ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10446993

ATTRIBUTE-AWARE HEAD SWAPPING GUIDED BY 3D MODELING

Wenbo Zhou[1], *Dongdong Chen*[2], *Jing Liao*[3], *Jie Zhang*[1], *Kejiang Chen*[1], *Weiming Zhang*[1], *Nenghai Yu*[1]

[1]University of Science and Technology of China, [2]Microsoft Cloud AI, [3]City University of Hong Kong

ABSTRACT

Face manipulation has ignited the interests of both academia and industry in very recent years. Existing face manipulation methods can be roughly categorized into two types: face attribute editing and face swapping. In this paper, we focus on swapping the identity. But unlike face swapping which only changes the face region, we attempt at a more challenging task: attribute-aware head swapping. Given a source video and a target video, we replace the whole target head with the whole source head while keeping the original target attributes. To address the inherent appearance gap (e.g., hairstyle, face shape), accompanying background incompatibility and lighting difference, our method consists of three key components: 1) a generative rendering-to-real-head model for source head modeling and attribute transfer; 2) a background modeling network to fix the background incompatibility during head swapping; 3) a deep harmonization network to fix remaining issues and makes the final composited result more realistic. We compare our approach to different face manipulation methods and the experimental results demonstrate its superiority for a lot of challenging cases.

Index Terms— Image Synthesis, face manipulation, head swapping, 3D face modeling

1. INTRODUCTION

Thanks to the tremendous success of deep generative models and extensive entertainment needs, face manipulation has been an emerging hot research topic in recent years and a variety of methods [1, 2, 3, 4, 5, 6, 7, 8] have been proposed. Depending on the manipulation goals, these methods can be roughly categorized into two types: *face attribute editing* [1, 5] that only aims to change the attribute of one face such as expression or hair style while keeping the original identity, and *face swapping* [2, 3, 8] that focuses on replacing the target identity with the source identity instead. In this paper, we also focus on the latter identity swapping task but with a more challenging goal, i.e., attribute-aware head swapping but not just face region swapping.

Head swapping is an important but unsolved problem because of its inherent challenges, including but not limited to significant appearance differences like hairstyle and face shape, accompanying background incompatibility after foreground swapping, and lighting difference. In contrast, face attribute editing and face swapping do not consider such challenges and most parts of the original head remain unchanged. For general cases when the source and target appearance is significantly different, existing face swapping techniques will fail and produce very unrealistic results.

Motivated by this, we present the attempt at attribute-aware head swapping. We propose a novel 3D modeling guided attribute-aware head swapping framework which consists of three key components. For the foreground head swapping, a generative rendering-to-real-head model is utilized to transfer the attribute-specific rendered source face into the corresponding real source head. We use a face reconstruction model to decompose each source frame into identity and attribute components and get the rendered head based on these two components. To address the background incompatibility issue, a background modeling network is trained on the target video by simulating different shapes of mask around the head region. Finally, we use a deep harmonization network to make the composition of the swapped source head and target background look more real and fix potential lighting and skin color incompatibility.

The main contributions of our method are as below:

- We creatively attempt attribute-aware head swapping, which can circumvent the limitations of existing face swapping methods and may inspire more promising works along this direction.
- We propose a novel 3D modeling guided head swapping framework consists of three key components to handle the head swapping, background incompatibility, and harmonization problems respectively.
- We have shown the superiority of our method over existing methods in a lot of challenging cases.

2. METHODS

To solve the challenging head swapping task, we use a coarse 3D modeling method but combine it with a generative model to learn the mapping between the coarsely rendered head and the real head. Then we render the source head with the target attribute and feed it into the generative model to obtain a high-quality transferred head. Finally, we use a head parsing

Thanks to the Natural Science Foundation of China under Grant 62372423, 62121002, U20B2047, 62072421, and 62002334, Key Research and Development program of Anhui Province under Grant 2022k07020008 for funding.

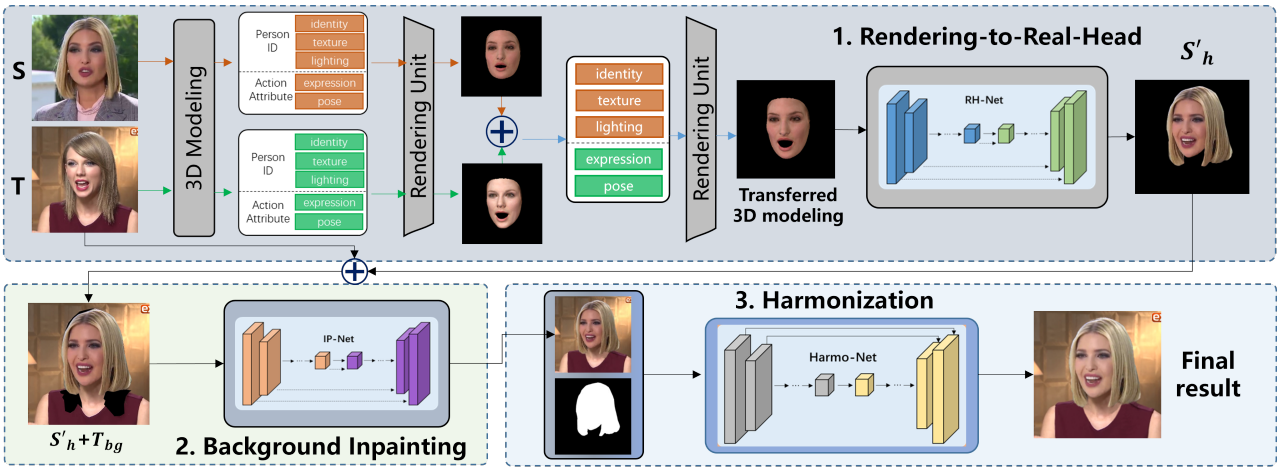

Fig. 1: The framework of our proposed method, which consists of three main components: a rendering-to-real-head model for realistic attribute-aware head modeling, a background inpainting network to fix the background incompatibility issue, and a deep harmonization network to make the final composited result more realistic.

network to get the target foreground mask and align the transferred source head to the corresponding target head position.

When composing the transferred source head and target background, since the source and target head usually have different shapes and hairstyles, some original regions of the foreground target head will be removed and those background pixels behind the target head become undefined and need to be inpainted. Therefore, a background modeling network is further leveraged. By the end of this step, our system has already achieved the basic head swapping function. But due to the appearance differences between the source and target such as potential skin color and lighting difference, a following harmonization network is further used to make the final composited result more realistic.

2.1. Coarse 3D Modeling and Rendering

In order to render the source head in an attribute-aware way, it needs to be modelled into two different types of components: identity related components and attribute related components. We adopt one existing state-of-the-art deep learning based face reconstruction method [9] for 3D modeling. For each frame of the source video or target video, we first use a face parsing model to segment the whole head based on BiSeNet [10] trained on the CelebAMask-HQ dataset [11], then use the 3D modeling method to fit a parametric face model.

Assume the face image to be I, this parametric model will decompose I into five independent components: identity I^{id}, expression I^{exp}, texture I^{tex}, illumination I_{ill} and pose I^{p}. These five components can be categorized into identity related components (I^{id}, I^{tex}) and attribute related components (I^{exp}, I^{ill}, I^{p}). The identity related components will be used to swap the source head to target, while the attribute related components will be used to keep the original target attributes.

We use the mesh renderer [12] to reconstruct the given head image, regarding the rendered head as the input and the original head image as the ground truth to train the following rendering-to-real-head network. The rendering process can be represented as:

$$\hat{I} = \mathcal{R}(I^{id}, I^{tex}, I^{exp}, I^{ill}, I^{p}) \quad (1)$$

When swapping the source head to target with the target attributes, different compontes are recombined as follows:

$$\hat{I_s}' = \mathcal{R}(I_s^{id}, I_s^{tex}, I_t^{exp}, I_t^{ill}, I_t^{p}) \quad (2)$$

Then we input this coarsly rendered face into the generative network to obtain the real head.

2.2. Rendering-to-Real-Head Translation

For the rendering-to-real-head translation network **RH-Net**, it adopts a UNet-like network structure which consists of a transformation network and a discriminator. This **RH-Net** mainly consists of an encoder part which computes the low-dimensional latent representation of the input, and a decoder part which synthesizes the output image. Given the rendered image $\hat{I}_s$ of size 256×256 as input, the encoder part consists of 8 downsampling blocks with 64-512 output channels and the innermost feature size becomes 1×1. Each downsampling block is ended with a batch normalization and a non-linear ReLU layer. In this way, the **RH-Net** can capture a very large receptive field, which significantly makes the source head modeling much easier.

Symmetrically, 8 upsampling blocks with 512-3 channels are used to output the reconstructed head image. And the final activation layer TanH brings the output to the normalized range [-1, +1] which helps the network to obtain better performance. We also adopt skip connections which can help the network to transfer the fine-scale structure of the input.

To train a high-quality **RH-Net**, the common pixelwise L_1 loss is utilized to penalize the distance between the synthesized output $RH(\hat{S})$ and the ground-truth image, and an adversarial loss is also used by using a PatchGAN [13] based adversarial network D to improve the quality of synthesized images. The loss function has the following form:

$$\mathcal{L}_{RH-Net} = \lambda_1 * \ell_{adv}(O_s, I_s) + \lambda_2 * \ell_1(O_s, I_s) \quad (3)$$

where O_s is the output of **RH-Net** when inputting $\hat{I}_s$, and λ_1, λ_2 are the balancing weights for these two terms. By default, $\lambda_1 = 0.01$ and $\lambda_2 = 1$ in our following experiments.

2.3. Background Inpainting

Due to the different shapes between the transferred source head and original target head, part of the foreground target head may be removed where needs to be further inpainted. Thus the initial composited image will be fed into the **IP-net** to inpaint the undefined background regions that are occluded by the target head. **IP-net** adopts a similar network architecture as **RH-Net**, but an extra hole mask (1 channel) is also included as the input in order to explicitly guide *IP-Net* which regions need to be inpainted.

We use the disjoint regions of the source and target head masks as approximated incompatible masks to simulate the possible incompatible background hole mask for training. The source head mask and the target head mask are randomly selected and combined thus the holes will appear in different regions of the target background. This enables the network can deal with various cases. We consider four types of loss functions for the objective function as below:

$$\begin{aligned}\mathcal{L}_{IP-Net} =& \lambda_3 * \ell_{adv}(O_t, I_t) + \lambda_4 * \ell_1(O_t, I_t) \\ &+ \lambda_5 * \ell_{vgg}(O_t, I_t) + \lambda_6 * \ell_{style}(O_t, I_t)\end{aligned} \quad (4)$$

where O_t is the output of **IP-Net**, and I_t is the ground truth target image. ℓ_{adv} and ℓ_1 are defined in the same way as Equation 3. The latter perceptual loss ℓ_{vgg} and style loss ℓ_{style} are inspired by existing style transfer methods [14, 15] to increase the texture fidelity. We set $\lambda_3 = 0.01$, $\lambda_4 = \lambda_5 = \lambda_6 = 1$ in our experiments.

2.4. Foreground and Background Harmonization

The differences in appearance attributes are common in arbitrary source and target video pairs, directly composing the transferred source head and the target background will make the results look a little unrealistic. To address such problems and make the final swapping results more natural, an end-to-end deep harmonization network **HM-Net** is leveraged.

To train such a network in a fully-supervised way, data acquisition plays the most important role. We take an inverse strategy to generate large-scale training pairs. In detail, we start from a real portrait image which is treated as the ground truth of the **HM-Net**. Then we randomly change the foreground and background appearance to generate a manipulated image that will be taken as input to the **HM-Net**. Specifically, after getting the foreground mask, we randomly change the foreground and background color based on the reference-based color transfer method [16] or adjust the brightness of the foreground and background. In this way, we can ensure that the ground-truth images are always real so that the **HM-Net** will learn to reconstruct a realistic output from a manipulated image. In addition, a 1-channel foreground mask is utilized as input to provide auxiliary guidance to help the network focus on harmonizing the edited foreground.

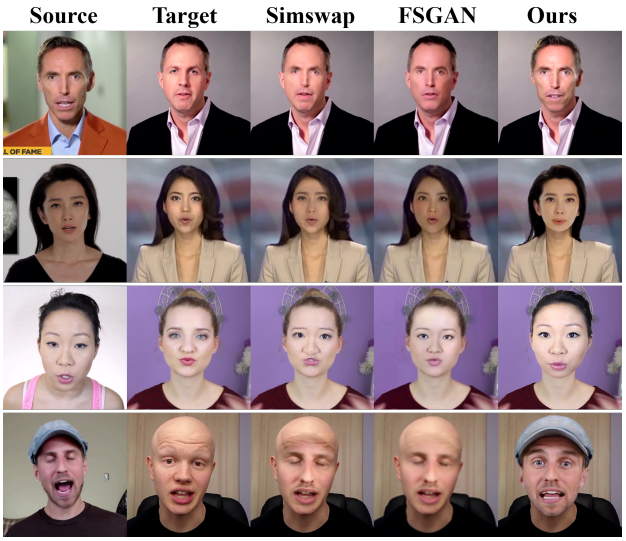

Fig. 2: Visual comparisons with face identity swapping methods. Obviously, the head swapping results of our method are much more natural and preserve the source identity better.

The harmonization process is also an image-to-image task, thus we utilize a simple UNet-256 as the the network structure of **HM-Net**. Similar to **RH-Net**, two types of loss functions ℓ_{adv} and ℓ_1 are adopted to comprise the objective function as Equation 3, with balancing weights of 0.01 and 1.

3. EXPERIMENT

In this section, we compare our method with several face manipulation methods both qualitatively and quantitatively. All the portrait videos shown in this paper are collected from YouTube and FaceForensic++ [17] for research purposes.

3.1. Qualitative Comparison

We first qualitatively compare our method with face swapping methods Simswap[3], FSGAN [2] and face attribute transfer method Head2Head++ [5] respectively. These methods are chosen as baselines because they are open-sourced and perform satisfied quality in the vast majority of cases. As shown in Figure 2, Simswap and FSGAN aim to swap the identity from source to target, however, due to the large difference in attributes between the source and target, the identity preservation of the swapped results is not high. Besides, a lot of texture details will be lost in these methods and incur blurry results. More importantly, some key identity properties cannot be swapped, such as the hairstyle in the first three examples. In contrast, our method highly maintains the source identity properties while transferring the target attributes. Moreover, our method performs well in some extreme cases. In the second case of Figure 2, the hairstyle and skin color of source and target portraits are very different. In the last case, the source portrait wears a hat but the target portrait does not. In fact, these are the most representative cases that existing face

Fig. 3: Visual comparisons with the face attribute transfer method. Our method can preserve the target attribute well.

swapping cannot handle. But thanks to the powerful modeling ability of generative networks, our method can still achieve excellent attribute-aware head swapping results.

Figure 3 is the visual comparison between our method and Head2head++. Head2head++ only reenacts the poses and expressions of the target on source portraits and preserves the source background in the generated image. Though head swapping is much more challenging than attribute transfer, our method can also preserve the target attribute very well thanks to the attribute-aware modeling mechanism.

3.2. Quantitative Comparison.

Currently, user study is a common way to judge the quality of face manipulation results. In this part, we conduct a user study. Given a total of 50 source&target image pairs and the corresponding swapping results of different methods, we ask 67 participants to judge the swapping results for each method from three aspects: 1) The identity fidelity of the swapping result with the source portrait; 2) The attribute similarity of the swapping result with target portrait; 3) The overall realism of the swapping result. The score range of each aspect is from 0 to 100, and the higher score represents better results.

Table 1: User study results of quantitative comparison for the swapping results. The higher scores mean better results.

Method	Identity	Attribute	Realism
Simswap	69.9	74.5	71.0
FSGAN	51.2	65.7	58.8
Ours	**91.1**	**86.02**	**94.45**

Since Head2head++ does not support identity swapping, we only compare with Simswap and FSGAN. In Table 1, our method achieves significantly better results by a large margin. Especially, though Simswap and FSGAN are both face identity-swapping methods, most users cannot even tell the source identity from the swapped results. This also justifies the necessity of head swapping. Moreover, we use several state-of-the-art deepfake detection models [18, 19] to evaluate the anti-detection ability. The average accuracy of used models significantly reduces (from 97% to 57%) when detecting the results generated by our method. The testing results are not listed because of space limitations.

Fig. 4: Intermediate results of our methods to show the functionality of each component.

3.3. Functionality Analysis for Each Component

We provide two groups of intermediate results for each component in Figure 4. Obviously, **RH-Net** can generate high-quality source heads, but some undefined background pixels might appear after aligning to the target background. By comparing the second and the third column, we can observe that although the incompatible background regions may produce a large hole, our **IP-Net** can inpaint these holed parts very well. Finally, **HM-Net** can fix the potential skin color and lighting differences and make the final results more realistic.

To further quantitatively assess the harmonization functionality of **HM-Net**, we calculate the color difference between randomly selected 100 target images and the outputs before/after **HM-Net** as [20], the numerical results are in Table 2. The smaller color difference represents more harmonious results. It shows that **HM-Net** can reduce the average color difference from 10.89 to 7.62, which is better than the other two methods and consistent with the visual results.

Table 2: Color difference results comparison among Simswap, FSGAN and our method. Smaller color difference represents more natural synthesis results.

Groups	Before **HM-Net**	Final
Simswap	/	11.22
FSGAN	/	14.95
Ours	10.89	**7.62**

4. CONCLUSION AND DISCUSSION

In this paper, we present a very challenging attempt at attribute-aware head swapping. This task is of great practical value especially for the cases where the source and target have significant appearance differences, such as hairstyle, skin color and face shape. We solve this problem by dividing it into three sub-tasks and propose three key components respectively. Both quantitative and qualitative comparisons demonstrate the superiority of our method over existing state-of-the-art methods. Although there are still limitations of our method, for example, we need a source video and target video rather not a single image for head and background modeling. However, this paper is still a great try along this direction and we hope it will inspire more promising works.

5. REFERENCES

[1] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Niebner, "Face2face: Real-time face capture and reenactment of rgb videos," pp. 2387–2395, 2016.

[2] Yuval Nirkin, Yosi Keller, and Tal Hassner, "Fsgan: Subject agnostic face swapping and reenactment," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 7184–7193.

[3] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge, "Simswap: An efficient framework for high fidelity face swapping," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2003–2011.

[4] Zhiliang Xu, Xiyu Yu, Zhibin Hong, Zhen Zhu, Junyu Han, Jingtuo Liu, Errui Ding, and Xiang Bai, "Facecontroller: Controllable attribute editing for face in the wild," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, vol. 35, pp. 3083–3091.

[5] Michail Christos Doukas, Mohammad Rami Koujan, Viktoriia Sharmanska, Anastasios Roussos, and Stefanos Zafeiriou, "Head2head++: Deep facial attributes re-targeting," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 1, pp. 31–43, 2021.

[6] Yuhao Zhu, Qi Li, Jian Wang, Cheng-Zhong Xu, and Zhenan Sun, "One shot face swapping on megapixels," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 4834–4844.

[7] Ang Li, Jian Hu, Chilin Fu, Xiaolu Zhang, and Jun Zhou, "Attribute-conditioned face swapping network for low-resolution images," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 2305–2309.

[8] Xuanhong Chen, Bingbing Ni, Yutian Liu, Naiyuan Liu, Zhilin Zeng, and Hang Wang, "Simswap++: Towards faster and high-quality identity swapping," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[9] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong, "Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 0–0.

[10] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 325–341.

[11] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo, "Maskgan: Towards diverse and interactive facial image manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5549–5558.

[12] Kyle Genova, Forrester Cole, Aaron Maschinot, Aaron Sarna, Daniel Vlasic, and William T. Freeman, "Unsupervised training for 3d morphable model regression," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[13] Phillip Isola, Junyan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-image translation with conditional adversarial networks," pp. 5967–5976, 2017.

[14] Justin Johnson, Alexandre Alahi, and Li Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711.

[15] Dongdong Chen, Lu Yuan, Jing Liao, Nenghai Yu, and Gang Hua, "Stylebank: An explicit representation for neural image style transfer," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1897–1906.

[16] Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley, "Color transfer between images," *IEEE Computer graphics and applications*, vol. 21, no. 5, pp. 34–41, 2001.

[17] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1–11.

[18] Ping Wang, Kunlin Liu, Wenbo Zhou, Hang Zhou, Honggu Liu, Weiming Zhang, and Nenghai Yu, "Adt: Anti-deepfake transformer," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 2899–1903.

[19] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu, "Multi-attentional deepfake detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 2185–2194.

[20] Harold Abelson, Gerald Jay Sussman, and Julie Sussman, *Industrial color differences evaluations*, Commission Internationale de L'Eclairage, Austria, 1995.