One-shot Face Reenactment Using Appearance Adaptive Normalization

Guangming Yao[1*], Tianjia Shao[*2], Yi Yuan [1†], Shuang Li[3], Shanqi Liu[3], Yong Liu[4], Mengmeng Wang[4], Kun Zhou[2]

[1] NetEase Fuxi AI Lab
[2] State Key Lab of CAD&CG, Zhejiang University
[3] School of Computer Science and Technology, Beijing Institute of Technology
[4] Institute of Cyber-Systems and Control, Zhejiang University
yaoguangming@corp.netease.com, tjshao@zju.edu.cn, yuanyi@corp.netease.com, shuangli@bit.edu.cn, shanqiliu@zju.edu.cn, yongliu@iipc.zju.edu.cn, mengmengwang@zju.edu.cn, kunzhou@acm.org

Abstract

The paper proposes a novel generative adversarial network for one-shot face reenactment, which can animate a single face image to a different pose-and-expression (provided by a driving image) while keeping its original appearance. The core of our network is a novel mechanism called appearance adaptive normalization, which can effectively integrate the appearance information from the input image into our face generator by modulating the feature maps of the generator using the learned adaptive parameters. Furthermore, we specially design a local net to reenact the local facial components (i.e., eyes, nose and mouth) first, which is a much easier task for the network to learn and can in turn provide explicit anchors to guide our face generator to learn the global appearance and pose-and-expression. Extensive quantitative and qualitative experiments demonstrate the significant efficacy of our model compared with prior one-shot methods.

Introduction

In this paper we seek a one-shot face reenactment network, which can animate a single source image to a different pose-and-expression (provided by a driving image) while keeping the source appearance (i.e identity). We start with the perspective that a face image can be divided into two parts, the pose-and-expression and the appearance, which is also adopted by previous work (Zhang et al. 2019). In face reenactment, the transferring of pose-and-expression is relatively easy because the training data can cover most possible poses and expressions. The main challenge of face reenactment is how to preserve the appearances of different identities. This insight motivates us to design a new architecture, which exploits a novel mechanism called the appearance adaptive normalization, to better control the feature maps of the face generator for the awareness of the source appearance. In general, the appearance adaptive normalization can effectively integrate the specific appearance information from the source image into the synthesized image, by modulating the feature maps of the face generator. Especially, the appearance adaptive normalization learns specific adaptive parameters (i.e., mean and variance) from the source image, which are utilized to modulate feature maps in the generator. In this way, the face generator can be better aware of the appearance of the source image and effectively preserve the source appearance.

The appearance adaptive normalization is inspired by recent adaptive normalization methods (Huang and Belongie 2017; Park et al. 2019), which perform cross-domain image generation without retraining for a specific domain. This attribute makes adaptive normalization potentially suitable for one-shot face reenactment, in which each identity could be seen as a domain. However, there exists a key challenge to apply these adaptive normalization methods to face reenactment. That is, these existing adaptive normalization methods are all designed to deal with the pixel-aligned image-to-image translation problems. For example, in (Park et al. 2019) they propose spatially-adaptive normalization for synthesizing photorealistic images given an input semantic layout. However, in the scenario of face reenactment, the source and driving images are not pixel-aligned. Such pixel misalignment makes it difficult to optimize the adaptive normalization layers during training in existing methods. Consequently, the existing methods will yield distorted images after reenactment, and we will show it in the experiments. To tackle this challenge, one key insight of our work is that instead of learning individual adaptive parameters for different adaptive normalization layers using independent architectures, we can use a unified network to learn all the adaptive parameters from the source image in a global way. The benefit of such paradigm is, by jointly learning the adaptive parameters, the different adaptive normalization layers can be globally modulated rather than being modulated locally. In this way, we can effectively optimize the adaptive normalization layers and control the feature maps of face generator to keep the source appearance. Specifically, we design a simple but effective skip-connected network to predict the adaptive parameters from the source image, which can explicitly promote the relations within adaptive parameters for different adaptive normalization layers, and thus effectively propagate the appearance information throughout the network during reenacting.

We make another key observation that, compared with reenacting the whole faces with largely varying appearances and expressions, reenacting the local facial components (i.e.,

[*]Both authors contributed equally to this research.
[†]Corresponding author

Figure 1: Generated examples by our method. The source image provides the appearance and different driving images provide different expressions and head poses. The reenacted face has the same appearance as the source and the same pose-and-expression as the driving. Both the source and driving images are unseen in the training stage.

eyes, nose, and mouth) is a much easier task for the network to learn. It is because the space of appearance and pose-and-expression is significantly reduced for these local regions. To this end, we can learn the reenactment of these local regions first, which can in turn provide explicit anchors to guide our generator to learn the global appearance and pose-and-expression. Especially, the landmarks are utilized to locate the source and target positions of each face component, so the network only needs to learn the reenactment of these components locally. After local reenacting, the synthesized face components are transformed to the target positions and scales with a similarity transformation and fed to the global generator for the global face synthesis.

In summary, we propose a novel framework for one-shot face reenactment, which utilizes appearance adaptive normalization to better preserve the appearance during reenacting and local facial region reenactment to guide the global synthesis of the final image. Our model only requires one source image to provide the appearance and one driving image to provide the pose-and-expression, both of which are unseen in the training data. The experiments on a variety of face images demonstrate that our method outperforms the state-of-the-art one-shot methods in both objective and subjective aspects (e.g., photo-realism and appearance preservation).

The main contributions of our work are:

1) We propose a novel method for one-shot face reenactment, which animates the source face to another pose-and-expression while preserving its original appearance using only one source image. In particular, we propose an appearance adaptive normalization mechanism to better retain the appearance.

2) We introduce the reenactment of local facial regions to guide the global synthesis of the final reenacted face.

3) Extensive experiments show that our method is able to synthesize reenacted images with both high photo-realism and appearance preservation.

Related Work

Face Reenactment

Face reenactment is a special conditional face synthesis task that aims to animate a source face image to a pose-and-expression of driving face. Common approaches to face reenactment could be roughly divided into two categories: many-to-one and many-to-many. Many-to-one approaches perform face reenactment for a specific person. ReenactGAN (2018) utilizes CycleGAN (2017) to convert the facial boundary heatmaps between different persons, and hence promote the quality of the result synthesized by an identity-specific decoder. Face2Face (2016) animates the facial expression of source video by swapping the source face with the rendered image. The method of Kim et al. (2018) can synthesize high-resolution and realistic facial images with GAN. However, all these methods require a large number of images of the specific identity for training and only reenact the specific identity. On the contrary, our method is capable of reenacting any identity given only a single image without the need for retraining or fine-turning.

To extend face reenactment to unseen identities, some many-to-many methods have been proposed recently. Zakharov et al. (2019) adopt the architecture of BigGAN (2018) and fashional meta-learning, which is capable of synthesizing a personalized talking head with several images, but it requires fine-tuning when a new person is introduced. Zhang et al. (2019) propose an unsupervised approach to face reenactment, which does not need multiple poses for the same identity. Yet, the face parsing map, an identity-specific feature, is utilized to guide the reenacting, which leads to distorted results when reenacting a different identity. Geng et al. (2018) introduce warp-guided GANs for single-photo facial animation. However, their method needs a photo with frontal pose and neutral expression, while ours does not have this limitation. (Pumarola et al. 2018) generates a face guided by action units (1978), which makes it difficult to handle pose changes. X2Face (2018) is able to animate a face under the guidance of pose, expression, and audio, but it can not generate face regions that do not exist in

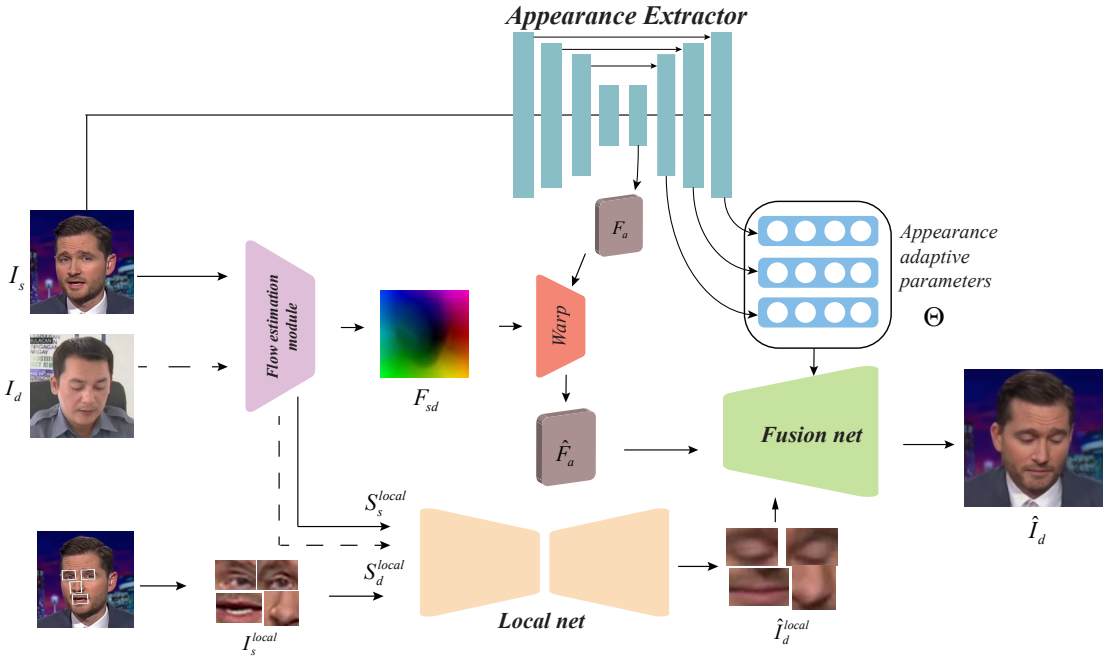

Figure 2: The architecture of generator of our proposed method.

original images. MonkeyNet (2019a) provides a framework for animating general objects. However, the unsupervised keypoints detection may lead to distorted results in the one-shot case. MarioNetTe (2019) proposes the landmark transformer to preserve the source shape during reenactment, but it does not consider how to retain the source appearance. Different from previous many-to-many methods, our goal is to synthesize a high-quality face image, by learning the appearance adaptive parameters to preserve the source appearance and utilizing the local component synthesis to guide the global face synthesis.

Adaptive normalization

The idea of adapting features to different distributions has been successfully applied in a variety of image synthesis tasks (Huang and Belongie 2017; Park et al. 2019). The adaptive normalization normalizes the feature to zero mean and unit deviation first, and then the normalized feature is denormalized by modulating the feature using the learned mean and standard deviation. In conditional BN (de Vries et al. 2017; Zhang et al. 2018), the fixed categorical images are synthesized using different parameters of the normalization layers for different categories. However, unlike the categorical image generation with fixed categories, the number of identities is unknown in the one-shot face reenactment. AdaIN (Huang and Belongie 2017) predicts the adaptive parameters for style transfer, which is spatially sharing. However, it is insufficient in controlling the global appearance, since the facial appearance is spatially varying. SPADE (Park et al. 2019) deploys a spatially varying normalization, which makes it suitable for spatially varying situations. However, SPADE (Park et al. 2019) is designed for the pixel-aligned image translation task which uses independent blocks to locally predict the adaptive parameters for different layers. In face reenactment, the source and driving images are not pixel-aligned, which makes it difficult to locally optimize the different adaptive normalization layers. Hence, we propose the appearance adaptive normalization mechanism to globally predict adaptive parameters of different layers using a skip-connected network, which better promotes the relations within the adaptive parameters for different layers during transferring.

Methodology

For convenience, we denote the images in the dataset as $I_{i\ i=1,\ldots,N_j}^{j\ j=1,\ldots,M}$, where j denotes the identity index and i denotes the image index of identity j. M is the number of identities and N_j is the number of images of identity j. $S_i^j \in \mathbb{R}^{68\times H\times W}$ denotes the corresponding heatmaps for the 68 facial landmarks of $I_i^j \in \mathbb{R}^{3\times H\times W}$, where H and W are the image height and width.

Overview

Our method is a generative adversarial method. We adopt a self-supervised approach to train the network in an end-to-end way, where the driving image I_d has the same identity as I_s in the training stage (i.e., two frames from a video). The landmark transformer (Ha et al. 2019) is utilized to improve the identity preservation. Fig.2 shows the architecture of the proposed generator, which takes as input the source image I_s and the driving image I_d. Our generator is composed of 4 sub-nets, and all the 4 sub-nets are jointly trained in an end-to-end way. First, to preserve the source appearance, we

send I_s to the appearance extractor to learn the appearance adaptive parameters Θ as well as the encoded appearance feature F_a, as shown at the top of Fig. 2. Second, to estimate the facial movements from the source image to the driving pose-and expression, the flow estimation module estimates the optical flow F_{sd} from I_s to I_d , which is then utilized to warp the encoded appearance feature, as shown in the middle of Fig. 2. Third, the local net is deployed to reenact the local facial regions, which provides essential anchors to guide the subsequent synthesis of the whole face, as shown at the bottom of Fig. 2. Finally, the fusion net fuses the adaptive parameters Θ, the reenacted local face regions $\hat{I}_d^{local}$ and the warped appearance feature $\hat{F}_a$, to synthesize the reenacted face. By modulating the distribution of feature maps in the fusion net using the appearance adaptive parameters, we let F_{sd} determine the pose-and-expression, and F_a and Θ retain the appearance.

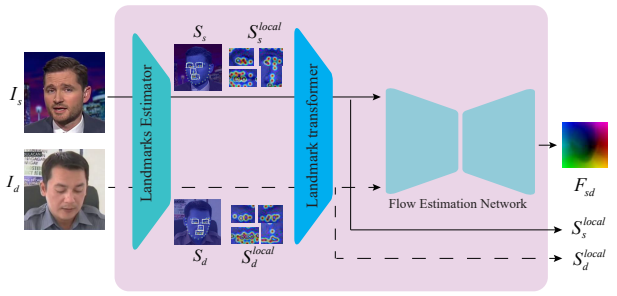

Figure 3: The procedure of flow estimation module.

Flow Estimation Module The procedure of flow estimation module is illustrated in Fig. 3. Firstly, we estimate landmarks for I_s and I_d to obtain the source heatmap S_s and the driving heatmap S_d respectively using OpenFace(Amos, Ludwiczuk, and Satyanarayanan 2016). We then feed S_s and S_d into the flow estimation net (FEN) to produce an optical flow $F_{sd} \in \mathcal{R}^{2\times H\times W}$, representing the motion of pose-and-expression. F_{sd} is then utilized to warp the appearance feature F_a. Bilinear sampling is used to sample F_{sd} to the spatial size of F_a. The warped F_a is denoted as $\hat{F}_a$, which is subsequently fed into the fusion net to synthesize the final reenacted face. Besides, we also build the heatmaps of local regions for source and driving images based on the landmarks, denoted as S_s^{local} and S_d^{local} respectively. The architecture of FEN is an hourglass net (Yang, Liu, and Zhang 2017), composed of several convolutional down-sampling and up-sampling layers. Notably, large shape differences between the source identity and the driving identity will lead to severe degradation of the quality of generated images, which is also mentioned by (Wu et al. 2018). To deal with this issue, we additionally adopt the landmark transformer (Ha et al. 2019), which edits the driving heatmap S_d so that S_d has a shape close to S_s. For more details, please refer to (Ha et al. 2019).

Local Net The local net G_{local} is built with the U-Net structure (Ronneberger, Fischer, and Brox 2015). We reenact the left eye, right eye, nose and mouth with 4 independent networks G_{eyel}, G_{eyer}, G_{nose}, and G_{mouth}. Each of them is a U-Net with three down-convolution blocks and three up-convolution blocks. The inputs of each local generator are I_s^{local}, S_s^{local} and S_d^{local}, where *local* refers to the corresponding parts (i.e., left eye, right eye, nose and mouth) on the image and heatmap. The reenacted face local regions serve as anchor regions that can effectively guide the fusion net to synthesize the whole reenacted face.

Appearance Extractor The source image I_s is fed into the appearance extractor $E_a(I_s)$ for predicting the adaptive parameters Θ and the appearance feature F_a. Here $\Theta = \{\theta_i = (\gamma_i, \beta_i), i \in \{1, 2, ..., N_a\}\}$, where i is the index of the adaptive normalization layer and N_a denotes the number of adaptive normalization layers in the fusion net. For a feature map $F_i \in \mathcal{R}^{c\times h\times w}$ in the fusion net, we have the corresponding $\gamma_i, \beta_i \in \mathcal{R}^{c\times h\times w}$ to modulate it. The encoded source appearance feature F_a is warped to $\hat{F}_a$ using the optical flow F_{sd}, and Θ and $\hat{F}_a$ are fed to the fusion net for face synthesis by controlling the distributions of feature maps. We employ the U-net (2015) architecture for the appearance extractor, because the skip-connection in appearance extractor can effectively promote the relations between adaptive parameters.

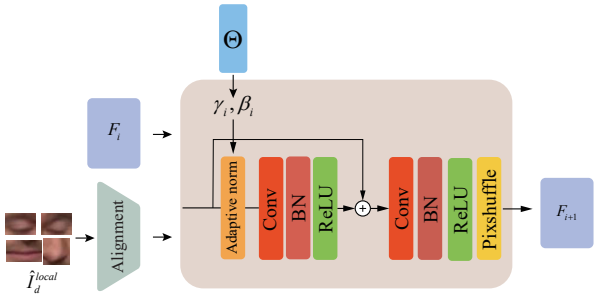

Figure 4: The fusion block of the proposed method.

Fusion Net The fusion net $\hat{I}_d = G_f(\hat{I}_d^{local}, \hat{F}_a, \Theta)$ aims to decode the reenacted local regions I_d^{local} and the warped appearance feature $\hat{F}_a$ to a reenacted face image $\hat{I}_d$ under the control of adaptive parameters Θ. G_f is a fully convolutional network, which performs decoding and up-sampling to synthesize the reenacted face. G_f consists of several fusion blocks to adapt the source appearance, followed by several residual-connected convolution layers to produce the final result. The architecture of fusion block is illustrated in Fig. 4. F_i denotes the input feature map of i-th fusion block, γ_i and β_i denote the i-th adaptive parameters and FB_i denotes the i-th fusion block. Before fed into the fusion block, the reenacted local regions $\hat{I}_d^{local}$ are similarly transformed to the target scale-and-position. In this way, the aligned face regions provide explicit anchors to the generator. These aligned $\hat{I}_d^{local}$ are then resized to the same spatial size as F_i using bilinear interpolation. At last, F_i and $\hat{I}_d^{local}$ are concatenated along the channel axis and fed into next

block of G_f. In this way, the formulation of fusion block can be written as:

$$F_{i+1} = FB_i([F_i, \hat{I}_d^{local}], \gamma_i, \beta_i). \quad (1)$$

The core of our fusion net is the appearance adaptive normalization mechanism. Specifically, the feature map is channel-wisely normalized by

$$\mu_c^i = \frac{1}{NH^iW^i} \sum_{n,h,w} F_{n,c,h,w}^i, \quad (2)$$

$$\sigma_c^i = \sqrt{\frac{1}{NH^iW^i} \sum_{n,h,w} [(F_{n,c,h,w}^i)^2 - (\mu_c^i)^2]}, \quad (3)$$

where $F_{n,c,h,w}^i$ is the feature map value before normalization, and μ_c^i and σ_c^i are the mean and standard deviation of the feature map in channel c. The index of the normalized layer is denoted as i. Notably, the denormalization in adaptive normalization is element-wise, where the normalized feature map is denormalized by

$$\gamma_{c,h,w}^i \frac{F_{n,c,h,w}^i - \mu_c^i}{\sigma_c^i} + \beta_{c,h,w}^i. \quad (4)$$

Here $\gamma_{c,h,w}^i$ and $\beta_{c,h,w}^i$ are the scale and bias learned by the appearance extractor from I_s. Besides, instead of using the transposed convolutional layer or the bilinear upsampling layer followed by a convolutional layer to expand the feature-map (Isola et al. 2017; Wang et al. 2018), we adopt the pixel-shuffle (Shi et al. 2016) to upscale the feature map.

Discriminator

There are two discriminators in our method, a discriminator D_L to discriminate whether the reenacted image and the driving heatmap are matched (pose-and-expression consistency) and a discriminator D_I to discriminate whether the source and reenacted image share the same identity (appearance consistency). D_L takes $\hat{I}_d$ and S_d as input, while D_I takes $\hat{I}_d$ and I_s as input. $\hat{I}_d$ is concatenated with S_d or I_s along the channel axis, before being fed into D_L or D_I respectively. To generate a sharp and realistic-looking image, the discriminators should have a large receptive field (Wang et al. 2018). In our method, instead of using a deeper network with larger convolutional kernels, we use a multi-scale discriminator (Wang et al. 2018) which can improve the global consistency of generated images in multiple scales.

Loss function

The total loss function is defined as:

$$L_{total} = \arg\min_G \max_{D_L, D_I} \lambda_{GAN} L_{GAN} + \lambda_c L_c + \lambda_{local} L_{local}., \quad (5)$$

where L_c denotes the content loss, L_{GAN} denotes the adversarial loss and L_{local} denotes local region loss. The adversarial loss is the GAN loss for D_L and D_I:

$$L_{GAN} = \mathbb{E}_{I_s, \hat{I}_d, S_d}[\log D_L(I_d, S_d) + \log(1 - D_L(\hat{I}_d, S_d))] + \mathbb{E}_{I_s, \hat{I}_d, I_d}[\log D_I(I_s, I_d) + \log(1 - D_I(I_s, \hat{I}_d, I_d))]. \quad (6)$$

The content loss is defined as:

$$L_c = L_1(I_d, \hat{I}_d) + L_{per}(I_d, \hat{I}_d), \quad (7)$$

where $L_1(I_d, \hat{I}_d)$ is the pixel-wise L1 loss, measuring the pixel distance between the generated image and the ground-truth image. $L_{per}(I_d, \hat{I}_d)$ is the perceptual loss (Johnson, Alahi, and Fei-Fei 2016), which has been shown to be useful for the task of image generation (Ledig et al. 2017). We make use of the pre-trained VGG (Simonyan and Zisserman 2014) to compute the perceptual loss, and L_{per} is written as:

$$L_{per}(I_d, \hat{I}_d) = \mathbb{E}_{i \in X}[||\Phi_i(I_d) - \Phi_i(\hat{I}_d)||_1], \quad (8)$$

where X represents the layers we use in VGG and $\Phi_i(x)$ denotes the feature map of the i-th layer in X.

The local region loss penalizes the perceptual differences between the reenacted local regions and the local regions on the ground-truth and is defined as:

$$L_{local} = L_{per}(I_{eyel}, \hat{I}_{eyel}) + L_{per}(I_{mouth}, \hat{I}_{mouth}) + L_{per}(I_{nose}, \hat{I}_{nose}) + L_{per}(I_{eyer}, \hat{I}_{eyer}). \quad (9)$$

Experiments

Implementation

The learning rate for the generator and discriminator are set to $2e^{-5}$ and $1e^{-5}$ respectively. We use Adam (Kingma and Ba 2014) as the optimizer. Spectral Normalization (Miyato et al. 2018) is utilized for each convolution layer in the generator. We set $\lambda_{GAN} = 10$, $\lambda_c = 5$ and $\lambda_{local} = 5$ in the loss function. The Gaussian kernel variance of heatmaps is 3.

Datasets and metrics

Both the FaceForensics++ (Rössler et al. 2019) and Celeb-DF (Li et al. 2019) datasets are used for quantitative and qualitative evaluation. The OpenFace (Amos, Ludwiczuk, and Satyanarayanan 2016) is utilized to detect the face and extract facial landmarks. Following the work of MarionNetTe(2019), we adopt the following metrics to quantitatively evaluate the reenacted faces of different methods. Frechet Inception Distance (FID) (Heusel et al. 2017) and structural similarity index (SSIM) (Wang et al. 2004) are utilized to measure the photographly similarity between the reenacted images and the ground-truth images. Those two metrics are only computed in the self-reenactment scenario since the ground-truth is inaccessible when reenacting a different person. Then we evaluate the identity preservation by calculating the cosine similarity (CSIM) of identity vectors between the source image and the generated image. The identity vectors are extracted by the pre-trained state-of-the-art face recognition networks (Deng et al. 2019). To inspect the model's capability of properly reenacting the pose and expression of driving image, we calculate PRMSE (Ha et al. 2019) and AUCON (Ha et al. 2019) between the generated image and the driving image to measure the reenacted pose and expression respectively.

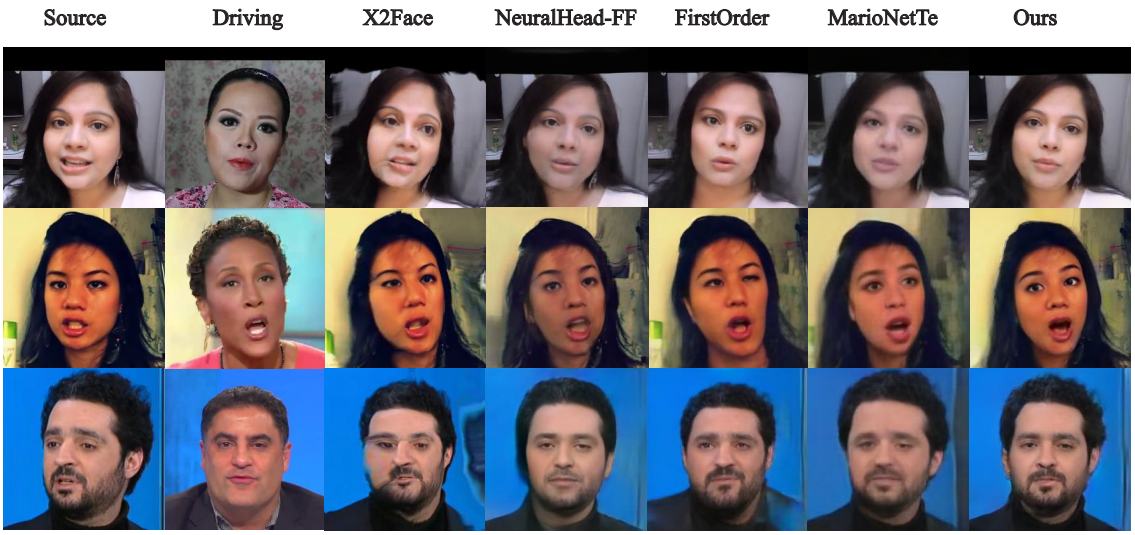

Figure 5: Qualitative comparison with state-of-the-art one-shot methods. Our proposed method generates more natural-looking and sharp results compared to previous methods.

Table 1: Quantitative comparison in the self-reenactment setting. Up/down arrows correspond to higher/lower values for better performance. Bold and underlined numbers represent the best and the second-best values of each metric respectively.

Model	CSIM↑	SSIM↑	FID↓	PRMSE↓	AUCON↑
FaceForensics++ (2019)					
X2face(2018)	0.689	0.719	31.098	3.26	0.813
NeuralHead-FF (2019)	0.229	0.635	38.844	3.76	0.791
MarioNETte (2019)	0.755	**0.744**	44.390	**3.13**	0.825
FirstOrder (2019b)	0.813	0.723	36.124	3.79	**0.886**
Ours	**0.823**	0.730	**30.394**	3.26	0.831
Celeb-DF (2019)					
X2face (2018)	0.676	0.473	14.186	4.10	0.679
NeuralHead-FF (2019)	0.511	0.586	17.973	6.09	0.747
MarioNETte (2019)	0.650	0.508	15.762	3.98	0.714
FirstOrder (2019b)	0.687	0.613	13.620	3.15	**0.839**
Ours	**0.753**	**0.667**	**12.597**	**3.12**	0.751

Table 2: Quantitative comparison of reenacting a different identity.

Model	CSIM↑	PRMSE↓	AUCON↑
Faceforensics++ (2019)			
X2face (2018)	0.604	9.80	0.697
NeuralHead-FF (2019)	0.381	6.82	0.730
MarioNETte (2019)	0.620	7.68	0.710
FirstOrder (2019b)	0.614	**6.62**	**0.734**
Ours	**0.658**	7.04	0.706
Celeb-DF (2019)			
X2face(2018)	0.400	6.52	0.400
NeuralHead-FF (2019)	0.352	8.30	0.480
MarioNETte (2019)	0.460	5.16	**0.662**
FirstOrder (2019b)	0.432	6.10	0.500
Ours	**0.463**	**5.10**	0.660

Quantitative and qualitative comparison

Table 1 lists the quantitative comparisons with existing one-shot reenactment methods when reenacting the same identity, and Table 2 reports the evaluation results when reenacting a different identity. It is worth mentioning that the method that, following (Ha et al. 2019), we re-implement (Zakharov et al. 2019) using only the feed-forward network in the one-shot setting. Differ from other competitors, FirstOrder (2019b) require two driving image to perform the relative motion transfer, one image provide the initial driving pose-and-expression and another one to provides the target driving pose-and-expression. We use the source image to provide the initial driving pose-and-expression when reenacting the same identity to perform the relative motion transfer, and the absolute motion transfer is adopted when reenacting different identities as the initial driving image is lacked for all competitors.

Notably, the results show that our method outperforms other methods in many metrics, demonstrating our method can synthesize highly realistic faces while effectively retaining the source appearance and faithfully reenacting the pose-and-expression. Fig. 5 illustrates typical qualitative examples, all of which are randomly selected from the testing set. We can see that X2face (2018) is unable to generate face regions that do not exist in the source images, so it may result in large artifacts. As the state of art, MarioNETte (2019) can effectively preserve the source shape, but there may still be some appearance artifacts in some regions. Our method fixes this issue by introducing the appearance adaptive normalization and local region reenacting.

We also qualitatively compare our method with recently proposed methods of Zhang et al. (2019) and FS-

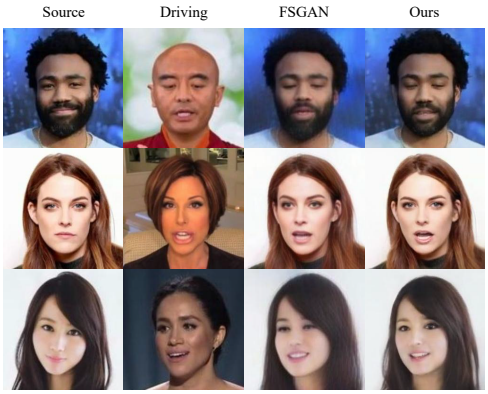

Figure 6: Comparison of our method with FSGAN(2019), source andn driving images are cited from FSGAN(2019)

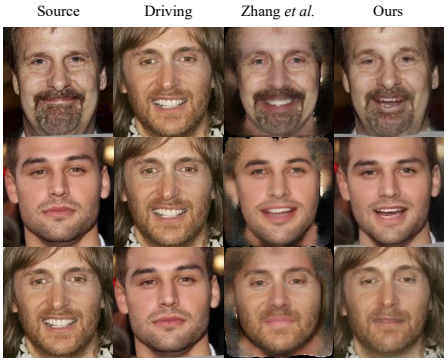

Figure 7: Comparison of our method with Zhang *et al.*(2019), source andn driving images are cited from Zhang *et al.*(2019).

GAN(2019), demonstrated in Fig. 6 and Fig. 7. We can observe blurriness and color-inconsistency in the results of FSGAN(2019). Also the images synthesized by Zhang et al. (2019) have distorted face shapes and artifacts in boundaries, because Zhang et al. (2019) utilize the face parsing map, which is an identity-specific feature, to guide the reenacting. On the contrary, with the help of appearance adaptive normalization and local region reenacting, our method can achieve more detailed and natural-looking results.

Ablation study

Table 3: Quantitative ablation study for reenacting a different identity on the Faceforensics++ dataset (Rössler et al. 2019).

Model	CSIM↑	PRMSE↓	AUCON↑
- local net	0.615	7.293	0.698
- AAN + SPADE	0.558	11.030	0.660
Ours	**0.658**	**7.04**	**0.706**

To better evaluate the key components within our network, we perform the ablation study by evaluating the following variants of our method:

- $-LocalNet$. The local net is excluded from the full model.
- $-AAN + SPADE$. To validate the effectiveness of appearance adaptive normalization, we use the spatially-adaptive normalization to replace it, and all the other components are the same as our model.

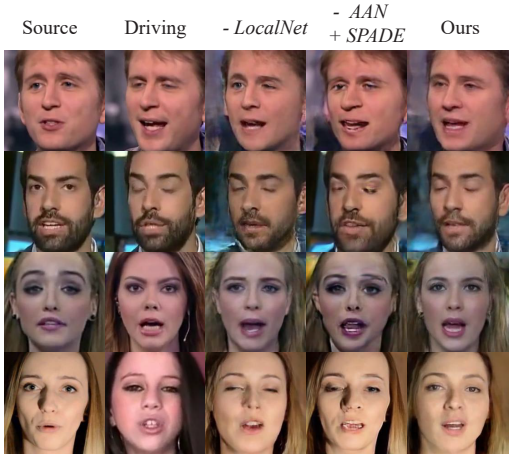

Figure 8: Qualitative results of the ablation study. Our full model leads to better results than other variants.

The qualitative results are illustrated in Fig. 8 and quantitative results are listed in Table 3. We can see that our full model presents the most realistic and natural-looking results. The local net can help reduce the pose-and-expression error, as it explicitly provides anchors for local face regions to guide the reenacting. The appearance adaptive normalization can effectively improve image quality and reduce artifacts by globally modulating the appearance features. Compared to the spatially-adaptive normalization (2019), our appearance adaptive normalization can better preserve the source appearance and leads to more realistic results. It validates our appearance adaptive normalization is more suitable for face reenactment.

Conclusion and future work

In the paper, we propose a novel method to deal with the challenging problem of one-shot face reenactment. Our network deploys a novel mechanism called appearance adaptive normalization to effectively integrate the source appearance information into our face generator, so that the reenacted face image can better preserve the same appearance as the source image. Besides, we design a local net to reenact the local facial components first, which can in turn guide the global synthesis of face appearance and pose-and-expression. Compared to previous methods, our network exhibits superior performance in different metrics. In the future, we plan to explore the temporal consistency in the network design to facilitate the face reenactment in videos.

Acknowledgments We thank anonymous reviewers for their valuable comments. This work is supported by National Key R&D Program of China (2018YFB1004300), NSF China (No. 61772462, No. U1736217) and the 100 Talents Program of Zhejiang University.

References

Amos, B.; Ludwiczuk, B.; and Satyanarayanan, M. 2016. OpenFace: A general-purpose face recognition library with mobile applications. Technical report, CMU-CS-16-118, CMU School of Computer Science.

Brock, A.; Donahue, J.; and Simonyan, K. 2018. Large Scale GAN Training for High Fidelity Natural Image Synthesis.

de Vries, H.; Strub, F.; Mary, J.; Larochelle, H.; Pietquin, O.; and Courville, A. 2017. Modulating early visual processing by language.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4690–4699.

Friesen, E.; and Ekman, P. 1978. Facial action coding system: a technique for the measurement of facial movement. *Palo Alto* 3.

Geng, J.; Shao, T.; Zheng, Y.; Weng, Y.; and Zhou, K. 2018. Warp-guided GANs for single-photo facial animation. In *SIGGRAPH Asia 2018 Technical Papers*, 231. ACM.

Ha, S.; Kersner, M.; Kim, B.; Seo, S.; and Kim, D. 2019. MarioNETte: Few-shot Face Reenactment Preserving Identity of Unseen Targets.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, 6626–6637.

Huang, X.; and Belongie, S. 2017. Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization. *2017 IEEE International Conference on Computer Vision (ICCV)* doi:10.1109/iccv.2017.167. URL http://dx.doi.org/10.1109/iccv.2017.167.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-Image Translation with Conditional Adversarial Networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* doi:10.1109/cvpr.2017.632. URL http://dx.doi.org/10.1109/cvpr.2017.632.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, 694–711. Springer.

Kim, H.; Carrido, P.; Tewari, A.; Xu, W.; Thies, J.; Niessner, M.; Pérez, P.; Richardt, C.; Zollhöfer, M.; and Theobalt, C. 2018. Deep video portraits. *ACM Transactions on Graphics (TOG)* 37(4): 163.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. 2017. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4681–4690.

Li, Y.; Yang, X.; Sun, P.; Qi, H.; and Lyu, S. 2019. Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics.

Miyato, T.; Kataoka, T.; Koyama, M.; and Yoshida, Y. 2018. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957* .

Nirkin, Y.; Keller, Y.; and Hassner, T. 2019. Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE International Conference on Computer Vision*, 7184–7193.

Park, T.; Liu, M.-Y.; Wang, T.-C.; and Zhu, J.-Y. 2019. Semantic Image Synthesis with Spatially-Adaptive Normalization.

Pumarola, A.; Agudo, A.; Martinez, A. M.; Sanfeliu, A.; and Moreno-Noguer, F. 2018. Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 818–833.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. *ArXiv* abs/1505.04597.

Rössler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; and Nießner, M. 2019. FaceForensics++: Learning to Detect Manipulated Facial Images. In *International Conference on Computer Vision (ICCV)*.

Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A. P.; Bishop, R.; Rueckert, D.; and Wang, Z. 2016. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1874–1883.

Siarohin, A.; Lathuilière, S.; Tulyakov, S.; Ricci, E.; and Sebe, N. 2019a. Animating Arbitrary Objects via Deep Motion Transfer. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Siarohin, A.; Lathuilière, S.; Tulyakov, S.; Ricci, E.; and Sebe, N. 2019b. First Order Motion Model for Image Animation. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* .

Thies, J.; Zollhofer, M.; Stamminger, M.; Theobalt, C.; and Nießner, M. 2016. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2387–2395.

Wang, T.-C.; Liu, M.-Y.; Zhu, J.-Y.; Tao, A.; Kautz, J.; and Catanzaro, B. 2018. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* doi:10.1109/cvpr.2018.00917. URL http://dx.doi.org/10.1109/cvpr.2018.00917.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; Simoncelli, E. P.; et al. 2004. Image quality assessment: from error visibility to

structural similarity. *IEEE transactions on image processing* 13(4): 600–612.

Wiles, O.; Koepke, A. S.; and Zisserman, A. 2018. X2Face: A network for controlling face generation by using images, audio, and pose codes.

Wu, W.; Zhang, Y.; Li, C.; Qian, C.; and Change Loy, C. 2018. Reenactgan: Learning to reenact faces via boundary transfer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 603–619.

Yang, J.; Liu, Q.; and Zhang, K. 2017. Stacked hourglass network for robust facial landmark localisation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 79–87.

Zakharov, E.; Shysheya, A.; Burkov, E.; and Lempitsky, V. 2019. Few-Shot Adversarial Learning of Realistic Neural Talking Head Models.

Zhang, H.; Goodfellow, I.; Metaxas, D.; and Odena, A. 2018. Self-Attention Generative Adversarial Networks.

Zhang, Y.; Zhang, S.; He, Y.; Li, C.; Loy, C. C.; and Liu, Z. 2019. One-shot Face Reenactment.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networkss. In *Computer Vision (ICCV), 2017 IEEE International Conference on*.