RSGAN: Face Swapping and Editing using Face and Hair Representation in Latent Spaces

Ryota Natsume
Waseda University
ryota.natsume.26@gmail.com

Tatsuya Yatagawa
Waseda University
tatsy@acm.org

Shigeo Morishima
Waseda University
shigeo@waseda.jp

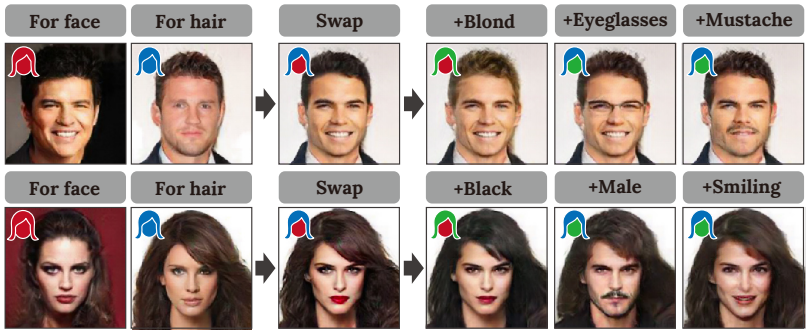

Figure 1: Results of face swapping and additional visual attribute editing with the proposed system. The face regions of images in the first column are embedded to the images in the second column. The face-swapped results are illustrated in the third column, and their appearances are further manipulated by adding visual attributes such as "blond hair" and "eyeglasses". Note that the input images in the first two columns are synthesized to reconstruct the original images by the proposed network. The original images are found in the URLs, goo.gl/qBv3a7, goo.gl/E7mhYm, goo.gl/1cSZSo, and goo.gl/3xH9jk.

CCS CONCEPTS

• **Computing methodologies** → **Computer graphics**; Image manipulation;

KEYWORDS

face, portrait, face swapping, image editing

ACM Reference Format:
Ryota Natsume, Tatsuya Yatagawa, and Shigeo Morishima. 2018. RSGAN: Face Swapping and Editing using Face and Hair Representation in Latent Spaces. In *Proceedings of SIGGRAPH '18 Posters.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3230744.3230818

1 INTRODUCTION

This abstract introduces a generative neural network for face swapping and editing face images. We refer to this network as "region-separative generative adversarial network (RSGAN)". In existing deep generative models such as Variational autoencoder (VAE) and Generative adversarial network (GAN), training data must represent what the generative models synthesize. For example, image inpainting is achieved by training images with and without holes. However, it is difficult or even impossible to prepare a dataset which includes face images both before and after face swapping because faces of real people cannot be swapped without surgical operations. We tackle this problem by training the network so that it synthesizes synthesize a natural face image from an arbitrary pair of face and hair appearances. In addition to face swapping, the proposed network can be applied to other editing applications, such as visual attribute editing and random face parts synthesis.

2 REGION-SEPARATIVE GAN

Figure 2 illustrates the proposed RSGAN that comprises two VAEs, which we refer to as *separator networks*, and one GAN, which we refer to as *composer network*. As shown in Fig. 2, the separator networks first encode image x and visual attributes c of the face and hair regions into different latent-space representations. Then, the composer network decodes the latent-space representations to a face image such that the original appearance is reproduced in it. The input and synthesized images are evaluated by global discriminator and local discriminator. The global discriminator distinguishes whether these images are real or fake as in standard GANs. On the other hand, the local discriminator distinguishes whether local patches in those images are real or fake. We train the separator networks with ℓ_1 and Kullback Leibler losses, and the composer network with ℓ_1 and adversarial losses.

In addition to real image samples, we also use random latent variables to train the network. The random variables are sampled from a multivariate normal distribution $\mathcal{N}(0, 1)$. The composer network generates a random image with sampled random variables. This training with random images enables the composer network to generate a natural image from arbitrary latent variables for

SIGGRAPH '18 Posters, August 12-16, 2018, Vancouver, BC, Canada

ACM ISBN 978-1-4503-5817-0/18/08.
https://doi.org/10.1145/3230744.3230818

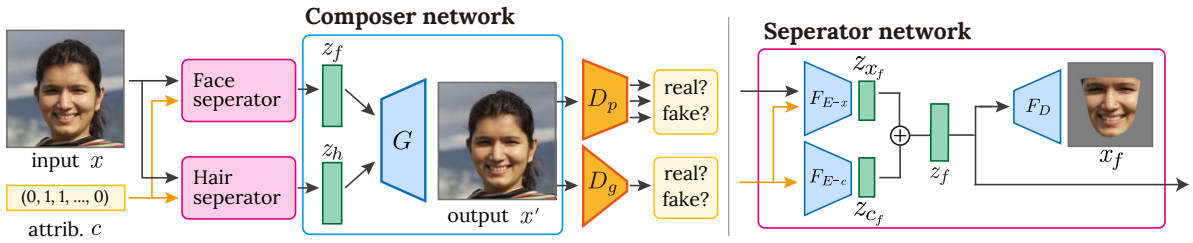

Figure 2: The network architecture of the proposed RSGAN that comprises two separator networks and a composer network. The separator networks extract latent-space representations z_f and z_h respectively for face and hair regions of input image x. The composer network reconstructs the input face image from the two latent-space representations. The face image in this figure courtesy of Flickr user "Megha" and is found in goo.gl/hzpqHq.

faces and hairs. As a result, the composer network can achieve face swapping with latent variables extracted from two different images.

We can perform face swapping with RSGAN in two steps. When a face region of a target image is replaced with the face region of a source image, face latent variables of the target image and hair latent variables of the source image are extracted by the separator networks. Then, the composer network generates a face-swapped result from these two latent variables.

For the purpose of preserving the hairstyle and background region of a target image, we apply gradient-domain image stitching [Levin et al. 2004] to a face-swapped image. In this operation, the face region of a face-swapped image is stitched with the target image. We denote these approaches with and without image stitching as "RSGAN" and "RSGAN-GD", respectively. The results of Fig. 1 were generated with RSGAN, and the results of Fig. 3 were generated with RSGAN-GD.

3 RESULTS AND CONCLUSION

The results of editing applications using RSGAN are shown in Fig. 1. In the supplementary video, we explained how these application can be achieved. Note that RSGAN can achieve both face swapping and visual attribute editing by passing two input images and modified visual attribute vectors through the network only once. We compared our face-swapping results with the state-of-the-art methods [Kemelmacher-Shlizerman 2016; Nirkin et al. 2018] in Fig. 3. While Shilzerman [2016] achieved natural face swapping, this method needs to select a target from a large-scale database such that the two image layouts are similar. Therefore, our proposed method is advantageous in practice because it can swap face regions between arbitrary images. On the other hand, Nirkin et al. [2018] swapped faces using 3D morphable models(3DMM). In their method, the face geometries and their corresponding texture maps are obtained by fitting 3DMM. Then, the texture maps of source and target images are swapped. Finally, the replaced face textures are re-rendered using lighting condition estimated for the target image. Compared to their results in Fig. 2, our results look more natural in terms that the proportion of sizes between facial parts and the entire face regions are similar to those in the source images. As reported in [Nirkin et al. 2018], these performance losses are due to their sensitiveness to the quality of landmark detection and 3DMM fitting.

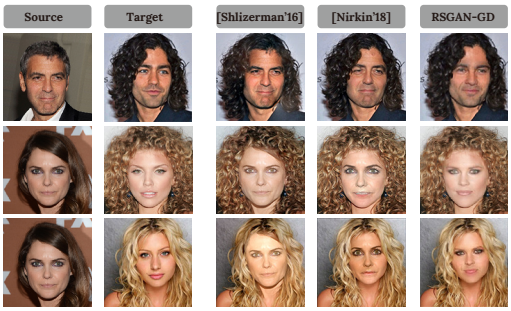

Figure 3: Comparisons of the proposed method to the state-of-the-art face swapping methods [Kemelmacher-Shlizerman 2016; Nirkin et al. 2018]. Images in the three left columns courtesy of [Kemelmacher-Shlizerman 2016].

In conclusion, we proposed an integrated editing system using a deep generative model that we refer to as RSGAN. The proposed method can achieve face swapping between arbitrary image pairs, and can robustly perform the swapping compared to previous methods using 3DMM.

ACKNOWLEDGMENTS

This study was granted in part by the Strategic Basic Research Program ACCEL of the Japan Science and Technology Agency (JPMJAC1602). Tatsuya Yatagawa was supported by a Research Fellowship for Young Researchers of Japan's Society for the Promotion of Science (16J02280). Shigeo Morishima was supported by a Grant-in-Aid from Waseda Institute of Advanced Science and Engineering. The authors would like to acknowledge NVIDIA Corporation for providing their GPUs in the academic GPU Grant Program.

REFERENCES

Ira Kemelmacher-Shlizerman. 2016. Transfiguring portraits. *TOG* 35, 4 (2016), 94:1–94:8. https://doi.org/10.1145/2897824.2925871

Anat Levin, Assaf Zomet, Shmuel Peleg, and Yair Weiss. 2004. Seamless image stitching in the gradient domain. In *Proc. of European Conference on Computer Vision (ECCV)*. 377–389. https://doi.org/10.1007/978-3-540-24673-2_31

Yuval Nirkin, Iacopo Masi, Anh Tuan Tran, Tal Hassner, and Gérard Medioni. 2018. On Face Segmentation, Face Swapping, and Face Perception. In *Proc. of IEEE Conference on Automatic Face and Gesture Recognition.*