SYNTHETIC THERMAL BACKGROUND AND OBJECT TEXTURE GENERATION USING GEOMETRIC INFORMATION AND GAN

V. A. Mizginov[a], S. Yu. Danilov[a,b]

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

Commission II, WG II/5

KEY WORDS: infrared images, augmented reality, object recognition, generative adversarial network

ABSTRACT:

Nowadays methods based on deep neural networks show the best performance among image recognition and object detection algorithms. Nevertheless, such methods require to have large databases of multispectral images of various objects to achieve state-of-the-art results. Therefore the dataset generation is one of the major challenges for the successful training of a deep neural network. However, infrared image datasets that are large enough for successful training of a deep neural network are not available in the public domain. Generation of synthetic datasets using 3D models of various scenes is a time-consuming method that requires long computation time and is not very realistic. This paper is focused on the development of the method for thermal image synthesis using a GAN (generative adversarial network). The aim of the presented work is to expand and complement the existing datasets of real thermal images. Today, deep convolutional networks are increasingly used for the goal of synthesizing various images. Recently a new generation of such algorithms commonly called GAN has become a promising tool for synthesizing images of various spectral ranges. These networks show effective results for image-to-image translations. While it is possible to generate a thermal texture for a single object, generation of environment textures is extremely difficult due to the presence of a large number of objects with different emission sources.
The proposed method is based on a joint approach that uses 3D modeling and deep learning. Synthesis of background textures and objects textures is performed using a generative-adversarial neural network and semantic and geometric information about objects generated using 3D modeling. The developed approach significantly improves the realism of the synthetic images, especially in terms of the quality of background textures.

1. INTRODUCTION

In modern computer vision systems (enhanced vision (Vygolov et al., 2017) (Kniaz, 2014) system, autonomous driving (Kniaz, 2015)) the ability is most demanded to detect and recognize various objects with high probability in degraded visual conditions, such as fog, rain, night. Infrared cameras solve the problem of acquiring images in such conditions, however, the objects could visually vary greatly due to weather conditions. Therefore, a robust algorithm is required for detecting and recognizing objects in multispectral images. Deep convolutional neural networks have proven to be a reliable algorithm for detecting and recognizing objects in images of the visible range. Also, the latest network architectures make it possible to use this algorithms on multiplespectral images. However, the most important factor in the success of DCNN (deep convolutional neural network) learning is large multispectral datasets, which are very difficult to obtain using experiments. Today, deep convolutional networks are increasingly used for the goal of synthesizing various images. Recently a new generation of such algorithms commonly called generative adversarial network has become a promising tool for synthesizing images of various spectral ranges. These networks show effective results for image-to-image translations. While it is possible to generate a thermal texture for a single object using 3D modelling, generation of environment textures for large scene is extremely difficult due to the presence of a large number of objects with different emission sources. Also, 3D modelling is not of high quality in terms of imitation of noise and distortion of real sensor. The combination of deep learning and 3D modelling solves this problem. This paper is focused on the development of the method to thermal image synthesis using a GAN and 3D modelling. The aim of the presented work is to expand and complement the existing datasets of real thermal images.

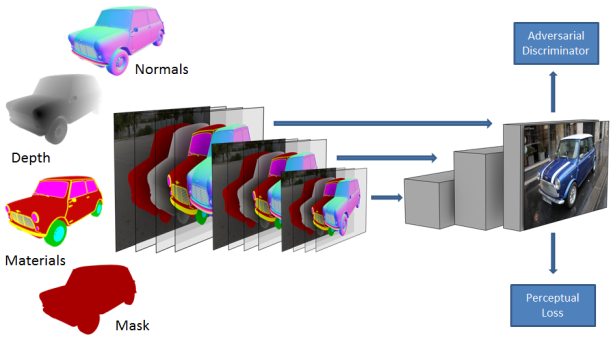


Figure 1. The pipeline of the GIS method

2. RELATED WORK

First research considering neural networks for image generation dates to 2013 (Zeiler and Fergus, 2013). Development of a new type of neural networks known as generative adversarial networks, made it possible to take a significant step forward in the field of synthesizing various images (Goodfellow et al., 2014). GAN consists of two deep convolutional neural networks: a Generator network tries to synthesize an image that visually indistinguishable from a given sample of images in the target domain. A discrimi-

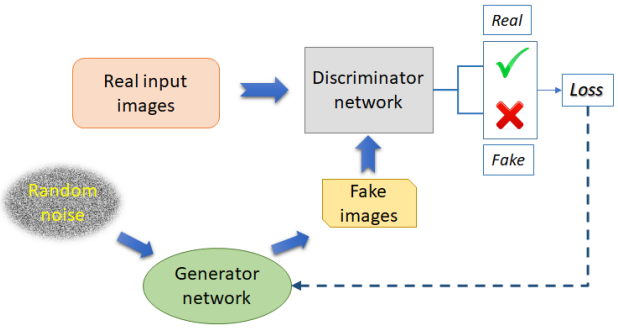


Figure 2. Basic idea of GAN

nator network tries to distinguish the “fake” images generated by the Generator network from the real image in the target domain. Generator and Discriminator networks are trained simultaneously. Such approach can be considered as an adversary game of two players (Figure 2)

The first goal that was performed using CNNs was the colorization of monochrome images (Zhang et al., 2016). Later CNNs has been used for simulating various artistic styles (Gatys et al., 2015) and transfiguration some objects in the image to others (Zhu et al., 2017). However, the realism and diversity of the results was insufficient. GANs significantly increased the quality of image-to-image translation (Isola et al., 2017). Recently GAN was applied to transform images from one spectral range to another. In (Liu et al., 2018) a method is proposed for converting near-infrared images to visible images without using paired pixel-wise aligned training dataset or rely on a colorful reference image(Figure 3). Our last papers presented the method for transformation of visible range images to infrared images (Kniaz et al., 2016). In (Kniaz and Mizginov, 2018) we present a new training method, which 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. In (Kniaz et al., 2018) two-step approach of image-to-image translation was developed. Firstly, we predict average object temperatures from an input color image. Secondly, we predict the relative local temperature contrasts, conditioned by a color image and thermal segmentation.

3. APPROACH

The proposed method is based on a joint approach that uses 3D modeling and deep learning (Alhaija et al., 2018) (Figure 1). Synthesis of background textures and objects textures is performed using a generative-adversarial neural network and semantic and geometric information about objects generated using 3D modeling. The developed approach significantly improves the realism of the synthetic images, especially in terms of the quality of objects textures.

3.1 Method

The task of converting images from one spectral range to another is ambiguous. If the translation of the infrared image into color is reduced to the problem of colorization, then the inverse transformation is multimodal. In other words, several synthesized images, the existence of which is physically possible in reality, can correspond to the original input color image. Another problem with GAN learning was that the thermal contrasts in the output image

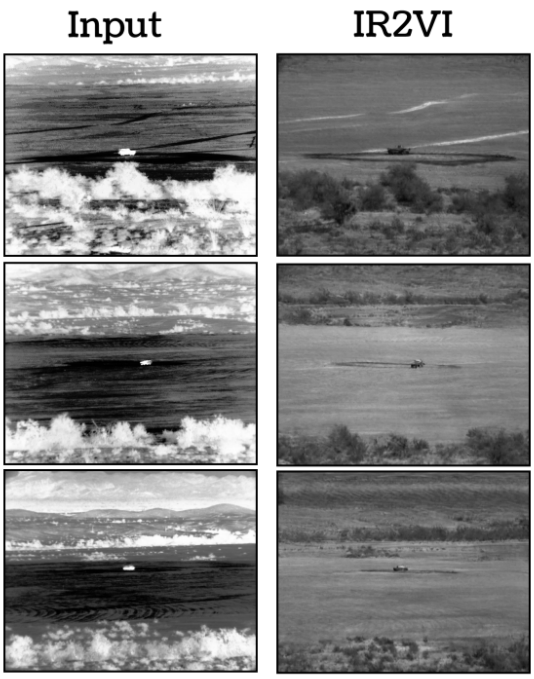


Figure 3. Infrared to visible range image translation

were averaged over the entire object, losing the characteristic thermal regions.

Based on the original method, we made the assumption that the use of segmentation of characteristic thermal zones of objects will increase the network’s ability to predict their location on the object. Accordingly, this will lead to an increase in the visual quality of the generated images. It turned out that some such thermal zones are almost unchanged in different weather conditions, different shooting conditions. As with semantic segmentation, thermal segmentation included the marking of thermal zones with certain labels that correspond to the degree of heating.

Unlike the original method, we will not need a normal map for each object, since the reflection of light does not play a significant role in the synthesis of thermal images. To generate maps of thermal zones of objects, as well as depth maps, we used realistic three-dimensional models created with the help of sophisticated three-dimensional modeling tools, since marking out thermal zones of objects on existing real images is a difficult task.

3.2 CNNs architecture

Our 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 9 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 8). Each level of the generator network has the dimension $W \times H \times R$, where W, H is the size of the attribute map (proportional to degree 2), R is its depth. The input images size is $512 \times 512 \times 7$, where

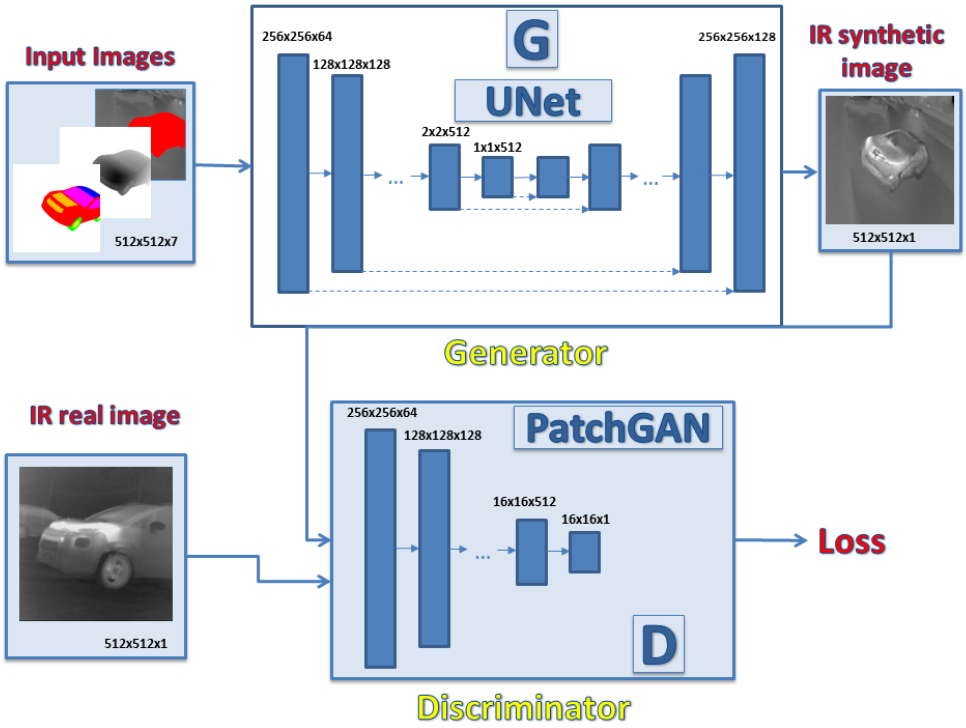


Figure 4. The proposed method for infrared image synthesis

the number of channels is determined by the label image, the thermal segmentation image and the depth map image. Convolutional layer consists of an input layer,intermediate layer and output layer. Its dimension is half the previous one.Each layer processes the input data with a filter with leaky ReLU nonlinearity. The output layer of the final module is followed by a convolution applied to the feature map and normalized to obtain the synthesized image. In other words, the U-Net is similar to a convolutional autoencoder with feedforward connections between the convolutional layers of the same depth. Such feed-forward connections increase the generator’s performance for restoration of small details and increase learning convergence. A discriminator network is based on the PatchGAN classifier (Li and Wand, 2016). The architecture consists of 6 fully connected layers having neurons with a leaky ReLU. The discriminator output is a one-dimensional binary map, where each value describes the discriminatory classification of a patch as real or synthesized by a generator. The final architecture of the generator network is presented in the Figure 4

3.3 Framework

The network was trained and tested using the PyTorch library (Ketkar, 2017).This framework was in 2017. It is an open source software designed to perform research on the design and training of deep neural networks.

4. EXPERIMENTS

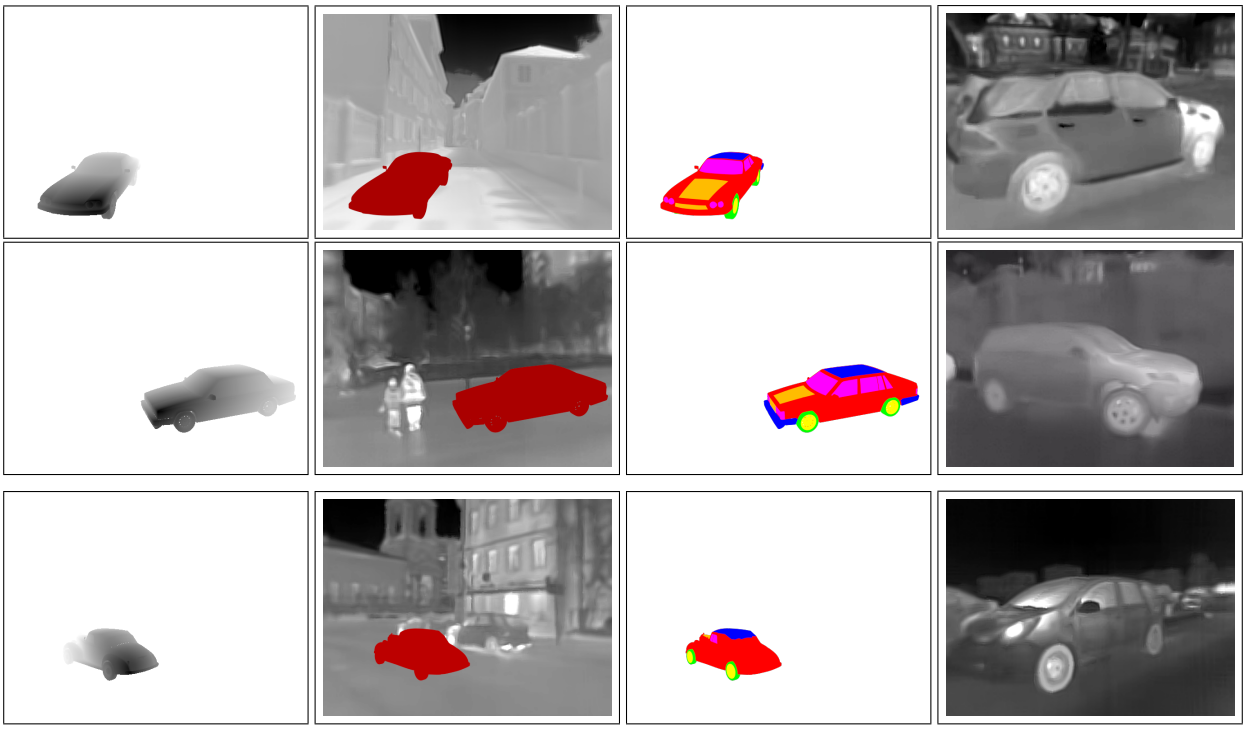
4.1 Input Image Dataset

We evaluate the developed method using a specially designed dataset. The images of cars were chosen as generated images. Perspective shooting looks like in KITTY 360 dataset (Alhaija et al., 2017). As input images we used the thermal contrasts map of objects, the depth map, the masks of objects.The samples were generated using the Blender 3D creation suite. It was prepared 6 three-dimensional models of cars. Two models were provided by the VoxelCity dataset (Knyaz et al., 2019). The heat contrasts map was formed by segmentation of several characteristic areas of vehicle heating while driving (hot motor, warm wheels, cold roof). Real infrared images of cars were used as a ground truth for training the network. Background images were obtained using the FLIR ONE PRO portable thermal imaging camera. The FLIR ONE camera produces as a standard output thermal preview images that present temperature of captured objects as monochrome (or pseudo-colors) images with a reference temperature scale bar. Also the FLIR ONE camera provides raw 16-bit data and the EXIF information for acquired images. Values of the raw data represent the object emission in the wavelengths 8–14 μ. The detailed technical specifications of the camera are presented in Table 1.

Parameter	Value
Visible range resolution	1440x1080
Infrared resolution	160x120
Field of view	43x55
Temperature range	-20…400 C
The spectral range	8 – 14 μm
Pixel size	12 μm

Table 1. FLIR One camera parameters.

The images were created in different places (city, park, country road, motorway) and different conditions of weather including snow, rain, fog. The images were scaled and cropped to square pictures 512×512 pixels. We intentionally captured all objects in similar conditions to provide uniform thermal contrast between the background and the object. The dataset includes 5000 thermal images, object masks, depth maps, thermal zone segmentations. Such approach provide semi unimodal distribution. Examples from the dataset are presented in Figure 5.



(a) Depth (b) Mask (c) Thermal segmentation (d) Groud thruth

Figure 5. Examples of input images dataset

4.2 Training CNN

We train the generator **G** synthesized images that were similar in appearance to real images. The goal of the network is to learn the process of generating images using same information about target objects. Unlike the original method, we train the network to transform model images of objects directly into real ifrared images by adding real background images of the environment to the input data. We use original `pix2pix` loss function. It provides state-of-the-art results for arbitrary image to image transforms on aligned image datasets. It uses conditional generator and conditional discriminator coupled with L1 loss:

$$L_{L1}(\mathbf{G}) = \mathbb{E}_{x,y,z}[||y - \mathbf{G}(x,z)||], \tag{1}$$

The final loss is given by:

$$\begin{aligned} L_{GAN3}(\mathcal{G},\mathcal{D}) =& \mathbb{E}_{\mathbf{X},\mathbf{Y}}[\log \mathcal{D}(\mathbf{X},\mathbf{Y})] + \\ & \mathbb{E}_{\mathbf{X},\mathbf{Z}}[\log(1 - \mathcal{D}(\mathbf{X},\mathcal{G}(\mathbf{X},\mathbf{Z}))]], \end{aligned} \tag{2}$$

where **Z** – is a random noise vector, that is used to avoid the deterministic output of the generator.

The network was trained using a NVIDIA RTX2080Ti captured GPU and was 200 epochs.This dataset was divided into independent training and test splits.

4.3 CNN evaluation

We used the independent test dataset to evaluate the GAN. To evaluate the generalization ability of the trained generator network we have performed generation of synthetic infrared images on samples of real background images. Some examples are presented in Figure 6. We evaluated our results using Learned Perceptual Image Patch Similarity (LPIPS) metric(Zhang et al., 2018). This method is a measure of ”perceptual distance” which measures how similar are two images in a way that coincides with human judgment. Since the use of the mean square error (RMS) metric to measure the difference between real infrared and synthesized images does not provide complete information about the similarity of synthesized and real images, we used the above method.The LPIPS metric estimates the distance between images, which ranges from 0 to 1.The less the value, it makes the images the more similar. For synthesized images shown in Figure 6, the following values are obtained (Table 2)

Image	Value
Image 1	0.309
Image 2	0.325
Image 3	0.195
Image 4	0.208
Image 5	0.256
Mean	0.258

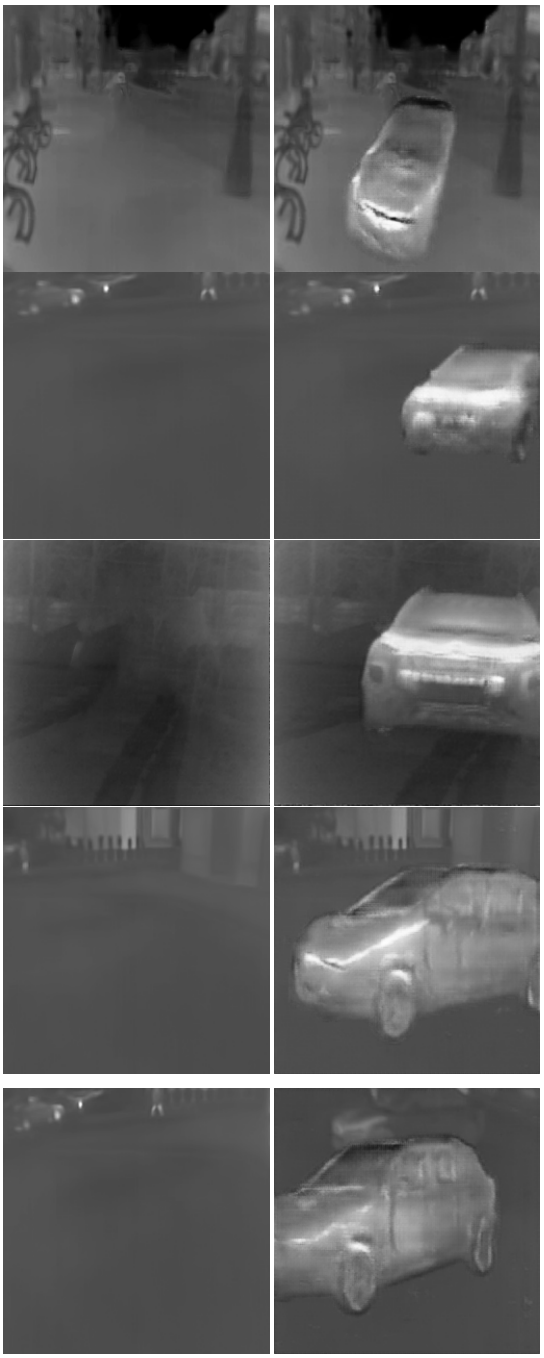
Table 2. Average LPIPS.

5. CONCLUSION

A new method for generation synthetic thermal images using a GAN was proposed. A training dataset was generated using the FLIR ONE PRO infrared camera and the Blender 3D creation suite. The size of the dataset is 5000 images. Also, The dataset includes 8 manualy created low-poly models of cars. To evaluate the proposed method the LPIPS metric was used. The evaluation of the generated infrared textures proved that they are similar to the ground truth model in both thermal emissivity and geometrical shape. The developed method allows for synthesizing realistic thermal images. The proposed approach can be used to supplement the existing training datasets with real infrared images.

6. ACKNOWLEDGEMENTS

The reported study was funded by Russian Foundation for Basic Research (RFBR) according to the research project № 17-29-03185



(a) Real background (b) GAN output IR

Figure 6. Examples of generated images

REFERENCES

Alhaija, H. A., Mustikovela, S. K., Geiger, A. and Rother, C., 2018. Geometric Image Synthesis. *CoRR*.

Alhaija, H. A., Mustikovela, S. K., Mescheder, L. M., Geiger, A. and Rother, C., 2017. Augmented reality meets deep learning. In: *British Machine Vision Conference 2017, BMVC 2017, London, UK, September 4-7, 2017*.

Gatys, L. A., Ecker, A. S. and Bethge, M., 2015. A Neural Algorithm of Artistic Style. *CoRR*.

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*.

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.

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

Kniaz, V., 2015. Fast instantaneous center of rotation estimation algorithm for a skied-steered robot. In: *Videometrics, Range Imaging, and Applications XIII*, Vol. 9528, International Society for Optics and Photonics, p. 95280L.

Kniaz, V., Gorbatsevich, V. and Mizginov, V., 2016. Generation of synthetic infrared images and their visual quality estimation using deep convolutional neural networks. *Scientific Visualization* 8(4), pp. 67–79.

Kniaz, V. V., 2014. A fast recognition algorithm for detection of foreign 3d objects on a runway. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XL-3, pp. 151–156.

Kniaz, V. V. and Mizginov, V. A., 2018. Thermal texture generation and 3d model reconstruction using sfm and gan. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLII-2, pp. 519–524.

Kniaz, V. V., Knyaz, V. A., Hladuvka, J., Kropatsch, W. G. and Mizginov, V., 2018. Thermalgan: Multimodal color-to-thermal image translation for person re-identification in multispectral dataset. In: *Computer Vision - ECCV 2018 Workshops - Munich, Germany, September 8-14, 2018, Proceedings, Part VI*, pp. 606–624.

Knyaz, V. A., Kniaz, V. V. and Remondino, F., 2019. Image-to-voxel model translation with conditional adversarial networks. In: L. Leal-Taixé and S. Roth (eds), *Computer Vision – ECCV 2018 Workshops*, Springer International Publishing, Cham, pp. 601–618.

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

Liu, S., John, V., Blasch, E., Liu, Z. and Huang, Y., 2018. IR2VI: enhanced night environmental perception by unsupervised thermal image translation. *CoRR*.

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

Vygolov, O. V., Gorbatsevich, V. S., Kostromov, N. A., Lebedev, M. A., Vizilter, Y. V., Knyaz, V. A. and Zheltov, S. Y., 2017. Semantic image segmentation for information presentation in enhanced vision. In: *Degraded Environments: Sensing, Processing, and Display 2017*, Vol. 10197, International Society for Optics and Photonics, p. 101970H.

Zeiler, M. D. and Fergus, R., 2013. Visualizing and Understanding Convolutional Networks. *arXiv.org* p. arXiv:1311.2901.

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

Zhang, R., Isola, P., Efros, A. A., Shechtman, E. and Wang, O., 2018. The unreasonable effectiveness of deep features as a perceptual metric.

Zhu, J., Park, T., Isola, P. and Efros, A. A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 2242–2251.