Synthesis and Visualization of Photorealistic Textures for 3D Face Reconstruction of Prehistoric Human*

Vladimir Kniaz[1,2] [0000−0003−2912−9986], Vladimir Knyaz[1,2] [0000−0002−4466−244X], and Vladimir Mizginov[1] [0000−0003−1885−3346]

[1] State Res. Institute of Aviation Systems (GosNIIAS), Moscow, Russia
[2] Moscow Institute of Physics and Technology (MIPT), Russia {knyaz, vl.kniaz, vl.mizginov}@gosniias.ru

Abstract. Reconstruction of face 3D shape and its texture is a challenging task in the modern anthropology. While a skilled anthropologist could reconstruct an appearance of a prehistoric human from its skull, there are no automated methods to date for automatic anthropological face 3D reconstruction and texturing. We propose a deep learning framework for synthesis and visualization of photorealistic textures for 3D face reconstruction of prehistoric human. Our framework leverages a joint face-skull model based on generative adversarial networks. Specifically, we train two image-to-image translation models to separate 3D face reconstruction and texturing. The first model translates an input depth map of a human skull to a possible depth map of its face and its semantic parts labeling. The second model, performs a multimodal translation of the generated semantic labeling to multiple photorealistic textures. We generate a dataset consisting of 3D models of human faces and skulls to train our 3D reconstruction model. The dataset includes paired samples obtained from computed tomography and unpaired samples representing 3D models of skulls of prehistoric human. We train our texture synthesis model on the *CelebAMask-HQ* dataset. We evaluate our model qualitatively and quantitatively to demonstrate that it provides robust 3D face reconstruction of prehistoric human with multimodal photorealistic texturing.

Keywords: Photogrammetry · 3D reconstruction · facial approximation · machine learning · generative adversarial networks · anthropology.

1 Introduction

Reconstruction of face 3D shape and its texture is a challenging task in the modern anthropology. While a skilled anthropologist could reconstruct an appearance of a prehistoric human from its skull, there are no automated methods to date for automatic anthropological face 3D reconstruction and texturing.

* The reported study was funded by Russian Foundation for Basic Research (RFBR) according to the research project 17-29-04509.

We propose a deep learning framework for synthesis and visualization of photorealistic textures for 3D face reconstruction of prehistoric human. Our framweork leverages a joint face-skull model based on generative adversarial networks. Specifically, we train two image-to-image translation models to separate 3D face reconstruction and texturing. The first model translates an input depth map of a human skull to a possible depth map of its face and its semantic parts labeling. The second model, performs a multimodal translation of the generated semantic labeling to multiple photorealistic textures.

We generate a dataset consisting of 3D models of human faces and skulls to train our 3D reconstruction model. The dataset includes paired samples obtained from computed tomography and unpaired samples representing 3D models of skulls of prehistoric human. We train our texture synthesis model on the *CelebAMask-HQ* dataset. We evaluate our model qualitatively and quantitatively to demonstrate that it provides robust 3D face reconstruction of prehistoric human with multimodal photorealistic texturing.

2 Related work

The development of modern technologies and the implementation of the new technologies in computer vision and deep learning have opened up wide opportunities for developing human face 3D reconstruction.

2.1 Human Face 3D Reconstruction

Manual facial approximation now is presented by the three main techniques: anthropometrical (American) method, anatomical (Russian) method, combination (British) method. The first one is based on soft tissue data and requires highly experienced stuff. Russian method [1] is performed by modeling muscles, glands and cartilage placing them onto a skull sequentially. This technique requires sufficient anatomical knowledge for accurate facial approximation. British method exploits the data of both soft tissue thickness and facial muscles.

The using of the computer-aided techniques for digital data processing has opened new possibilities for achieving realistic facial reconstruction. The facial approximation can be carried out through a programmatic face modeling by a surface approximation based on a skull 3D model and tissue thickness [2,3]. The 3D reconstruction of the face of Ferrante Gonzaga (1507 – 1557) has been performed using the physical model of the skull obtained by methods of computed tomography of his embalmed body and rapid prototyping [4]. The facial approximation of a 3,000-year-old ancient Egyptian woman [5] has been made with the use of medical imaging data.

Recent possibilities for collecting and processing big amounts of digital anthropological data allow to involve statistical and machine learning techniques for face approximation problem. The applying statistical shape models representing the skull and face morphology for the face approximation problem has been studied [6,7] by fitting them to a set of magnetic resonance images of the head. A large scale facial model – a 3D Morphable Model [8] has been automatically constructed from 9663 distinct facial identities. The 3D Morphable Model contains statistical information about a huge variety of the human population. A novel method for co-registration of two independent

statistical shape models was presented in [9].A face model is consistent with a skull model using stochastic optimization based on Markov Chain Monte Carlo (MCMC). A facial reconstruction is posed as a conditional distribution of plausible face shapes given a skull shape. Also deep learning models appear that are capable of multi-modal data translation [10,11] or generating object's shape 3D reconstruction basing on a single image [12,13].These approaches are also can be applied for facial approximation.

2.2 Generative Adversarial Networks

A new type of neural networks known as generative adversarial networks (GANs) [14] made it possible to take a significant step forward in the field of image processing. GANs consist of two deep convolutional neural networks: a Generator network tries to synthesize an image that visually indistinguishable from a given sample of images from the target domain. A Discriminator network tries to distinguish the 'fake' images generated by the Generator network from the real images in the target domain. Generator and Discriminator networks are trained simultaneously. This approach can be considered as an adversarial game of two players.

One of the first goals solved using GANs was image synthesis. Image-to-image translation problem was solved using conditional GAN termed `pix2pix` [15]. Such network learns a mapping $G : (x, z) \rightarrow y$ from observed image x and random noise vector z, to output y. This method also uses a sum of two loss functions: a conditional adversarial objective function and an L1 distance. However, for many tasks it is not possible to generate paired training datasets for image-to-image translation tasks.

To overcome this difficulty a `CycleGAN` [16] was proposed. The `CycleGAN` leverage a cycle consistency loss for learning a translation from a source domain X to a target domain Y in the absence of paired examples. Therefore, the `CycleGAN` model detects special features in one image domain and learns to translate them to the target domain. A new `StyleGAN` model was proposed in [17] that provides a superior performance in the perceptual realism and quality of the reconstructed image. Unlike the common generator architecture that feeds the latent code through the input layer, the `StyleGAN` appends a mapping of the input to an intermediate latent space, which controls the generator. Moreover, an adaptive instance normalization (AdaIN) is used at each convolution layer. Gaussian noise is injected after each convolution facilitating generation of stochastic features such as hair-dress or freckles. The problems of the first `StyleGAN` model were partially eliminated in the second `StyleGANv2` model [18]. In this model parameters are optimized and the neural network training pipeline was adjusted. The changes made have improved the quality of the results.

3 Method

Our aim is training two deep generative adversarial model for joint 3D face reconstruction and photorealistic texturing of prehistoric human. We use `pix2pixHD` [19] and `MaskGAN` [20] models as a starting point to develop our `skull2photo` framework. We also use assumptions of Knyaz et al. [21]. We provide two key contribution to the

original `skull2face` framework. Firstly, we add a new GAN model for photorealistic multimodal texturing of the reconstructed 3D face. Secondly, we replace the original `pix2pix` generator with a deeper `pix2pixHD` model.

3.1 `skull2photo` Framework Overview

Our aim is 3D reconstruction and texture generation of a prehistoric human face from a single depth map of its skull. We consider four domains: the skull depth map domain $\mathcal{A} \in \mathbb{R}^{W \times H}$, the face depth map domain $\mathcal{B} \in \mathbb{R}^{W \times H}$, the face semantic labeling domain $\mathcal{C} \in \mathbb{R}^{W \times H \times 3}$, and the face texture domain $\mathcal{D} \in \mathbb{R}^{W \times H \times 3}$.

We train two generator models: depth map generator G_1, and texture generator G_2. The aim of our depth map generator G_1 is learning a mapping $G : (\boldsymbol{A}, N) \rightarrow (\boldsymbol{B}, \boldsymbol{C})$, where N is a random vector drawn from a standard Gaussian distribution $\mathcal{N}(0, I))$, $\boldsymbol{A} \in \mathcal{A}$ is the input the skull depth map, $\boldsymbol{B} \in \mathcal{B}$ is the output face depth map, and $\boldsymbol{C} \in \mathcal{C}$ is the semantic labeling of the face parts similar to [20]. Our texture generator G_2 aims learning a mapping $G : \boldsymbol{C} \rightarrow \boldsymbol{D}$ from the semantic labeling $\boldsymbol{C}$ to the photorealistic face texture $\boldsymbol{D} \in \mathcal{D}$.

The multimodal adversarial loss governs the training process of our texture generator G_2

$$G^*, E^* = \arg\min_{G,E} \max_D \begin{array}{l} \mathcal{L}_{\mathrm{GAN}}^{\mathrm{VAE}}(G, D, E) + \lambda \mathcal{L}_1^{\mathrm{VAE}}(G, E) \\ + \mathcal{L}_{\mathrm{GAN}}(G, D) \\ + \lambda_{\mathrm{latent}} \mathcal{L}_1^{\mathrm{latent}}(G, E) + \lambda_{\mathrm{KL}} \mathcal{L}_{\mathrm{KL}}(E) \end{array}, \tag{1}$$

where $E(\mathbf{D})$ is the latent code generated by an encoder network similar to [22], and $\mathcal{L}_{\mathrm{KL}}$ is the Kullback–Leibler-divergence (KL-divergence) loss

$$\mathcal{L}_{\mathrm{KL}}(E) = \mathbb{E}_{\mathbf{D} \sim p(\mathbf{D})} \left[\mathcal{D}_{\mathrm{KL}}(E(\mathbf{D}) \| \mathcal{N}(0, I)) \right], \tag{2}$$

and $\mathcal{D}_{\mathrm{KL}}(p\|q)$ is an integral over a latent distribution encoded by $E(\mathbf{D})$

$$\mathcal{D}_{\mathrm{KL}}(p\|q) = -\int p(z) \log \frac{p(z)}{q(z)} dz \tag{3}$$

Overview of the proposed framework is presented in Figure 1.

3.2 Dataset Generation

For training the developed `skull2photo` framework a special crania-to-facial (*C2F*) dataset was created [21]. The *C2F* dataset includes data of two modalities: skull 3D models and face 3D models. For model training these 3D model were translated in depth map form. The *C2F* dataset has two parts. The first part is paired samples subset, containing the corresponding 3D models of a face and a skull, generated by processing computer tomography data. The paired samples subset contains 24 pairs of skull and face 3D models.

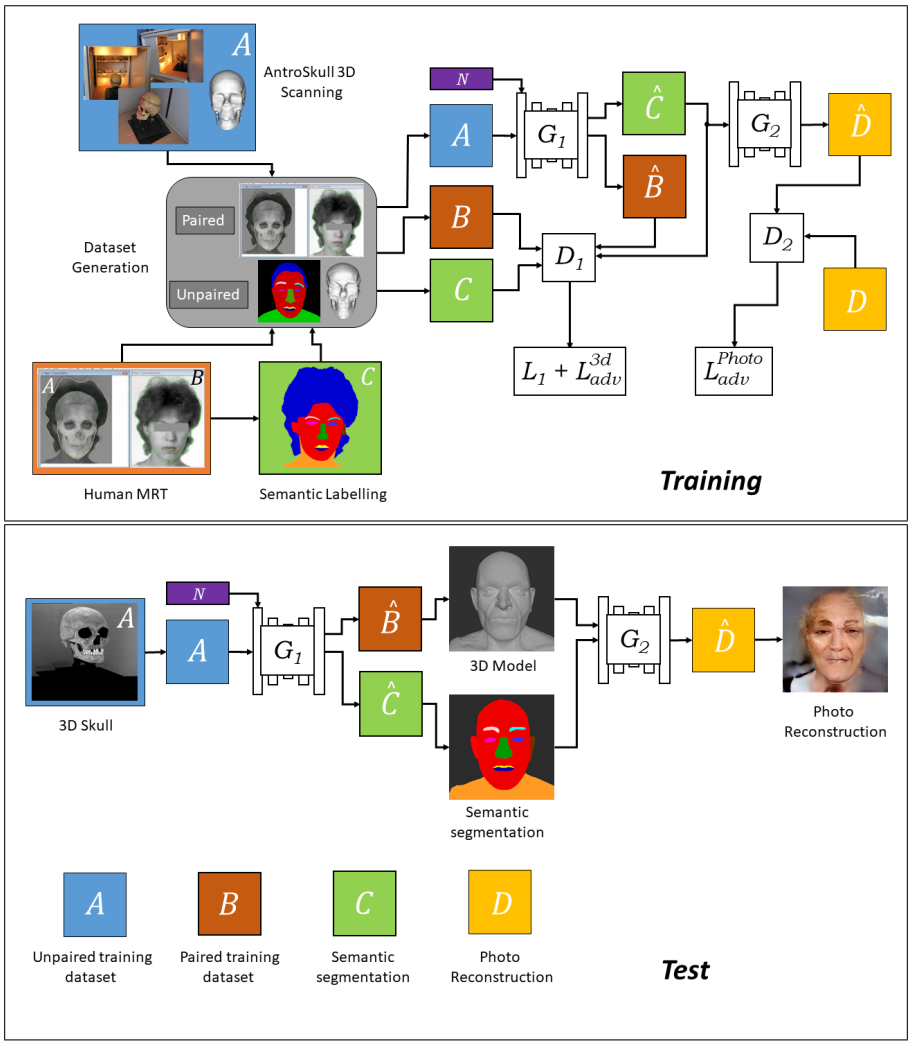

Fig. 1. Overview of the proposed model: training on the paired and unpaired dataset from skull depth map; testing the model.

4 Experiments

We evaluate our `skull2photo` framework qualitatively and quantitatively using the *C2F* and the *CelebAMask-HQ* [20]. Firstly, we present implementation details in Section 4.1. After that, we demonstrate qualitative results for face 3D reconstruction and texturing in Section 4.2. Finally, we explore quantitative results in terms of 3D shape accuracy in Section 4.3.

4.1 Network Training

Our framework is contained two GAN networks. The first of them is the `pix2pixHD` framework [19]. The `pix2pixHD` framework was designed to perform an arbitrary image-to-image transformation.We train the generator G_1 synthesized face depth map and semantic labels. The input images were skull depth map. We collected the original dataset that includes paired and unpaired skull depth map images. The unpaired samples subset contains 316 skull depth map images. The paired samples subset contains 200 pair depth map images.

The second neural networks is the `MaskGAN` [20]. The generator G_2 trained to reconstruct realistic photographs of human faces from semantic segmentation images. For this goals we used the CelebA dataset [23]. It is a large-scale face image dataset that has 30,000 high-resolution face images. Each image has a segmentation mask of facial attributes.

The network was trained and tested using the PyTorch library. It was trained using two NVIDIA RTX2080Ti captured GPU and was 200 epochs.This dataset was divided into independent training and test splits. The training of the generator G_1 was completed in 27 hours and the generator G_2 in 45 hours.

4.2 Qualitative Evaluation

The trained model was tested on independent testing dataset to reconstruct unseen faces. Firstly, for the qualitative evaluation we reconstructed modern human faces. We used a small part of *CelebAMask-HQ* dataset. Secondly, we tried to reconstructed the ancient man's face. This task is not easy because there are significant differences between modern man's face and ancient man's face.

Initially we generated face depth map images and semantic segmentation images using generator G_1. Then, we used previous received images as input for generator G_2 and reconstructed face photo texture. Finally, we selected several random style codes of face and synthesized several face samples. Examples were presented in Figure 2

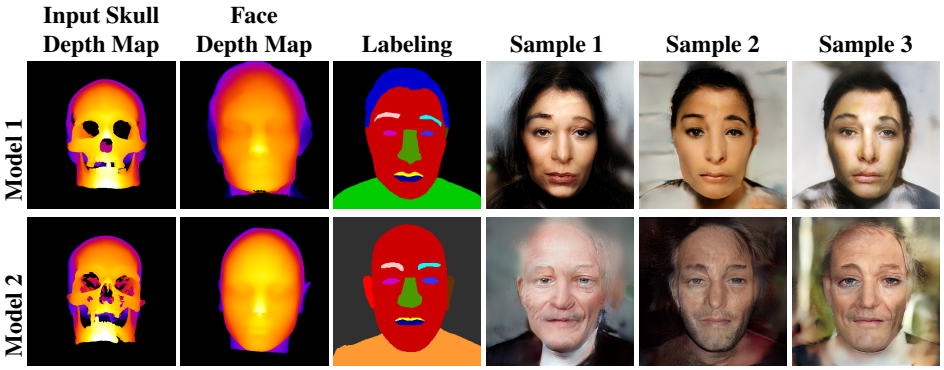

Fig. 2. Examples of input data and the results of the neural network.

4.3 Quantitative Evaluation

We present quantitative results on the independent test split of our *C2F* dataset in in Table 1. Depth maps predicted by the network are normalized to the range $[0, 1]$, where 0 is the front clipping plane located at 0 mm from virtual camera, and 1 is the far clipping plane located at 100 mm from camera. We use the L2 distance between the ground truth face depth map and the reconstructed depth map. We compare our `skull2photo` model to the `skull2face` [21] baseline. Experimental results demonstrate the the modified generator G_1 improves the quality of 3D reconstruction by 11%.

Table 1. Quantitative results on the independent test split of our *C2F* dataset.

L2 distance (mm)						
	Female		*Male*		*Average*	
	mean	**std**	**mean**	**std**	**mean**	**std**
`skull2face` [21]	15.31	1.26	16.37	1.25	15.84	1.26
Ours	**14.01**	**1.41**	**14.45**	**1.42**	**14.23**	**1.42**

5 Conclusion

We demonstrated that generative adversarial models can learn a challenging task of 3D face reconstruction and texturing of prehistoric human. Furthermore, we explored the possibility to generate different possible faces from a single skull using the KL-divergence loss function. Our main observation that the multimodal texture reconstruction model trained on images of the modern people can generalize to prehistoric human. We developed a two-stage framework for reconstruction of depth map and texture of a prehistoric human from a single depth map of its skull. The model was implemented using the PyTorch library and trained using three datasets. A paired dataset consisting of depth maps of human faces and corresponding skull was generated from computed tomography data. An unpaired dataset was developed by generating 3D reconstructions of skulls of prehistoric humans. A publicly available dataset *CelebAMask-HQ* dataset was used for training texture generation model. Both qualitative and quantitative evaluation proved that the our framework is capable of generating realistic 3D reconstructions of prehistoric human faces from a single depth map of a skull.

Acknowledgements

The reported study was funded by Russian Foundation for Basic Research (RFBR) according to the research project 17-29-04509.

References

1. Gerasimov, M.: The face finder. London: Hitchinson & Co (1971)

2. Knyaz, V.A., Zheltov, S.Y., Stepanyants, D.G., Saltykova, E.B.: Virtual face reconstruction based on 3D skull model. In: Corner, B.D., Pargas, R.P., Nurre, J.H. (eds.) Three-Dimensional Image Capture and Applications V. vol. 4661, pp. 182 – 190. International Society for Optics and Photonics, SPIE (2002), https://doi.org/10.1117/12.460172
3. Knyaz, V.A., Maksimov, A.A., Novikov, M.M.: Vision based automated anthropological measurements and analysis. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLII-2/W12, 117–122 (2019), https://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XLII-2-W12/117/2019/
4. Benazzi, S., Bertelli, P., Lippi, B., Bedini, E., Caudana, R., Gruppioni, G., Mallegni, F.: Virtual anthropology and forensic arts: the facial reconstruction of ferrante gonzaga. Journal of Archaeological Science 37(7), 1572 – 1578 (2010), http://www.sciencedirect.com/science/article/pii/S0305440310000233
5. Lindsay, K.E., Ruhli, F.J., Deleon, V.B.: Revealing the face of an ancient egyptian: Synthesis of current and traditional approaches to evidence-based facial approximation. The Anatomical Record 298(6), 1144–1161 (2015), https://anatomypubs.onlinelibrary.wiley.com/doi/abs/10.1002/ar.23146
6. Paysan, P., Lüthi, M., Albrecht, T., Lerch, A., Amberg, B., Santini, F., Vetter, T.: Face reconstruction from skull shapes and physical attributes. In: Denzler, J., Notni, G., Süße, H. (eds.) Pattern Recognition. pp. 232–241. Springer Berlin Heidelberg, Berlin, Heidelberg (2009), https://doi.org/10.1007/978-3-642-03798-6{_}24
7. Paysan, P., Knothe, R., Amberg, B., Romdhani, S., Vetter, T.: A 3d face model for pose and illumination invariant face recognition. In: 2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance. pp. 296–301 (Sep 2009)
8. Booth, J., Roussos, A., Ponniah, A., Dunaway, D., Zafeiriou, S.: Large scale 3d morphable models. International Journal of Computer Vision 126(2), 233–254 (Apr 2018), https://doi.org/10.1007/s11263-017-1009-7
9. Madsen, D., Lüthi, M., Schneider, A., Vetter, T.: Probabilistic joint face-skull modelling for facial reconstruction. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. pp. 5295–5303 (2018), http://openaccess.thecvf.com/content_cvpr_2018/html/Madsen_Probabilistic_Joint_Face-Skull_CVPR_2018_paper.html
10. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-Image Translation with Conditional Adversarial Networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5967–5976. IEEE (2017)
11. Kniaz, V.V., Knyaz, V.A., Hladůvka, J., Kropatsch, W.G., Mizginov, V.: Thermalgan: Multimodal color-to-thermal image translation for person re-identification in multispectral dataset. In: Leal-Taixé, L., Roth, S. (eds.) Computer Vision – ECCV 2018 Workshops. pp. 606–624. Springer International Publishing, Cham (2019), https://link.springer.com/chapter/10.1007/978-3-030-11024-6_46
12. Kniaz, V.V., Remondino, F., Knyaz, V.A.: Generative adversarial networks for single photo 3d reconstruction. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLII-2/W9, 403–408 (2019), https://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XLII-2-W9/403/2019/
13. Knyaz, V.: Machine learning for scene 3d reconstruction using a single image. Proc. SPIE 11353, Optics, Photonics and Digital Technologies for Imaging Applications VI 11353, 1135321 (2020), https://doi.org/10.1117/12.2556122
14. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Ghahramani, Z., Welling, M., Cortes, C.,

Lawrence, N.D., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 27, pp. 2672–2680. Curran Associates, Inc. (2014), http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf
15. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. CVPR (2017)
16. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Computer Vision (ICCV), 2017 IEEE International Conference on (2017)
17. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. CoRR abs/1812.04948 (2018), http://arxiv.org/abs/1812.04948
18. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan (2019)
19. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)
20. Lee, C.H., Liu, Z., Wu, L., Luo, P.: Maskgan: Towards diverse and interactive facial image manipulation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
21. Knyaz, V.A., Kniaz, V.V., Novikov, M.M., Galeev, R.M.: Machine learning for approximating unknown face. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLIII-B2-2020, 857–862 (2020), https://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XLIII-B2-2020/857/2020/
22. Zhu, J., Zhang, R., Pathak, D., Darrell, T., Efros, A.A., Wang, O., Shechtman, E.: Toward multimodal image-to-image translation. In: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA. pp. 465–476 (2017), http://papers.nips.cc/paper/6650-toward-multimodal-image-to-image-translation
23. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision (ICCV) (December 2015)