ATTRIBUTE-CONDITIONED FACE SWAPPING NETWORK FOR LOW-RESOLUTION IMAGES

Ang Li[1]*, *Jian Hu*[2]*, *Chilin Fu*[1], *Xiaolu Zhang*[1], *Jun Zhou*[1]†

[1]Ant Group, China
[2]Queen Mary, University of London, UK
{liang268038,chilin.fcl}@antgroup.com, {yueyin.zxl,jun.zhoujun}@antfin.com, jian.hu@qmul.ac.uk

ABSTRACT

Deep learning based face swapping technologies have opened new frontiers for entertainment industries while pose novel threats to identity security. Applying face swapping to real-world products, as well as defending against its misuse, rely on the capacity to generate high quality face swapped images from realistic scenarios where high resolution images are hard to come by. To this end, we need to address the drawbacks of existing methods, especially on their lacking on maintaining the detail attributes and their dependency on high resolution images as inputs. In this paper, we propose a novel Attribute-Conditioned Face Swapping Network (AFSNet) to preserve attributes and handle low resolution images. Specifically, we use an Image Enhancement Network (IEN) to restore high resolution images from low resolution images and a Face Exchange Module (FEM) to swap the faces. In the FEM, we improve the fidelities of the generated images by using a novel multi-domain feature fusion module (MDFFM) to integrate the identity feature, context feature, IEN feature, and attribute vector to obtain the final image. We also design an attribute transfer loss to promote the consistency of the attributes such as beards and youth between the source and swapped images. The experiments demonstrate our method's superior performance compared with the state-of-the-art methods.

Index Terms— Face Swapping, face editing, face SR, deepfake, deep learning

1. INTRODUCTION

Face swapping aims to swap the face from source image to the target one, keeping the same identity as the source image while maintaining the target's pose and non-face area. Because of the potential damage to payment security [1] and privacy protection [2], face swapping algorithms have received widespread attention. To train a stronger detection network[3], it is necessary to add numerous face swapping images to the training set. Specifically, not only more realistic face swapping images is imperative, in which the detailed attributes are consistent with the source image. Images from a wider source, as those generated by low-resolution or blurry images, are also essential.

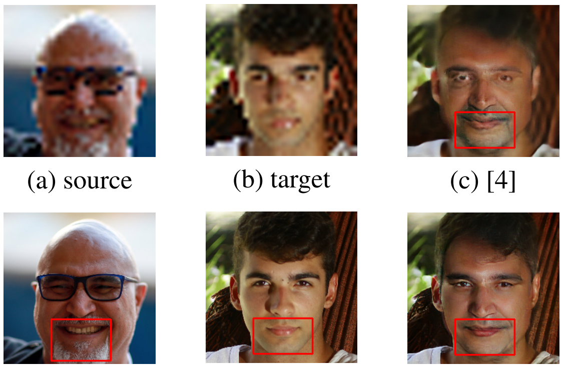

(a) source (b) target (c) [4]
(d) source (e) target (f) [4]

Fig. 1. Example of face swapping images. The source has a white mustache, but the result has a black one. The attribute is not consistent.

Face swapping algorithms can be divided into two categories: 3D-Based methods and GAN-based ones. In 3D-Based methods, Volker et al. [5] use an algorithm to predict the 3D shape, texture and scenes. However, it still needs manual interaction. Thies et al. [6] propose to reconstruct the face shape by a new global non-rigid model-based bundling method. GAN [7][8] advances the development in the fields of image generation and restoration. Similarly, GAN has also achieved good performance in the territory of face swapping. FSGAN [9] proposes a multi-step face-swapping algorithm that disassembled swapping into reenactment, swapping, and blending. FaceShifter [4] presents a high-fidelity identity swapping algorithm, where a novel AAD Layer is utilized to fuse features of target and identity of the source. Besides, it handles facial occlusions by a novel HEAR-Net. However, existing face swapping algorithms still have two problems. On the one hand, all methods assume that the images are of high resolution, and the face swapping models trained on high-resolution images do not work well for low-resolution images. On the other hand, some face attributes after swapping are not consistent with source images. As shown in

*equal contribution
†corresponding author

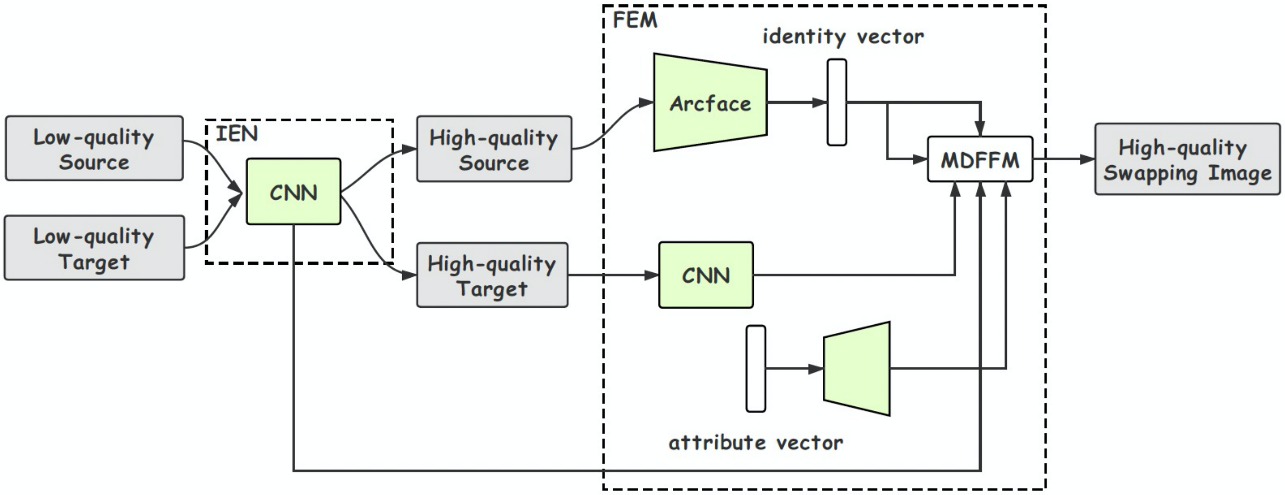


Fig. 2. Architecture of our AFSNet. Low-resolution source and target images are sent to Image Enhancement Network(IEN) to get high-resolution images. Then Face Exchange Module(FEM) takes them as input to get the swapping image, in which a novel multi-domain feature fusion module(MDFFM) is utilized to mix features from the IEN, source and target images.

Figure 1, the man in the source image has a white mustache, while the boy in the target one does not. After swapping the face, the attribute of the mustache should be consistent with the source. However, existing methods can not achieve this. If someone is familiar with the real face in the source image, it is easy to tell that the image is a fake based on the differences in the attributes.

To solve these difficulties, we put forward a novel two-stage framework named attribute-conditioned face swapping network(AFSNet) for low-resolution images. It contains two parts, including Image Enhancement Network(IEN) and Face Exchange Module(FEM). IEN is responsible for restoring high-resolution images from low-resolution images while FEM focuses on embedding the source identity to the target image. It is hard to fully interact the cross domain features by directly connecting the two modules, since the structure of features from both modules may be destroyed by simple combination of features. Hence, based on feature combination, we further retained the intermediate features of the IEN and utilized them in FEM. Moreover, a novel multi-domain feature fusion module (MDFFM) is developed to better integrate the feature from the identity feature, context feature, IEN feature, and attribute vector. Meanwhile, we propose an Attribute Transfer Loss that captures finer grained attributes and maintains consistency to adjust the attributes correctly. Specifically, our main contributions in this work are fourfold:

- Firstly, we introduce a novel two-stage framework named attribute-conditioned face swapping network (AFSNet) for low-resolution images. It is the first face swapping network which can control attributes and designed for low-resolution images. It contains two modules, Image Enhancement Network(IEN) and Face Exchange Module(FEM).
- Secondly, a novel Multi-Domain Feature Fusion Module(MDFFM) is introduced to mix the feature from the identity feature, context feature, IEN feature, and attribute vector.
- Thirdly, to control the face attributes more accurately, we present an Attribute Transfer Loss to constrain the attributes of final result to be consistent with the source image.
- Lastly, experiments on different datasets show that our model outperforms state-of-the-art methods, both quantitatively and qualitatively.

2. PROPOSED METHODS

2.1. Problem Definition

Given a low-resolution source image s, a low-resolution target image t, our goal is to obtain a high-resolution face swapping image i. Our proposed two-stage AFSNet is shown in Figure 2. It consists of two modules: Image Enhancement Network(IEN) and Face Exchange Module(FEM), where a novel multi-domain feature fusion module(MDFFM) is utilized to fuse features from the identity feature, context feature, IEN feature, and attribute vector. Attribute Transfer Loss is designed to constrain the attributes of the result image to be consistent with the attributes of the source one.

2.2. Image Enhancement Network(IEN)

Our Image Enhancement Network(IEN) aims to restore the high-resolution image from the low-resolution one. We select a standard Encoder-Decoder network here. Specifically, the encoder contains three identical blocks. The block is composed of a 3x3 convolutional layer and three Resblocks[10]. The stride of the latter two 3x3 convolutional layers is 2. The goal is to change the spatial resolution of input features to 1/2, reducing the amount of calculation, and expanding the receptive field. The decoder and encoder are mirrored network, but the 3x3 convolutional layer is replaced with a 4x4 transposed convolution in the decoder. Besides, we also utilize skip-connection in both encoder and decoder, which can benefit the gradient transfer and network convergence greatly. Here, the loss function is the Mean square error(MSE), using the output image and the ground truth.

2.3. Face Exchange Module(FEM)

The Face Exchange Module(FEM) is designed to swap the identity from the high-resolution source image to the target one.

We propose a novel Multi-Domain Feature Fusion Module (MDFFM) as shown in Figure 3 to fuse four different features, including the identity feature, the context feature, the attribute vector and the IEN feature. Specifically, the identity feature is an embedding to represent the identity of one face, we use a pretrained ArcFace [11] to extract the identity vector of the source image. Secondly, the context feature obtained by CNN refers to the non-face area and the pose features of the target facewhich are beneficial to keep the swapping pictures look as natural as possible. The attribute vector is a multi-hot vector to represent the characteristic attributes, which can capture correlations between different properties. In this paper, we select mustache, eye bags, gender, and youth as the attributes of the face. The ground truth label is provided in the CelebA [12] dataset. We pretrain a classification network to extract the attribute vector of the source image. Finally, the IEN feature is the intermediate features from IEN, which has a positive effects on the swapping images.

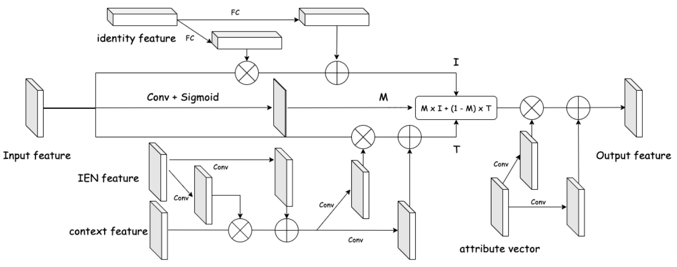


Fig. 3. Architecture of our proposed MDFFM.

Furthermore, we found that simply combining attribute feature can not completely correct the attributes. As the beard shown in Figure 1, the source image has a white beard, while attribute editing can only handle black beard. If we further analyze the process of face exchange, it is not hard to perceive the attribute vector is a completely high-level representation, while the source image is almost not involved in face exchange except for the identity. Therefore, we develop an Attribute Transfer Loss to deal with this issue. We conduct attribute training on the classification network, and use the final output of the convolutional layer as the corresponding feature of the attribute, then Attribute Transfer Loss is introduced to restrict the consistency of the features of the swapped images with those of the source images.

The final overall loss function includes GAN loss, attribute classification loss, Attribute Transfer Loss, identity similarity loss and reconstruction loss.

$$L_{GAN} = \min_G \max_D E_{x\sim P_r}[\log D(x)] + E_{z\sim P_g}[1 - \log D(z)] \tag{1}$$

$$L_{CE} = \sum_n -(y_n \log x_n + (1 - y_n) \log(1 - x_n)) \tag{2}$$

$$L_{at} = ||p(I_{source}) - p(I_{swapped})|| \tag{3}$$

$$L_{cos} = 1 - cos(q(I_{source}), q(I_{swapped})) \tag{4}$$

$$L_{rec} = ||I_{swapped} - I_{source}||_2 \tag{5}$$

$$L_{swap} = \gamma_1 * L_{GAN} + \gamma_2 * L_{CE} + \gamma_3 * L_{AT} + \gamma_4 * L_{cos} + \gamma_5 * L_{rec} \tag{6}$$

where L_{cos} is the cosine similarity loss, cos is the cosine similarity, $p(.)$ is the pretrained classification network for attributes, $q(.)$ is the pretrained ArcFace network, I_{source} is the source image, $I_{swapped}$ is the swapped image. L_{at} is the proposed Attribute Transfer Loss, L_{rec} is the reconstruction loss when the source image is the same as the target image. L_{swap} is the total loss function of our face swapping network. γ_1, γ_2, γ_3, γ_4 and γ_5 are the weights to balance all losses. In our experiments, we set 1.0, 1.0, 0.001, 1.0 and 100.

3. EXPERIMENTS

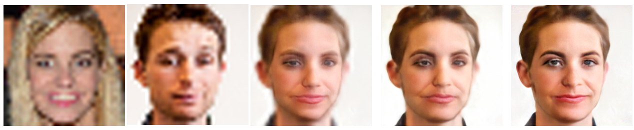
In this section, we first introduce the experimental settings in Section3.1, including datasets and parameters setting. Then, we present the ablation study and analysis the contribution of different components. Finally, we provide a comparison with other methods.

3.1. Experimental Settings

CelebA [12] is a popular face swapping dataset with 40 face attributes annotations per image. We select four attributes including mustache, eye bags, gender and youth; Then we train an attribute classification network based on them, so that given an image, we can get the attribute vector. In our experiments, we use bicubic to downsample the image to 1/8 to get the low-resolution and high-resolution training pair. To train our AFSNet, we use CelebA-HQ [13] and FFHQ [14] datasets. 10000 pair images and 5000 pair images are selected as the training and testing set respectively. There is no overlap between the training and testing sets. All face images are aligned and resized to 256x256x3 using [15]. We set γ_1, γ_2, γ_3, γ_4 and γ_5 as 1.0, 1.0, 0.001, 1.0 and 100.

3.2. Ablation Study

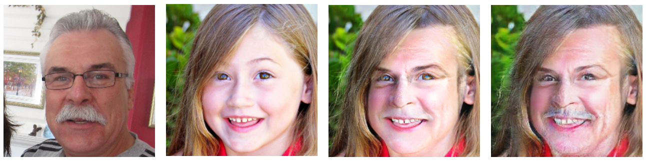
We conduct an ablation study to verify the effects of different modules. Firstly, IEN aims to obtain the high-resolution image from the low-resolution one. As shown in Figure 4(c) and 4(e), the quality of the swapping image deteriorates significantly with only FEM. One question is that directly connecting the IEN and FEM may be enough, however, we proved that features in IEN is helpful to FEM according to the comparison between Figure 4(d) and 4(e). 4(e) is more clear than 4(d), this is also one of the evidences that the novel Multi-Domain Feature Fusion Module(MDFFM) we developed is effective, because the features of IEN are fused through MDFMM.



(a) source (b) target (c) FEM(d) w/o IEN fea (e) ours

Fig. 4. Qualitative comparison of the ablation study to prove the importance of the interaction between IEN and FEM. (c) FEM means that source and target images are sent into FEM without IEN, (d) means that our methods without the intermediate feature from IEN, (e) IEN + FEM is our whole network.

We also notice that the Attribute Transfer Loss preserve the properties of the attributes. For example, figure 5(c) is generated by the model without Attribute Transfer Loss, which leaders to the incorrect mustache.



(a) source (b) target (c) without L_{at} () Ours

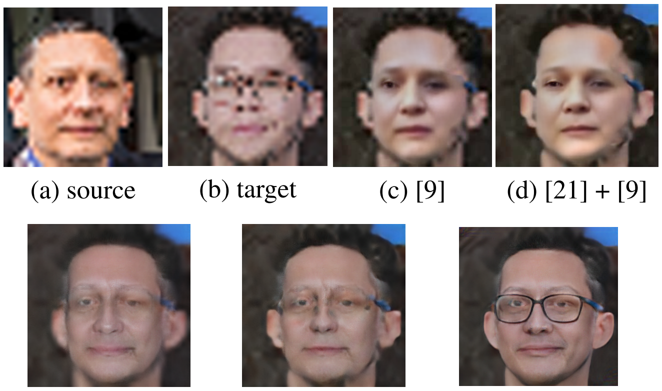
Fig. 5. Ablation study to prove the significance of the Attribute Transfer Loss. Our methods obtain more similar beard.

3.3. Comparisons with other methods

Since there are no method focusing on the low-quality face swapping, we use the state-of-the-art natural face swapping network[4, 9] in our comparison. Besides, there are many deep learning based SR methods[13, 16, 17, 18, 19, 20, 21]. We select recent SOTA face SR method[21] to add before face swapping methods. We set identity similarity, attribute accuracy, landmark similarity as our objective metrics. The identity similarity is calculating the cosine similarity of the ground truth source image and the swapped image. The attribute accuracy is calculating the accuracy of mustache, eye bags, gender, and youth between the ground truth source image and the swapped image. Landmark similarity is calculating the MSE between the landmark of the target image and swapped image. To use PSNR and SSIM, we select two images of one identity from CelebA as the source and target image. The experimental results are shown in Table 1 and Figure 6. Our AFSNet can not only surpass SOTA methods in objective metrics, but also has advantages over other methods in subjective vision.

Method	identity similarity	attribute MSE	landmark	PSNR	SSIM
[9]	0.12	0.73	9.69	22.67	0.65
[21] + [9]	0.25	0.79	6.98	25.66	0.72
[4]	0.56	0.85	10.85	23.15	0.68
[21] + [4]	0.68	0.87	7.12	30.98	0.89
Ours	**0.73**	**0.96**	**6.54**	**31.40**	**0.93**

Table 1. Quantitative comparison of experiments. Our methods are the best compared with SOTA methods in all metrics.



(a) source (b) target (c) [9] (d) [21] + [9]

(e) [4] (f) [21] + [4] (g) Ours

Fig. 6. Qualitative comparison with the state-of-the-art methods. Our results are the best.

4. CONCLUSION

In this paper, we introduce a novel attribute-conditioned face swapping network(AFSNet) , which is the first specific face swapping model for low-quality images which can control attributes. It contains two modules, including Image Enhancement Network(IEN) and Face Exchange Module(FEM). Firstly, through IEN, high quality images and features are obtained from the low-quality images. Secondly, in FEM, a novel multi-domain feature fusion module(MDFFM) is designed to mix the feature from the identity feature, context feature, IEN feature, and attribute vector. Finally, an attribute transfer loss is designed to constrain the details of attributes to be consistent between the source and target images. Experiments prove that our method can generate high-quality face swapping images compared with the state-of-the-art methods.

5. REFERENCES

[1] Marco Schreyer, Timur Sattarov, Bernd Reimer, and Damian Borth, "Adversarial learning of deepfakes in accounting," in *arxiv:cs.LG*, 2019.

[2] Arun Ross and Asem Othman, "Visual cryptography for biometric privacy," *IEEE transactions on information forensics and security*, vol. 6, no. 1, pp. 70–81, 2010.

[3] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo, "Face x-ray for more general face forgery detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5001–5010.

[4] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen, "Advancing high fidelity identity swapping for forgery detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5074–5083.

[5] Volker Blanz, Kristina Scherbaum, Thomas Vetter, and Hans-Peter Seidel, "Exchanging faces in images," in *Computer Graphics Forum*. Wiley Online Library, 2004, vol. 23, pp. 669–676.

[6] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner, "Face2face: Real-time face capture and reenactment of rgb videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2387–2395.

[7] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial networks," *arXiv preprint arXiv:1406.2661*, 2014.

[8] Jian Hu, Hongya Tuo, Chao Wang, Lingfeng Qiao, Haowen Zhong, Junchi Yan, Zhongliang Jing, and Henry Leung, "Discriminative partial domain adversarial network.," in *ECCV (27)*, 2020, pp. 632–648.

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

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[11] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.

[12] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang, "Large-scale celebfaces attributes (celeba) dataset," *Retrieved August*, vol. 15, pp. 2018, 2018.

[13] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[14] Tero Karras, Samuli Laine, and Timo Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 4401–4410.

[15] Jia Xiang and Gengming Zhu, "Joint face detection and facial expression recognition with mtcnn," in *2017 4th International Conference on Information Science and Control Engineering (ICISCE)*. IEEE, 2017, pp. 424–427.

[16] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang, "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2015.

[17] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 136–144.

[18] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681–4690.

[19] Maitreya Suin, Kuldeep Purohit, and AN Rajagopalan, "Degradation aware approach to image restoration using knowledge distillation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 2, pp. 162–173, 2020.

[20] Xingqian Xu, Zhangyang Wang, and Humphrey Shi, "Ultrasr: Spatial encoding is a missing key for implicit image function-based arbitrary-scale super-resolution," *arXiv preprint arXiv:2103.12716*, 2021.

[21] Jichun Li, Bahetiyaer Bare, Shili Zhou, Bo Yan, and Ke Li, "Organ-branched cnn for robust face super-resolution," in *2021 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2021, pp. 1–6.