Using Fully Connected and Convolutional Net for GAN-Based Face Swapping

Bo-Shue Lin, Ding-Wen Hsu, Chin-Han Shen, Hsu-Feng Hsiao
Department of Computer Science
National Chiao Tung University
HsinChu, Taiwan
Email: leo0507.cs06@nctu.edu.tw, aevin880713.cs06@nctu.edu.tw, fionaschs@nctu.edu.tw, hillhsiao@cs.nctu.edu.tw

***Abstract*—The lifelike results of using face swapping have contributed greatly to the research in computer vision. In this work, we extend the architecture of faceswap-GAN in order to obtain more natural results compared to the original framework. In the original architecture, the self-attention module usually converts the facial features from a source face to the target face with artificial distortion around the facial features. We use a structure of fully connected convolutional layers as a discriminator to approach the problem. The outcome can be smoother and more natural perceptually compared to the results using the original faceswap-GAN.**

Keywords—Deepfake, Generative adversarial network (GAN), fully-connected and convolutional network

I. INTRODUCTION

Deepfake was first known of faking the faces of famous movie stars in porn videos and has been banned in a lot of social media applications. It is able to swap the appearance from one person to another in multimedia contents such as images or videos with lifelike results efficiently. It is sometimes difficult to authenticate whether a video has been tampered with. The concept of deepfake is to relate the connection between two human faces. Its learning model is based on the autoencoder, which includes an encoder to process the data into features and a decoder to perform the deconvolution to generate the output as much alike to the original input as possible. The faceswap uses denoising autoencoder (Fig. 1) as the training model. The input to the encoder shall contain artificial noise. The decoder will try to recover the distorted input back to the original data without the artificial noise. The artificial noise can usually make the training model generate the relatively robust results compared to a normal autoencoder.

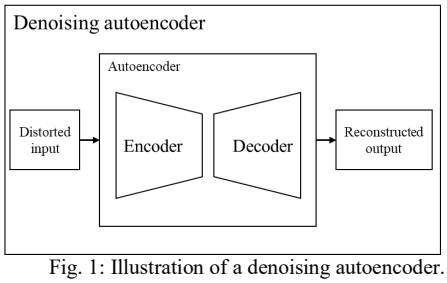

Fig. 1: Illustration of a denoising autoencoder.

For a single autoencoder, it actually processes only one specific face using feature generation (encoder) and face reconstruction (decoder). In order to build such 'connection' between two faces, the autoencoder shown in Fig. 2 uses the same encoder for both A and B to encode the input, while recovering the encoded data with different decoders for specific faces. As the instance showed in Fig. 2, the encoder relates two faces of A and B, while separate decoders are used to reconstruct different faces. After the training procedure, the model will contain three parts, an encoder for obtaining both A and B features, a decoder to reconstruct A's face, and a decoder to reconstruct B's face. A's face can be swapped to B's body, and similarly, B's face can be swapped to A's body. However, a face from a third person cannot be swapped to either A or B using the trained model.

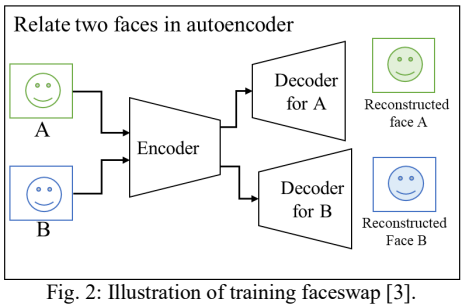

Fig. 2: Illustration of training faceswap [3].

In this paper, we extend the architecture of faceswap-GAN to obtain a more natural transformation of faces. We propose to use fully connected layers to modify the discriminator to avoid the over-processing issues resulted from the self-attention module in the original method. The following section will introduce several works related to face swapping. Then, the details of the proposed architecture will be described in section 3. The experimental results will be shown in section 4. In the last section, we will conclude our work.

II. RELATED WORKS

For the development of deepfake, the techniques of facial imagery synthesis and manipulating are the important references. We will briefly review the facial manipulating methods categorized into encoder-decoder (autoencoder) methods and the GAN-based methods. Also, the methods of deepfake detection will be reviewed in the end of this section.

A. Encoder-decoder (autoencoder) methods

Liu et al. [5] introduced featuring dual encoders and decoders based on the VAE-GAN framework, which is the developing foundation of the encoder-decoder based methods. The key idea of their work is the concept of a shared latent space enforced via tied weights in several of the layers of the encoders and decoders closest to the encoded bottleneck.

A recent work has expanded the capabilities for creating three-dimensional morphable models, allowing them to be learned from sets of 2D images using deep encoder-decoder networks [6]. Korshunova et al. [7] approached the problem of face swapping using the perspective of style transfer. They use a multiscale texture network with both content and style losses measured in a VGG-19 feature space. Yan et al. [8] explore a Y-shaped, single-encoder architecture. During training, they introduce warp distortions to the input images

978-1-7281-9396-0/20/$31.00 ©2020 IEEE

while tasking the decoders with reconstructing the undistorted images, as the core concept of a denoising autoencoder. Zhao et al. [9] used an encoder-decoder architecture with a multitask objective including face alignment and segmentation goals. Their model requires extensive labeled training data. Natsume et al. [10] employ several encoder-decoder networks. Each of them specializes in different features extracted from an input image such as binary mask, isolated face, and facial landmarks. A separate generator is then used to combine the target face with a source image. Kim et el. [11] proposed a method to synthesize the photo-real video portraits in front of general static backgrounds based on a novel rendering-to-video translation network.

B. *GAN-based facial manipulating methods*

Generative adversarial networks (GANs) have become popular for image synthesis. The general approaches that show impressive improvement in face swapping use conditional GANs [12, 13, 14]. Natsume et al. [15] compose the output using two separator networks—one for the face and one for hair. They use a GAN to verify and fine-tune the generated results. Shu et al. [16] treat face representation as a rendering problem and use a GAN to create surface normal, albedo, lighting, and alpha matte information from input images to allow for more compelling image edits. Choi et el. [17] proposed a scalable image-to-image translation model among multiple domains using a GAN architecture with the multi-task learning setting.

C. *Detection of swapped faces*

For the detection of swapped faces, several methods for detecting the inauthentic facial images and videos have been proposed. The conventional methods are mostly based on finding inconsistencies in the content. However, it is getting harder to detect due to the non-linearity and complexity of the learning process. To deal with the issue, a specific generative model [18] based on deep learning method and semantic approaches that evaluate the generated faces' realism [19] was proposed. Recently, Ciftci et el. [20] have proposed a relatively powerful fake portrait video detector based on biological signals. The method examines the signal discontinuity in spatial and temporal domain. The work has shown that spatial coherence and temporal consistency of biological signals are not well preserved in GAN-generated content. With the analysis of the signal transformations and corresponding feature sets, they utilize the findings to formulate a generalized classifier for fake content. In [20], signal maps are generated and a CNN is employed to improve the traditional classifier for detecting synthetic content.

III. THE PROPOSED METHODS

Fig. 3 shows the structure of the faceswap-GAN. The faceswap was originally developed using the autoencoder architecture, which is an unsupervised learning network that aims to compress the input to a low dimensional representation and then reconstruct it. It only takes two videos as inputs with source face and target face in each. First, the two videos will be extracted into frames and the faces are located inside a frame using a trained multi-task cascaded convolutional networks (MTCNN). Then the faces will be located as the input features of training procedure. The training of face swapping is achieved using the denoising autoencoder to obtain more robust reconstruction.

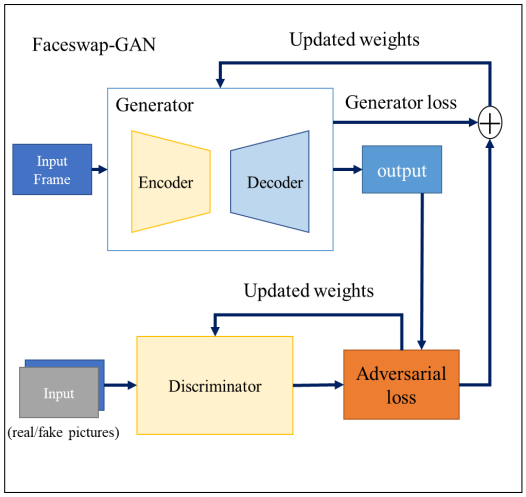

Fig. 3: Flowchart for the faceswap-GAN.

In the original version of face swapping model, the self-attention module is used to highlight the facial features to convert the features to target face precisely. It helps build a huge range of connections and remain the holistic of the object. The flowchart (Fig. 4) below shows the procedure of self-attention [4]. However, the result of the face swapping can be distorted significantly around the facial features, which results from the over-processing at a decoder.

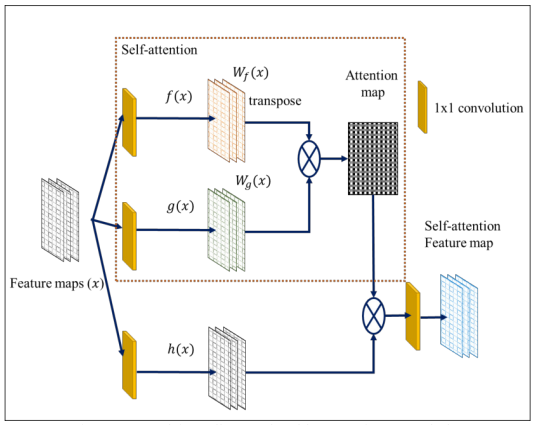

Fig. 4: Working flow of self-attention module.

To improve the results generated using an autoencoder structure, the GAN architecture has been applied. The general model is the original autoencoder with frame picture as input, it learns the facial information to generate the faked pictures. The faked pictures will be sent to the discriminator next. The discriminator will judge whether a picture is real or faked from the generator (autoencoder). Then the result of the judgement (adversarial loss) will update the generator. In addition to the adversarial loss which will affect the generator, the other losses consisting of adversarial loss, reconstruction loss, edge loss, perceptual loss and cyclic loss will also have an influence on the weights in the training process.

Fig. 5 shows the encoder-decoder structure, which is the content generator used in faceswap-GAN. The encoder consists of the combination of conv2D layers and the self-attention module from [1]. The feature map will be up-scaled using a conv2D and pixelShuffler [21] method. On the other

hand, the first process of a decoder is to upscale with several self-attention modules involved. After the third time of upscaling in the decoder, the residual block structure is used with the leaky ReLU being the activation function. Leaky-ReLU is used to avoid vanishing gradients problem. The denoising autoencoder makes the model have better ability to reconstruct a human face. After the final self-attention block, the steps of upscaling and convolution are used to match the original data size. It finally generates the mask and RGB outputs. These two outputs will be concatenated together in order to simplify the inputs for discriminator.

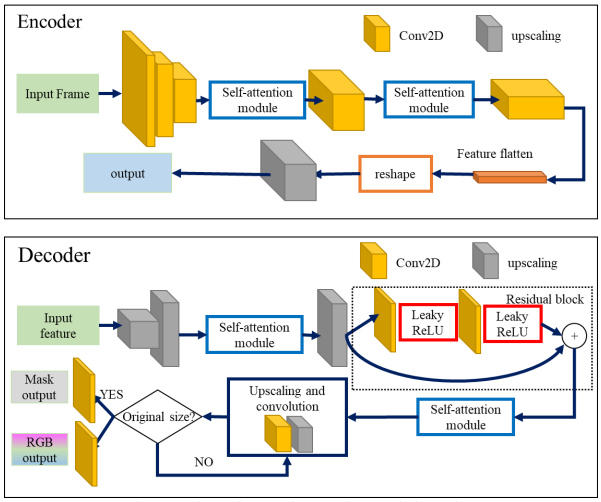

Fig. 5: Flowchart of encoder and decoder.

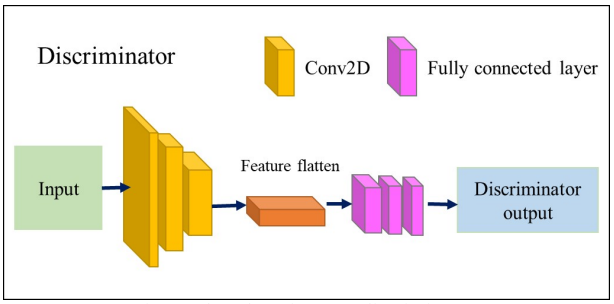

Fig. 6: Flowchart for FCC discriminator.

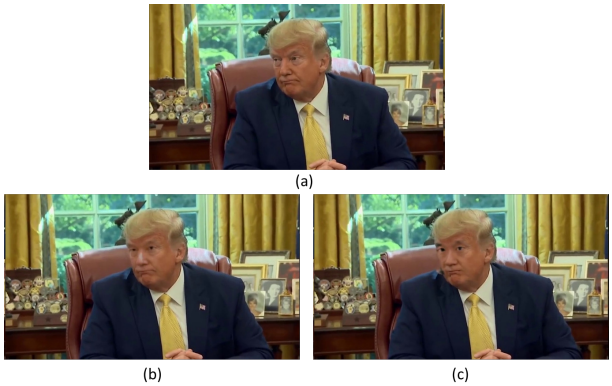

Fig. 7. (a) original frame, (b) the swapped frame using optimizer Adam, (c) using RMSprop.

For the discriminator, we adopt the discriminator in the FCC-GAN [3] structure to replace the original discriminator, which reconstructs the generator outputs with combination of 2d convolution and self-attention block. The FCC-GAN incorporates multiple deep fully connected and convolution layers shown in Fig. 6 in both the GAN generator and discriminator. A faster learning curve can be achieved compared to the conventional architecture. Images with better quality can also be produced. In our proposed method, the discriminator composes of conv2D layers and fully connected layers with Adam optimizer. As a side note, Adam optimizer is better than RMSprop in a GAN structure for face swapping. The comparison of using Adam optimizer and using RMSprop can be seen in Fig. 7.

The fully-connected layers in the discriminator have two benefits: (i) to avoid the discriminator loss from decreasing too quickly, which will make the gradient either too small or vanished. (ii) not having shared weights can help the network more independent for learning the required non-spatial dimensional mapping task. These benefits of our proposed networks are the main reason to have more natural results.

IV. Experiment Results

To show the performance of our proposed method, we demonstrate the performance on two sets of videos. The same initial parameters are used for both sets. We set the learning rate for the generator as 0.0001, and the learning rate for discriminator is 0.0002. Since the performance of the networks are subjective in nature, the visualized results are shown here for the evaluation.

Fig. 8. The source (left) and the target (right) face.

We use our model to swap two oriental faces. The source face is a famous Japanese actor, Kimura Takuya, whose facial features are highly identical. The target face is a famous male Taiwanese singer Jay Chou. In this experiment, we used a 96-second video with the actor and another 83-second video with the singer respectively in the training process.

In this experiment, we evaluate the swapped results using the videos in Fig. 8. We shall be able to notice the difference between the original frames (shown in Fig. 9-a) and the generated frames using both the original method (shown in Fig. 9-b) and our proposed method (shown in Fig. 9-c). In the original method, more facial details from the source face have been remained in the swapped results. The results using the proposed FCC discriminator show a smoother swapped face. The facial features of Kimura Takuya has been altered into the features of Jay Chou better .

Next, we do the evaluation of swapping the face in another video shown in Fig. 10. Both the generated results using the original faceswap-GAN and the proposed FCC discriminator method contain the perceptible discontinued distortion. The distortion generated using the original faceswap-GAN method already impairs most of the faces in the testing video. In the visualization result, the unnatural distortion can be noticed around the eyebrows. The double eyebrows can be observed from the Fig. 10-b. In comparison, our proposed FCC discriminator method have smoother and more natural faces. The facial features contain less over-processed distortion while we can maintain the reasonable level of quality of the swapped videos. The double eyebrows haven't appeared in the results using our method.

V. Conclusion

In this work, we have improved the faceswap-GAN using the fully connected convolutional discriminator. In the experiment results, our proposed method can generate relatively smoother and more natural faces compared to the results using the original structure. It is supposed that the fully connected layers can make the model more sensitive to the details in an image overall. For the future work, we expect to improve our work using a better discriminator that can reduce the flaws of deepfake technology further.

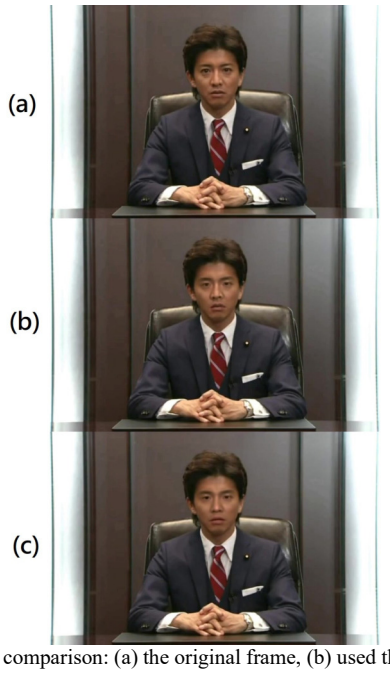

Fig. 9. Visual comparison: (a) the original frame, (b) used the faceswap-GAN, (c) proposed faceswap-GAN's generator with FCC discriminator.

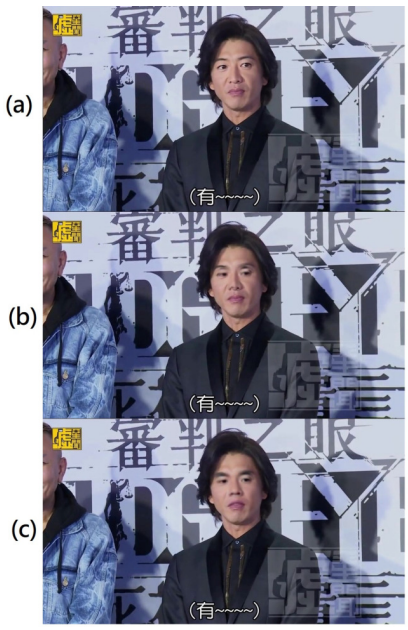

Fig. 10: Visual comparison: (a) the original frame, (b) used the faceswap-GAN, (c) proposed faceswap-GAN's generator with FCC discriminator.

References

[1] Deepfakes/faceswap, GITHUB, 'faceswap', 2018. [Online]. Available: https://github.com/deepfakes/faceswap. [Accessed: 16-Jul- 2020]

[2] shaoanlu, GITHUB, "faceswap-gan.", 2019. [Online]. Available: https://github.com/shaoanlu/faceswap-GAN [Accessed: 21 Mar 2019]

[3] Barua, S., Erfani, S. M. & Bailey, J. (2019). FCC-GAN: A Fully Connected and Convolutional Net Architecture for GANS. ArXiv. Abs/1905.02417. Retrieved 17 November 2019 from https://arxiv.org/pdf/1905.02417.pdf

[4] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).

[5] Liu, Ming-Yu, Breuel, Thomas, and Kautz, Jan. "Unsupervised Image-to-Image Translation Networks". Advances in Neural Information Processing Systems 30. Ed. by GUYON, I., LUXBURG, U. V., BENGIO, S., et al. Curran Associates, Inc., 2017, 700–708 2.

[6] Tran, Luan and Liu, Xiaoming. "Nonlinear 3D face morphable model". Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018, 7346–7355 3.

[7] Korshunova, Iryna, Shi, Wenzhe, Dambre, Joni, and Theis, Lucas. "Fast Face-Swap Using Convolutional Neural Networks". 2017 IEEE International Conference on Computer Vision (ICCV) (2017), 3697–3705 2.

[8] Yan, Shuqi, He, Shaorong, Lei, Xue, et al. "Video Face Swap Based on Autoencoder Generation Network". 2018 International Conference on Audio, Language and Image Processing (ICALIP). IEEE. 2018, 103–108 2.

[9] Zhao, Yucheng, Tang, Fan, Dong, Weiming, et al. "Joint face alignment and segmentation via deep multi-task learning". Multi media Tools and Applications (Jan. 2018). ISSN: 1573-7721. DOI: 10.1007/s11042-018-5609-1 2.

[10] Natsume, Ryota, Yatagawa, Tatsuya, and Morishima, Shigeo. "FSNet: An Identity-Aware Generative Model for Imagebased Face Swapping". Proc. of Asian Conference on Computer Vision (ACCV). Springer, Dec. 2018 2, 3.

[11] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Niessner, P. Perez, ´C. Richardt, M. Zollhofer, and C. Theobalt, "Deep video portraits," ¨ ACM Trans. Graph., vol. 37, no. 4, pp. 163:1–163:14, Jul. 2018.

[12] Dong, Hao, Neekhara, Paarth, Wu, Chao, and Guo, Yike. "Unsupervised Image-to-Image Translation with Generative Adversarial Networks". CoRR abs/1701.02676 (2017)

[13] Isola, Phillip, Zhu, Jun-Yan, Zhou, Tinghui, and Efros, Alexei A. "Image-to-image translation with conditional adversarial networks". Proceedings of the IEEE conference on computer vision and pattern recognition. 2017, 1125–1134

[14] Wang, Ting-Chun, Liu, Ming-Yu, Zhu, Jun-Yan, et al. "Video-to-Video Synthesis". Advances in Neural Information Processing Systems (NeurIPS). 2018

[15] Natsume, Ryota, Yatagawa, Tatsuya, and Morishima, Shigeo. "RSGAN: face swapping and editing using face and hair representation in latent spaces". Aug. 2018, 1–2.

[16] Shu, Zhixin, Yumer, Ersin, Hadap, Sunil, et al. "Neural face editing with intrinsic image disentangling". Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017, 5541–5550

[17] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-toimage translation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[18] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-forensics: Using capsule networks to detect forged images and videos," in ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 2307–2311

[19] F. Matern, C. Riess, and M. Stamminger, "Exploiting visual artifacts to expose deepfakes and face manipulations," in IEEE Winter Applications of Computer Vision Workshops (WACVW), Jan 2019, pp. 83–92.

[20] U. A. Ciftci, I. Demir and L. Yin, "FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals," in IEEE Transactions on Pattern Analysis and Machine Intelligence, doi: 10.1109/TPAMI.2020.3009287.

[21] W. Shi et al., "Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 1874-1883, doi: 10.1109/CVPR.2016.207.