Intelligent Automation & Soft Computing

Tech Science Press

DOI: 10.32604/iasc.2023.037270
Article

FSA-Net: A Cost-efficient Face Swapping Attention Network with Occlusion-Aware Normalization

Zhipeng Bin[1], Huihuang Zhao[1,2,*], Xiaoman Liang[1,2] and Wenli Chen[1]

[1]College of Computer Science and Technology, Hengyang Normal University, Hengyang, 421002, China
[2]Hunan Provincial Key Laboratory of Intelligent Information Processing and Application, Hengyang, 421002, China
*Corresponding Author: Huihuang Zhao. Email: happyzhh@hynu.edu.cn
Received: 28 October 2022; Accepted: 04 February 2023

Abstract: The main challenges in face swapping are the preservation and adaptive superimposition of attributes of two images. In this study, the Face Swapping Attention Network (FSA-Net) is proposed to generate photorealistic face swapping. The existing face-swapping methods ignore the blending attributes or mismatch the facial keypoint (cheek, mouth, eye, nose, etc.), which causes artifacts and makes the generated face silhouette non-realistic. To address this problem, a novel reinforced multi-aware attention module, referred to as RMAA, is proposed for handling facial fusion and expression occlusion flaws. The framework includes two stages. In the first stage, a novel attribute encoder is proposed to extract multiple levels of target face attributes and integrate identities and attributes when synthesizing swapped faces. In the second stage, a novel Stochastic Error Refinement (SRE) module is designed to solve the problem of facial occlusion, which is used to repair occlusion regions in a semi-supervised way without any post-processing. The proposed method is then compared with the current state-of-the-art methods. The obtained results demonstrate the qualitative and quantitative outperformance of the proposed method. More details are provided at the footnote link and at https://sites.google.com/view/fsa-net-official.

Keywords: Attention face-swapping; neural network; face manipulation; identity swap; image translation

1 Introduction

Face swapping is the process of substituting the face of a given person in a video with the face of another person while preserving facial attributes such as the facial pose, background lighting, and face silhouette. The main challenge of face swapping is how to generate a photorealistic and attribute-consistency face. Korshunova et al. [1] proposed a fast face-swapping method based on convolutional neural networks, in which a generative adversarial network is used to train the model. Despite the effectiveness improvement provided by the existing methods, they usually use complicated model architectures [2,3] and tremendous loss functions to change the face identity. In addition, the face identity is an important feature of a person, and thus changing it is a challenging task. Moreover,

changing the face identity usually causes megapixels change, while no spatial information can be given due to the lack of generalization ability. Therefore, the existing studies mainly focus on texture-based methods [4] or GAN-based methods [5] that mainly require human manual operations or canonical masks. The photorealistic image is trained by a long short-term memory neural network [6] to synthesize the attribute-consistency face. Face swapping methods [7,8] have been tried out in the film and television industries over the years, and the ones that are now in use are often complex and time-consuming computer graphics techniques. They need to be handled carefully on site and with a lot of frame-by-frame animation and post-processing by experts in digital effects. Only recently have the techniques improved to the point that moviemakers are increasingly prepared and display close-ups of computer-generated actors. Jon Hamm in Black Mirror and Chris Evans and Paul Rudd in Captain America are two examples.

Even though the swapping results are excellent, it costs extensive annotations and takes many hours of training to generate only a few seconds of footage. Deep-learning methods [9,10] for face swapping, as opposed to these computer-graphics methods, have garnered a lot of attention recently. These methods provide a pipeline of intelligent, efficient-driven face swapping.

The emergence of deep learning makes the video editing no longer rely on professional editing tools such as Adobe After Effects (AAE) and DaVinci Resolve (DR), or complex professional skills and extremely high-end hardware set-up. The latent code can assist the semi-automatic training process through a user-friendly graphical interface. The application of the neural-based technology [11] and the popularization of GPU have been accelerating the popularization of the deepfake technology. In this work, we present FSA-Net to generate high-fidelity, photo-realistic, and perceptually coherent face swapping. We achieve this through the following key contributions:

- FSA-Net is proposed to generate photorealistic and attribute-consistency faces. The latter contains an RMAA framework for handling facial pose and expression occlusion flaws by introducing the swin transformer to effectively integrate the source face into the target image, and a novel stochastic Error Refinement (SRE) module to further refine the contours for generating high-fidelity swapped results.
- Two novel loss functions are proposed: adversarial loss for training occlusion aware network with smaller steps, and attributes preservation loss for seamlessly and more accurately integrating the identity-consistency face into the target face.
- The Bountiful experiments demonstrate that the proposed method for generating high-fidelity faces qualitatively and quantitatively outperforms the state-of-the-art methods.

The remainder of this paper is organized as follows. In Section 2, some related works are introduced. The proposed approach is detailed in Section 3. The experimental results are presented in Section 4. Finally, the conclusions are drawn in Section 5.

2 Related Works

3D model and optical flow-based method. The 3D face swapping method proposed by Rahman et al. [12] relies on a monocular camera to reconstruct the 3D face model and map the distortion parameters of the source face to the target face. Face2Face [13] feeds the source image and target image into a 3DMM model, combining 3D transfer to generate attribute consistency swapped faces. Naruniec et al. [14] proposed a 3D method for automatic face replacement in videos, which does not require lots of manual operations and hardware acquisition, but only a single camera video, by tracking the face in the video with a 3D multilinear model and imitating the source face to the target face with

the corresponding 3D shape. Cai et al. [15] designed a system to change the face and thus match the mouth movements using a high-quality 3D face capture technique. Rossler et al. [16] performed segmentation to facilitate face swapping by estimating the 3D face shape from the face segmented by the network, and finally fusing the two aligned sources and targeting 3D face shapes.

Style transfer-based and GAN-based method. In contrast to the previous methods, the recent pretrained model uses face segmentation to obtain realistic results during the conversion phase, which is trained with a combination of photometric reconstruction loss and adversarial loss. Zhang et al. [17] used a multi-cross architecture to preserve the target attribute by applying a perceptual loss. Karras et al. [18] proposed an image-to-image interpolation network to convert latent code of faces into target embedder, while the latent code assists the semi-automatic training process through an encoder-decoder structure. Fdftnet [19] used an automatic training of end-to-end transfer models to synthesize high-quality results. It used a cycle consistency loss to preserve the face shape of source and target images.

Neural transfer-based methods. The neural transfer-based methods allow the generator to continuously learn the characteristics of the target image data, including the rendering network, which is a pre-trained combination of photometric reconstruction loss and adversarial loss. During the implementation procedure, they apply the patch-based GAN loss to the neural-based method [20]. Many studies tackled the facial expression transfer, involving the evaluation and processing mechanism of facial expressions, the influencing factors of face recognition [21], and the facial expression recognition emotion type effect.

Graphics-Based Method. FaceSwap [22] is a graphics-based face swapping method that first acquires face keypoints, and then renders the their locations through a graphics model, which continuously reduces the difference between the target shape and the keypoint location. Finally, it blends the images of the rendered model and acquires the final image using color correction techniques. In the graphics segmentation network of the identity-consistency face swapping techniques, a source domain and a target domain are used to generate identity-consistency face, and the data in the target domains can be directly considered as the generated results.

3 Proposed Method

3.1 Pipeline of the Workflow

In this study, the FSA-Net is proposed for photorealistic face swapping. The ability of state-of-the-art facial swapping architectures to generate high-fidelity faces is leveraged. Two novel frameworks are first proposed: Reinforced Multi-Aware Attention (RMAA) and Stochastic Error Refinement (SER).

Given an image with 3 channels, the latent inversion represents tokenizing the source input face (Fig. 1). It is then fed into latent representation reconstruction guided by an optical flow harmonization, generating all the style views that produce the corresponding offsets with encoding and decoding that are compressed by latent spatial backbone, in order to generate the embedding vectors. Afterwards, the RMAA framework fuses the identity features into the embedding consistency layer with the source image and target image to generate the photorealistic results. An overview of FSA-Net is presented in Fig. 1.

An elaboration of this complicated structure consists in the skip pooling connection, max pooling, and a convolution block. The polling connection is used to connect along the longitudinal channel, and then feed into the channel dimension. It is blended by the facial spatio-temporal smooth loss function after the convolution operation. In the calculation process, pixels of the unknown area are

selected as the pixel grid, and the foreground area and the background are used. The contrast of the area determines the threshold. The union of all the masks of each seed pixel is used to obtain the total growth area of all the pixel seeds, that is, the entire foreground area. The accurate foreground area is then obtained by determining their intersection. The mask only grows in the foreground area, and does not grow in the background area. In addition, the unknown area that is incorrectly classified as a certain foreground area can be corrected by intersection. Therefore the transparency mask map is finally solved.

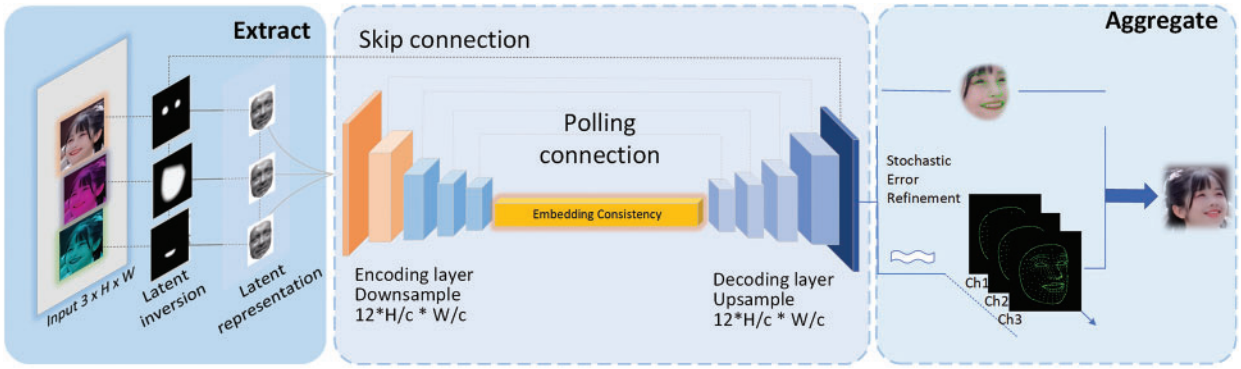

Figure 1: Overview of FSA-Net. The framework has three parts. The first part on the left is for extracting attention maps from the source image frames. In the second part, the RMAA module learns the identity attributes sent from the left input feed. The aggregate part combines three channels appended with a canonical mask to generate photorealistic results

Before integration, batch normalization is performed in a depth feature map with relatively concentrated image information and low resolution. In the up-sampling process, the spatial domain information is conducive to the restoration of target images. Finally, a fully connected layer network structure is used to classify each pixel.

3.2 Reinforced Multi-Aware Attention Framework

The framework contains one encoder and one decoder. The middle layer connects the encoding and decoding smooth layers, so that the network can be forced to learn and summarize the high-level semantic concepts of the training samples with a distribution embedding. For the latent space, existing methods usually use multiple encoders or decoders to manipulate the encoding in order to affect the output. If the encoder and decoder are symmetric and the network is trained with an objective, the network output is the reconstruction result.

The earliest replacement-based studies simply replace the pixels one by one, which requires a high degree of picture angle and human posture. However, the 3D-based methods use trainable models to deal with the pose problem of pictures. These models are very effective in inferring the texture of faces and reconstructing portraits of people. However, they hardly take into account the material gap between pictures when performing face swap, such as facial occlusion, ambient lighting, and image styles. The region-aware GAN is introduced to repair the distorted occlusions. The structure of the latter is presented in Fig. 2.

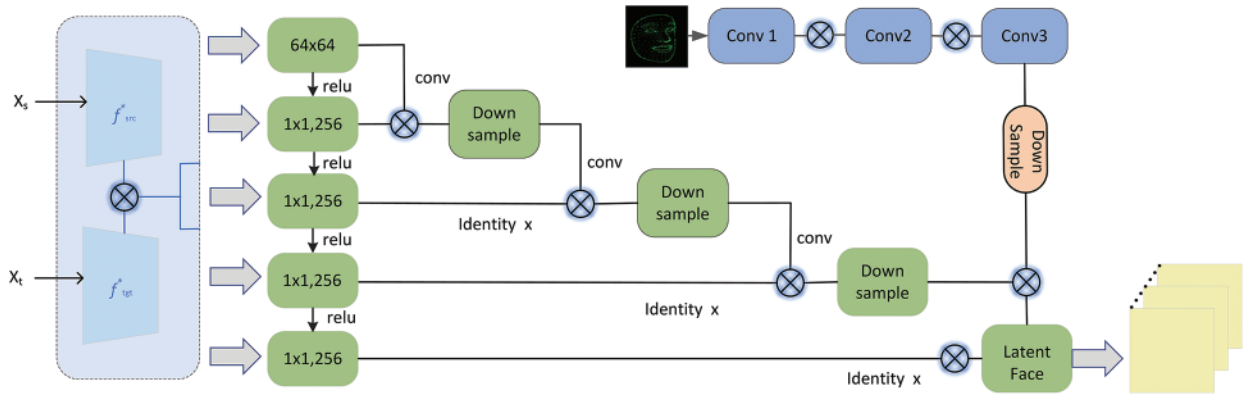

Figure 2: Structure of RMAA. It is composed of a series of residual blocks that integrate the gained features in order to refine the preprocessed image. The identity loss x is leveraged to generate results having the same identity as the source face

The RMAA structure is mainly divided into 2 sections. The left part consists of autoencoder networks that learn image features in an unsupervised manner. They are usually divided into encoders and decoders. The encoder encodes the image into latent variables, and the decoder reconstructs the latent variables into the original data. The autoencoder is a multi-layer convolutional neural network that compresses the input RGB color channel image through layer-by-layer convolution and compresses it into a small latent space vector z.

Given the source face X_s and the target face X_t, the face identity transformation fits the source face to the target face. The identity x is a deconvolution operation that reverses vector z layer by layer. Finally, an RGB channel image having the same size as the input image is generated.

3.3 Stochastic Error Refinement (SER) Module

An SER module is proposed to finally refine the facial occlusions. In contrast to the modules in FaceSwap, the SER module performs the embedding of two separate facial identities into an autoencoder, and the two autoencoders share the encoder gradients, as shown in Fig. 3. The training process consists in first encoding the facial images of two people into a latent space vector, and then reconstructing the input images through their respective decoders to minimize the reconstruction error. Inspired by the study of Nirkin [23], the SER module is used to generate the fine-grained face image.

The SER module first inputs a latent image, and its output should be infinitely close to the real samples. The input of the discriminator is the real samples and the outputs of the generator are alternately trained in a zero-sum game. The discriminator is trained to maximize the discrimination error, and the generator is trained to minimize it, so that the discriminator is able to distinguish the generated samples from the real samples. The output of the generator distribution is consistently covering the whole face region. Moreover, in contrast to the existing methods, the SER module can generate high-resolution images and keep more identity-consistency to various input scenarios. Let $F_{s,t}$ the intermediate result and ΔS_t the stochastic error, it is then fed into a parallel-net like structure to obtain the refined images, which is expressed as:

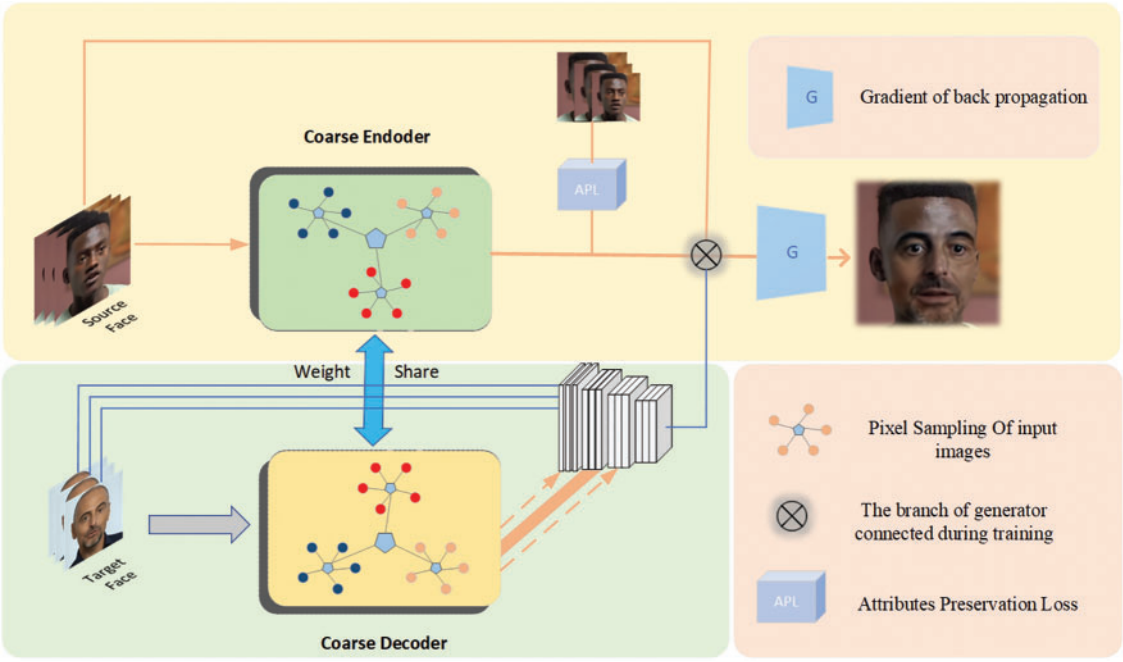

Figure 3: Overview of the stochastic error refinement (SER) module, which is based on an encoder-decoder structure with weight share framework

$$F_{s,t} = SER - \sqrt{\left(\hat{F}_{s,t}, \Delta S_t\right)}, \tag{1}$$

The matching embedding integration is given by:

$$\mathrm{C}^{\Phi} = \sum_{i=1}^{n} \frac{1}{2} \left\{\beta_{att}^{i} + \log\left(1 - \beta_{att}^{i}\right)\right\}, \tag{2}$$

where β_{att}^{i} denotes the attributes embedding on the i-th feature level and C^{ϕ} represents the consistency attribute extracted from the attributes embedding layer.

In the proposed framework, since there is no ground truth reference in the face swapping process, the source target image is used to replace its position. The first upper layers are then removed, and the last few layers are only used to calculate the weak feature matching loss, which can be written as:

$$\mathrm{F}_{om}\left(P_t\right) = \left\{\mathrm{y}_{st}^{1}\left(P_t\right), y_{\mathrm{st}}^{2}\left(P_t\right), \cdots y_{\mathrm{st}}^{n}\left(P_t\right)\right\}, \tag{3}$$

where $\mathrm{F}_{om}\left(P_t\right)$ represents the object module map and $\mathrm{y}_{st}^{1\ldots n}$ is the number of attribute levels.

Note that the embedding network encodes the variant input predicted frames.

3.4 Overall Loss Function

The extracted facial attributes such as cheek, mouth, expression, background lighting, and spatial polling information identity, are fed into the face model for binary classification training. The backbone network consists of four layers, while each one consists of a different number of bottleneck

modules. More precisely, the attributes embedding as the feature maps generated from the decoder are defined as:

$$\mathcal{L}_{WFM} = \mathcal{L}_{rec}\left(I_{last}, \tilde{I}_{upper}\right) + \mathcal{L}_{rec}\left(I_{last}\right) + \mathcal{L}_{gen}, \tag{4}$$

where I_{last} represents the last layer module map, $\mathcal{L}_{rec}$ is the reconstruction objective layer, and $\mathcal{L}_{gen}$ denotes the generative predicted frames.

It extracts the features using self-supervised learning, while the generated swapped face and the target image should have the same attributes embedding.

Adversarial loss: A novel adversarial loss is proposed as follows. Let $\mathcal{L}_{src}$ the adversarial loss for the source sampled canonical mask, an identity preservation loss is then used to preserve the identity of the source:

$$\mathcal{L}_{src} = \mathrm{De}[\![1 - \cos\left(\mathbf{z}_{id}\left(I_{src}\right), \mathbf{z}_{id}\left(I_{tgt}\right)\right)]\!] \tag{5}$$

where cos represents the cosine similarity of two vectors.

The attributes preservation loss is also defined as $\mathcal{L}_{att}$, which represents the evaluation loss between ∂^{k}_{att} and ∂^{p}_{att}:

$$L_{\mathrm{att}}(D) = \sum_{i=1}^{K} \frac{1}{N_i} \left\| \partial^{k}_{att}\left(\mathrm{X}_s\right) - \partial^{p}_{att}\left(X_t\right) \right\|_1, \tag{6}$$

Attributes preservation loss: Inspired by the stepwise consistency loss [24], an attribute preservation loss is proposed. Given an input pair from the data source and data target, let X_s the frames from the data source and X_t the data target, then:

$$\mathcal{L}_{ap} = \begin{cases} 1 - \sin\left(\|X_s - X_t\|_2^2\right) & \text{if } X_t = X_s \\ 0 & \text{otherwise} \end{cases}. \tag{7}$$

Afterwards, two embeddings are integrated into the swapped face results, while Rt and Mt aggregate them through feature concatenations as:

$$\mathcal{L}_{\mathrm{output}} = Mt \odot \frac{1}{2}\lambda_{src}\mathcal{L}_{att} + Rt \odot \frac{1}{2}\lambda_{\mathrm{src}}\mathcal{L}_{\mathrm{rec}}, \tag{8}$$

where Mt represents the number of source subject masks, $\odot$ represents the dot product, Rt is the number of recurrent channels, and $\mathcal{L}_{\mathrm{output}}$ denotes the attribute spatial dimensions result.

4 Experiment and Comparison

4.1 Dataset and Experiment

A generated dataset of 2010 images, which contains 43 types of poses on 620 face swapping video frequencies is used. The CelebA dataset [25] contains 50,000 swapped images and 500 videos. The test dataset based on the FaceSwap open-source program is used. 16 pairs of similar portraits are selected from the latter, and each image is divided into 64 × 64 and 128 × 128 pixels. Each angle has 10 videos in the dataset, and 320 videos are finally generated by changing faces. The generated dataset contains 900 videos, of which 700, 100, and 100 are used for training, validation, and generating test results, respectively. The datasets of video face swapping are shown in Table 1. For the training process, we use lntel(R) Xeon(R) Silver 4210R CPU @ 2.40 GHz (40 CPUs),~2.4 GHz. In the current model, we use Dropout to avoid the model overfitting problem, which is a regularization method that is used to randomly disable neural network units. It can be implemented on any hidden or input layer, but not

on the output layer. This method can eliminate the dependence on other neurons so that the network can learn independent correlation, so as to avoid overfitting the model.

Table 1: Datasets used in the experiments

Datasets	Year	Total videos	Source	Participants consent	Tools
Celeb-DF	2021	1203	Youtube	N	FakeApp
FaceForensics	2018	2008	Youtube	N	Face2Face
Deepfake-TIMIT	2019	620	Vid	N	Faceswap-GAN
FaceForensics++	2019	997	Vid and Youtube	N	Faceswap DeepFake Face2Face NeuralTextures
DeepFakeDetection (Part of FaceFornsics++)	2021	3363	Actors	Y	A refined version of the DeepFake
DFDC preview Dataset	2020	5214	Actors	Y	Unknown
UADFV	2019	3419	Youtube	N	Face2Face

4.2 Visual Comparison

To evaluate the performance of the proposed model, it is applied to select: i) the image having the most similar identity with the source face; ii) the one sharing the most similar head pose, face expression, and scene lighting with the target image; iii) the most realistic one. In each evaluation unit, two real face images, the source and the target, and six face swapping results generated by HiFace [26], MegaFS [27], DeepFakes [28], Faceshifter [29], FaceSwap [30], and the proposed model, are presented (Fig. 4).

It can be seen from Fig. 4 that the proposed method can better preserve the source face attributes and better retain the lighting conditions. We use a novel reinforced multi-aware attention module, referred to as RMAA for handling facial fusion and expression occlusion flaws, which is used to repair occlusion regions in a semi-supervised way without any post-processing. which our method is better at edge optimization and preserving its facial attributes, such as facial expression, head posture and background.

More experimental results are shown in Fig. 5. It can also be observed from Fig. 5 that the proposed method can lead to the best results.

Due to the large training set of the experiment, setting a smaller batch will not only lengthen the training time of the network model, but also be detrimental to its convergence. Under the same epoch, a larger batch size will reduce the number of iterations and reduce the detection performance of fake faces. By considering the choice of the optimizer in the experiment, it is deduced that the proposed method has a high convergence speed in HiFace. The loss value will decrease to a minimum value, and stop at the local optimal solution of the model. The generalization performance on the test set is reduced, while the convergence speed of the optimizer is relatively stable, and it can effectively solve

extreme cases such as the local optimal solution of the model. Therefore, the proposed method can better preserve the face shape of the source identity.

Figure 4: Comparison between the results obtained by HiFace, MegaFS, DeepFakes, Faceshifter, FaceSwap, and the proposed method using a male set from the Celeb-DF face images

Figure 5: (Continued)

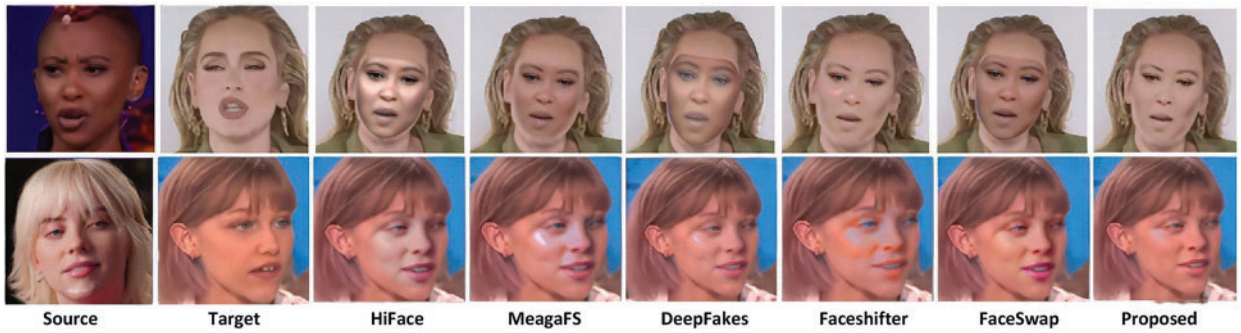

Figure 5: Comparison between the results obtained by HiFace, MegaFS, DeepFakes, Faceshifter, FaceSwap, and the proposed method using images involving long-hair. Note that the proposed method preserves more multi-level attributes

4.3 Quantitative Comparison

To further analyze the reliability of FSA-Net, the SEE is computed to assess the shape and expression estimation. It can be deduced from the obtained results that the proposed method achieves a higher score (Table 2). To evaluate the structure similarity, the SSIM method is used to analyze the similarity between the two images generated by the six methods. In fact, the SSIM is a local quality score that combines local image structure, brightness, and contrast. The structures in this metric are patterns of pixel intensities, primarily among neighboring pixels, after linearization for luminance and contrast. To evaluate the pose and landmark, the evaluation criterion presented in [31] is used. In addition, a pose estimator is used to estimate the head pose, and a 3D face model is used to retrieve expression vectors. It can be seen from Table 2 that the proposed method is advantageous in terms of expression preservation, while it is comparable to the other methods in pose preservation marks.

Table 2: Quantitative analysis results on FaceForensics++

Method	PSNR↑	SSIM↑	Pose marks↓
HiFace	0.76 ± 0.07	0.61 ± 0.04	5.12 ± 1.07
MegaFS	0.72 ± 0.09	0.73 ± 0.04	5.06 ± 0.97
DeepFakes	0.72 ± 0.11	0.61 ± 0.04	5.33 ± 1.21
FaceShifter	0.79 ± 0.05	0.58 ± 0.03	5.00 ± 0.97
FaceSwap	0.78 ± 0.06	0.51 ± 0.03	$\mathbf{4.90 \pm 0.88}$
FSA-Net	$\mathbf{0.82 \pm 0.11}$	$\mathbf{0.74 \pm 0.04}$	5.03 ± 1.01

In Table 2, the arrow key ↑ denotes that a higher score indicates a better result, ↓ denotes that a lower score indicates a better result, while the best results are marked in bold. It can be deduced from Table 2 that higher PSNR and SSIM values indicate the preservation of more details with the same structure. The pose mark results can have a better performance to get rid of the influence of source identity. In addition, FSA-Net effectively relieves the instability of GAN, and it leads to a more photo-realistic result without degradation.

4.4 Ablation Studies

In this section, the performance of each part of the proposed method is evaluated. Thus, ablation studies are conducted with three configurations of the proposed method to evaluate their impact on the performance. Q_b is then defined as the baseline method, the RMAA module is denoted by Qr, and the SER module is represented by Qs. The three configurations are denoted by $Q_b + Q_r$, $Q_b + Q_s$, and $Q_b + Q_r + Q_s$. The qualitative comparison results are shown in Fig. 6.

Figure 6: Ablation experiments. From left to right: the source face, the target image frame, the image generated without RMAA, the image generated without SER module, and the combination of RMAA and SER

To analyze the influence of the RMAA module, experiments are performed on the source identity set. It is deduced that the RMAA effectively relieves the instability of GAN, through which a more photo-realistic result without degradation is then achieved. In addition, it can be seen from the pose marks metrics and PSNR scores in Table 2 that the target poses and expressions are best retained with the full structure, and the error differences are not extreme. To evaluate the effectiveness of SER, Qs and Qr are added to the pipeline. It can be observed that the inpainting and blending generators led to the best results, while efficiently preserving the pose and expression similarly. Finally, there is a small decrease in the SSIM, due to the single networks and processing training time added to the whole pipeline.

5 Conclusion

This paper proposes an efficient face swapping method, referred to as FSA-Net, for generating high-photorealistic swap images. The proposed method can also solve the occlusion of faces in different occluded regions. This is achieved by combining the attentional feature map with a separate aggregation module. To ensure the diversity of the attention map and deal with the occluded parts, FSA-Net further adjusts the loss function on the occluded face. The experiments show that the proposed method outperforms all the baselines on multiple benchmarks. Moreover, FSA-Net chooses

relatively concise features generated by different scales to better solve the contours occlusion problem. We consider this study can make it possible to tackle increasingly difficult face-swapping challenges by greatly simplifying them. One may think about expanding the challenge scope to represent face-swapping on videos end-to-end with less effort for balancing the components and less memory utilization. Our proposed model's performance in maintaining the position and expression suffers somewhat as a result. In the future, we assume that a straightforward fine-tuning or alternative hyperparameter selection would be adequate to achieve the purpose.

Funding Statement: This work was supported by the National Natural Science Foundation of China (No.61772179), the Hunan Provincial Natural Science Foundation of China (No.2020JJ4152, No.2022JJ50016), the science and technology innovation Program of Hunan Province (No.2016TP10 20), the Scientific Research Fund of Hunan Provincial Education Department (No.21B0649), and the Double First-Class University Project of Hunan Province (Xiangjiaotong [2018]469).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] I. Korshunova, W. Shi, J. Dambre and L. Theis, "Fast face-swap using convolutional neural networks," in *Proc. ICCV*, Venice, VN, IT, pp. 3677–3685, 2017.

[2] R. Tolosana, V. Rodriguez, J. Fierrez, A. Morales and J. O. Garcia, "An introduction to digital face manipulation," *Handbook of Digital Face Manipulation and Detection*, vol. 12, no. 4, pp. 3–26, 2022.

[3] C. Xu, J. Zhang, M. Hua, Q. He, Z. Yi *et al.,* "Region-aware face swapping," in *Proc. CVPR*, New Orleans, LA, USA, pp. 7632–7641, 2022.

[4] D. Ulyanov, L. Vadim, V. Andrea and S. Victor, "Texture networks: Feed-forward synthesis of textures and stylized images," in *Proc. ICML*, New York, NY, USA, pp. 1349–1357, 2016.

[5] Y. Huang, J. F. Xu, Q. Guo, Y. Liu and G. Pu, "FakeLocator: Robust localization of GAN-based face manipulations," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 2, pp. 2657–2672, 2022.

[6] S. Kankanamge, C. Fookes and S. Sridharan, "Facial analysis in the wild with LSTM networks," in *Proc. ICIP*, Beijing, BJ, CHN, pp. 1052–1056, 2017.

[7] J. X. Sun, X. Wang, Y. Zhang, X. Li, Q. Zhang *et al.,* "Fenerf: Face editing in neural radiance fields," in *Proc. CVPR*, New Orleans, LA, USA, pp. 7672–7682, 2022.

[8] N. Nejatishahidin, P. Fayyazsanavi and J. Kosecka, "Object pose estimation using mid-level visual representations," in *Proc. IROS*, Kyoto, Japan, pp. 13105–13111, 2022.

[9] F. Heydarpoor, M. Karbassi and S. Bidabadi, "Solving multi-objective functions for cancer treatment by using metaheuristic algorithms," *International Journal of Combinatorial Optimization Problems and Informatics*, vol. 11, no. 3, pp. 61–75, 2020.

[10] E. Javaheri, V. Kumala, A. Javaheri, R. Rawassizadeh, J. Lubritz *et al.*, "Quantifying mechanical properties of automotive steels with deep learning based computer vision algorithms," *Metals*, vol. 10, no. 2, pp. 163–175, 2020.

[11] M. Ebadi, J. Alireza and R. Ebrahimi, "Video data compression by progressive iterative approximation," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 6, pp. 189–195, 2021. https://doi.org/10.9781/ijimai.2020.12.002

[12] A. Rahman, M. J. Islam, T. Tasnim, N. Siddique, M. Shahiduzzaman *et al.,* "A qualitative survey on deep learning based deep fake video creation and detection method," *Australian Journal of Engineering and Innovative Technology*, vol. 4, no. 1, pp. 13–26, 2022.

[13] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt and M. Niener, "Face2face: Real-time face capture and reenactment of rgb videos," in *Proc. CVPR*, Las Vegas, NEV, USA, pp. 2387–2395, 2016.

[14] J. Naruniec, H. Leonhard, S. Christopher and M. W. Romann, "High resolution neural face swapping for visual effects," *Computer Graphics Forum*, vol. 39, no. 4, pp. 173–184, 2020.
[15] W. J. Cai, Y. F. Wang, J. H. Ma and Q. Jin, "CAN: Effective cross features by global attention mechanism and neural network for ad click prediction," *Tsinghua Science and Technology*, vol. 27, no. 1, pp. 186–195, 2022.
[16] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies *et al.,* "Faceforensics++: Learning to detect manipulated facial images," in *Proc. ICCV*, Seoul, SEL, Kr, pp. 1–11, 2019.
[17] K. Zhang, Z. Zhang, Z. Li and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503,2016, 2016.
[18] T. Karras, S. Laine and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. CVPR*, Long Beach, CA, USA, pp. 4401–4410, 2019.
[19] H. Jeon, B. Youngoh and S. Simon, "Fdftnet: Facing off fake images using fake detection fine-tuning network," *ICT Systems Security and Privacy Protection*, vol. 12, no. 3, pp. 416–430, 2020.
[20] P. Salehi and A. Chalechale, "Pix2pix-based stain-to-stain translation: A solution for robust stain normalization in histopathology images analysis," in *Proc. ICMV*, Tehran, TH, IR, pp. 1–7, 2020.
[21] W. Zhang and W. Wang, "Broad learning system for tackling emerging challenges in face recognition," *CMES-Computer Modeling in Engineering and Sciences*, vol. 12, no. 3, pp. 1–23, 2022.
[22] B. Peng, H. Fan, W. Wang, J. Dong and S. Yu, "A unified framework for high fidelity face swap and expression reenactment," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 6, pp. 3673–3684, 2021.
[23] G. Li, Z. Ji and X. Qu, "Stepwise domain adaptation (SDA) for object detection in autonomous vehicles using an adaptive centerNet," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 10, pp. 1–15, 2022.
[24] Z. Liu, P. Luo, X. Wang and X. Tang, "Large-scale celebfaces attributes (celeba) dataset," in *Proc. ICCV*, Santiago, ST, CL, pp. 3730–3738, 2015.
[25] Y. Nirkin, I. Masi, A. T. Tuan, T. Hassner and G. Medioni, "On face segmentation, face swapping, and face perception," *IEEE Automatic Face and Gesture Recognition*, vol. 4, no. 3, pp. 98–105, 2018.
[26] Y. Zhu, Q. Li, J. Wang, C. Z. Xu and Z. Sun, "One shot face swapping on megapixels," in *Proc. CVPR*, Nashville, NV, TN, pp. 4834–4844, 2021.
[27] Z. Xu, H. Zhou, Z. Hong, Z. Liu, J. Liu *et al.,* "StyleSwap: Style-based generator empowers robust face swapping," in *Proc. ECCV*, Milan, ML,IT, pp. 661–677, 2022.
[28] L. Li, J. Bao, H. Yang, D. Chen and F. Wen, "Advancing high fidelity identity swapping for forgery detection," in *Proc. CVPR*, Seattle, WA, USA, pp. 5074–5083, 2020.
[29] S. Mahajan, C. Ling and T. Chieh, "SwapItUp: A face swap application for privacy protection," *Information Networking and Applications*, vol. 13, no. 2, pp. 46–50, 2017.
[30] Y. Li, X. Yang, P. Sun, H. Qi and S. Yu, "Celeb-df: A large-scale challenging dataset for deepfake forensics," in *Proc. CVPR*, Seattle, WA, USA, pp. 3207–3216, 2020.
[31] P. Lukac, R. Hudec, M. Benco, P. Kamencay, Z. Dubcova *et al.,* "Simple comparison of image segmentation algorithms based on evaluation criterion," in *Proc. ICR*, Bratislava, BTL, SK, pp. 1–4, 2011.