cloud is presented in figure 4.

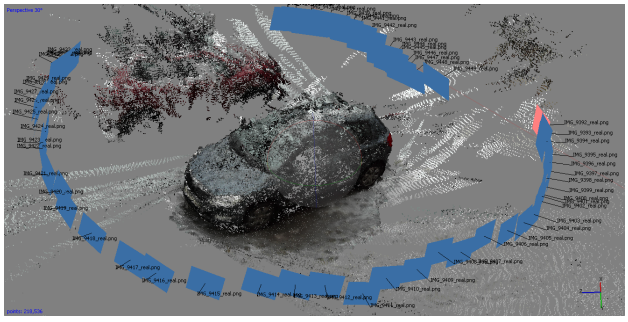


Figure 4. Dense point cloud with the visible texture

We evaluated the generated 3D model using the methodology proposed in (Remondino et al., 2014) to measure the accuracy of the generated surface. We use a 3D model of a car obtained using a fringe projection scanner (Knyaz, 2010) as the ground truth. The final accuracy of the reconstructed model in the object space was 9 cm for the Škoda model and 12 cm for the Citroën model. The distance between the ground truth model and the SfM reconstruction is presented in pseudo-color in figure 5.

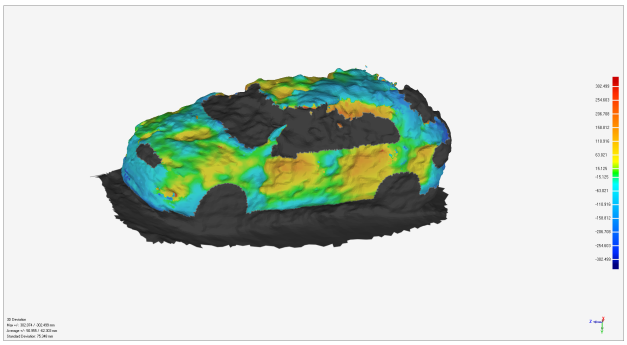


Figure 5. The distance between the ground truth model and the SfM reconstruction for the test object Škoda Fabia

4.2 GAN training

The U-net generator was trained using two player and three player methods to produce the infrared images. To train the GAN, we used the PyTorch framework (Ketkar, 2017). We have modified the original `pix2pix` framework to add the third player to the training process. To train the GAN we used NVIDIA 1080 GPU. The training dataset included 1000 pairs of images of 5 object classes. The training was completed in 4 hours and 200 epochs.

We term a single training process of `pix2pix` with 200 completed epochs as a single iteration of the three player GAN training process. After the first iteration, the whole training dataset is fed to the generator to obtain the true negative examples $\hat{\mathbf{Y}}$. These examples are used in the next iteration of training. The training of the GAN using the developed method was completed in 10 iterations (40 hours and 2000 epochs in total).

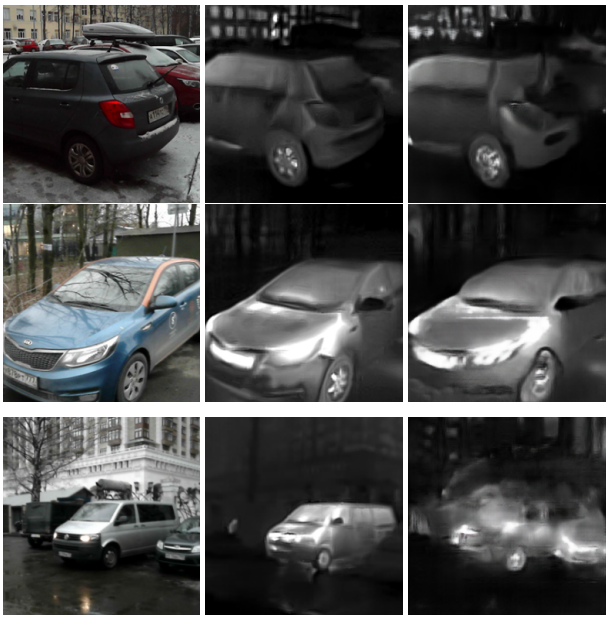
4.3 GAN evaluation

We used the independent test dataset to evaluate the GAN and measure the reconstruction error. We calculated the average standard deviation of a difference of a real infrared image and a synthesized infrared image in analog-to-digital units (ADU). The final thermal reconstruction errors are presented in Table 2. The results of the reconstruction are presented in figure 6.

To evaluate the generalization ability of the trained generator network we have performed generation of synthetic infrared images on samples from PASCAL VOC 2012 dataset. Since no ground truth infrared images are available for this dataset, we can only evaluate the result by visual inspection of synthesized images. Some examples of the reconstruction are presented in figure 7.

Method	pix2pix	Ours
Car	18.191	14.740
House	24.691	20.384
Human	25.744	23.143
Cat	51.033	50.328
Dog	55.221	53.143

Table 2. Standard deviation of the difference between the real infrared image and the synthesized infrared image in ADU



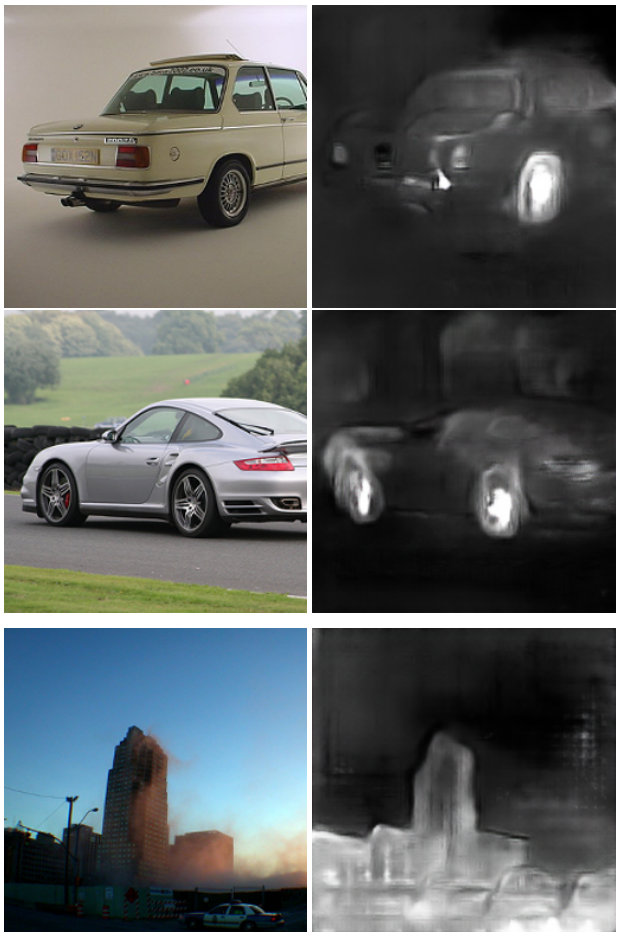
(a) Visible range (b) Infrared (c) GAN output IR

Figure 6. Examples of generated images

4.4 Infrared textures generation

We fed the visible range images into the generator network to create infrared textures for the generated 3D model. The resulting 3D model is presented in figure 8. To evaluate the accuracy of the generated textures, we rendered the output 3D model using camera external and internal orientation parameters estimated by the Agisoft Photoscan. Using this technique we obtain synthesized infrared images geometrically aligned with the ground truth infrared images from the FLIR ONE PRO camera. We use an RMS error between brightness values of synthesized images and the ground truth infrared images to evaluate the quality of synthesized textures.

To render realistic infrared images, we used Blender 3D creation suite. We imported 3D models and camera parameters from Agisoft PhotoScan to Blender and applied the textures with an emissive shader. To avoid bias in the RMS error caused by the background reconstruction quality we calculate the RMS error only over the object region. The resulting RMS error for the Škoda model was 8 ADU, and the error for the Citroën model was 6 ADU.



(a) Visible range (b) GAN output IR

Figure 7. Examples of generated images

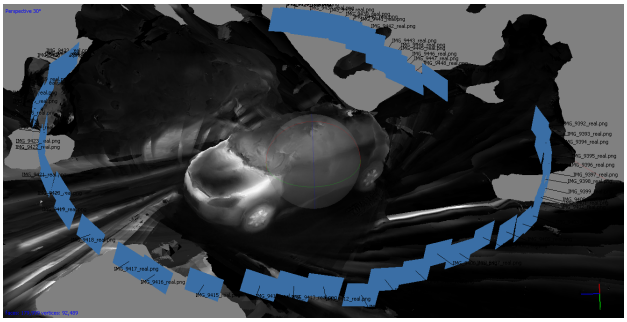


Figure 8. The resulting 3D model with the infrared texture

5. CONCLUSION

A new technique for generation of realistic 3D models with synthesized thermal textures was developed. The technique uses Structure from Motion for generation of realistic 3D models from the visible range imagery. The thermal textures are generated using Generative Adversarial Network. A modified `pix2pix` framework was used to train the GAN for a transformation from visible range images to thermal range images.

To overcome the difficulties of training GAN for a spectral range transformation a new training method was developed. The method

extends the traditional GAN training pipeline from the antagonistic game of two players to the game of three players. The third player represents an “expert” that provides the true negative samples to the discriminator network.

An extensive training dataset was generated using the FLIR ONE PRO infrared camera to evaluate the proposed method. The dataset includes 2000 pairs of geometrically aligned images pairs of visible and infrared range. Images include samples of five object classes: person, cat, dog, car, building.

The proposed method was implemented using PyTorch library for GAN training and AgiSoft Photoscan software for a 3D model reconstruction. The evaluation of the generated 3D models and infrared textures proved that they are similar to the ground truth model in both thermal emissivity and geometrical shape.

ACKNOWLEDGEMENTS

The reported study was funded by Russian Foundation for Basic Research (RFBR) according to the research project N^{o} 16-08-01260 and by Russian Science Foundation (RSF) according to the research project N^{o} 16-11-00082.

REFERENCES

Geiger, A., Lenz, P., Stiller, C. and Urtasun, R., 2013. Vision meets robotics - The KITTI dataset. *I. J. Robotics Res.* 32(11), pp. 1231–1237.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C. and Bengio, Y., 2014. Generative Adversarial Networks. *CoRR*.

Hajebi, K. and Zelek, J. S., 2008. Structure from Infrared Stereo Images. In: *2008 Canadian Conference on Computer and Robot Vision*, IEEE, pp. 105–112.

Hosseinyalamdary, S. and Yilmaz, A., 2015. Surface Recovery: Fusion of Image and Point Cloud. In: *2015 IEEE International Conference on Computer Vision Workshop (ICCVW*, IEEE, pp. 175–183.

Isola, P., Zhu, J.-Y., Zhou, T. and Efros, A. A., 2017. Image-to-Image Translation with Conditional Adversarial Networks. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, pp. 5967–5976.

Kehl, W., Milletari, F., Tombari, F., Ilic, S. and Navab, N., 2016. Deep Learning of Local RGB-D Patches for 3D Object Detection and 6D Pose Estimation. *ECCV* 9907(7), pp. 205–220.

Ketkar, N., 2017. Introduction to pytorch. In: *Deep Learning with Python*, Springer, pp. 195–208.

Kluger, F., Ackermann, H., Yang, M. Y. and Rosenhahn, B., 2017. *Deep Learning for Vanishing Point Detection Using an Inverse Gnomonic Projection*. Springer International Publishing, Cham, pp. 17–28.

Kniaz, V. V., Gorbatsevich, V. S. and Mizginov, V. A., 2017. THERMALNET: A DEEP CONVOLUTIONAL NETWORK FOR SYNTHETIC THERMAL IMAGE GENERATION. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLII-2/W4, pp. 41–45.

Knyaz, V. A., 2010. Multi-media Projector-Single Camera Photogrammetric System For Fast 3d Reconstruction. . . . *Archives of Photogrammetry* 38(PART 5), pp. 343–348.

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 Infrared Range. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2155–2164.

Li, C. and Wand, M., 2016. Precomputed Real-Time Texture Synthesis with Markovian Generative Adversarial Networks. *arXiv.org*.

Menze, M. and Geiger, A., 2015. Object scene flow for autonomous vehicles. *CVPR* pp. 3061–3070.

Nex, F., Gerke, M., Remondino, F., Przybilla, H. J., Bäumker, M. and Zurhorst, A., 2015. ISPRS BENCHMARK FOR MULTI-PLATFORM PHOTOGRAMMETRY. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences* II-3/W4, pp. 135–142.

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

Ronneberger, O., Fischer, P. and Brox, T., 2015. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. Springer International Publishing, Cham.

Wulff, J., Butler, D. J., Stanley, G. B. and Black, M. J., 2012. Lessons and Insights from Creating a Synthetic Optical Flow Benchmark. In: *Computer Vision – ECCV 2012. Workshops and Demonstrations*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 168–177.

Yamaguchi, M., Saito, H. and Yachida, S., 2017. Application of LSD-SLAM for Visualization Temperature in Wide-area Environment. *VISIGRAPP* pp. 216–223.

Zhang, R., Isola, P. and Efros, A. A., 2016. Colorful Image Colorization. *ECCV* 9907(Chapter 40), pp. 649–666.

Zhu, J.-Y., Park, T., Isola, P. and Efros, A. A., 2017. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In: *2017 IEEE International Conference on Computer Vision (ICCV)*, IEEE, pp. 2242–2251.