MarioNETte: Few-Shot Face Reenactment Preserving Identity of Unseen Targets

Sungjoo Ha,[*] Martin Kersner,[*] Beomsu Kim,[*] Seokjun Seo,[*] Dongyoung Kim[†]
Hyperconnect
Seoul, Republic of Korea
{shurain, martin.kersner, beomsu.kim, seokjun.seo, dongyoung.kim}@hpcnt.com

Abstract

When there is a mismatch between the target identity and the driver identity, face reenactment suffers severe degradation in the quality of the result, especially in a few-shot setting. The identity preservation problem, where the model loses the detailed information of the target leading to a defective output, is the most common failure mode. The problem has several potential sources such as the identity of the driver leaking due to the identity mismatch, or dealing with unseen large poses. To overcome such problems, we introduce components that address the mentioned problem: image attention block, target feature alignment, and landmark transformer. Through attending and warping the relevant features, the proposed architecture, called MarioNETte, produces high-quality reenactments of unseen identities in a few-shot setting. In addition, the landmark transformer dramatically alleviates the identity preservation problem by isolating the expression geometry through landmark disentanglement. Comprehensive experiments are performed to verify that the proposed framework can generate highly realistic faces, outperforming all other baselines, even under a significant mismatch of facial characteristics between the target and the driver.

Introduction

Given a *target* face and a *driver* face, face reenactment aims to synthesize a *reenacted* face which is animated by the movement of a driver while preserving the identity of the target.

Many approaches make use of generative adversarial networks (GAN) which have demonstrated a great success in image generation tasks. Xu et al.; Wu et al. (2017; 2018) achieved high-fidelity face reenactment results by exploiting CycleGAN (Zhu et al. 2017). However, the CycleGAN-based approaches require at least a few minutes of training data for each target and can only reenact predefined identities, which is less attractive in-the-wild where a reenactment of unseen targets cannot be avoided.

The few-shot face reenactment approaches, therefore, try to reenact any unseen targets by utilizing operations such as adaptive instance normalization (AdaIN) (Zakharov et al. 2019) or warping module (Wiles, Koepke, and Zisserman 2018; Siarohin et al. 2019). However, current state-of-the-art methods suffer from the problem we call *identity preservation problem*: the inability to preserve the identity of the target leading to defective reenactments. As the identity of the driver diverges from that of the target, the problem is exacerbated even further.

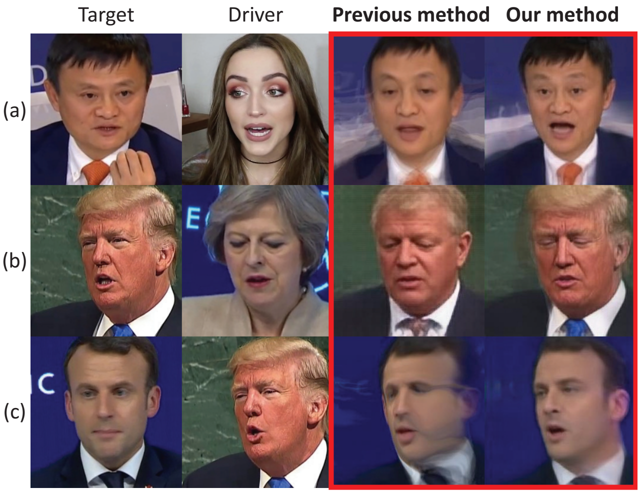

Figure 1: Examples of identity preservation failures and improved results generated by the proposed method. Each row shows (a) driver shape interference, (b) losing details of target identity, and (c) failure of warping at large poses.

Examples of flawed and successful face reenactments, generated by previous approaches and the proposed model, respectively, are illustrated in Figure 1. The failures of previous approaches, for the most part, can be broken down into three different modes [1]:

1. Neglecting the identity mismatch may lead to a identity of the driver interfere with the face synthesis such that the generated face resembles the driver (Figure 1a).

[*] Equal contributions, listed in alphabetical order.

[†] Corresponding author.

[1] Additional example images and videos can be found at the following URL: http://hyperconnect.github.io/MarioNETte

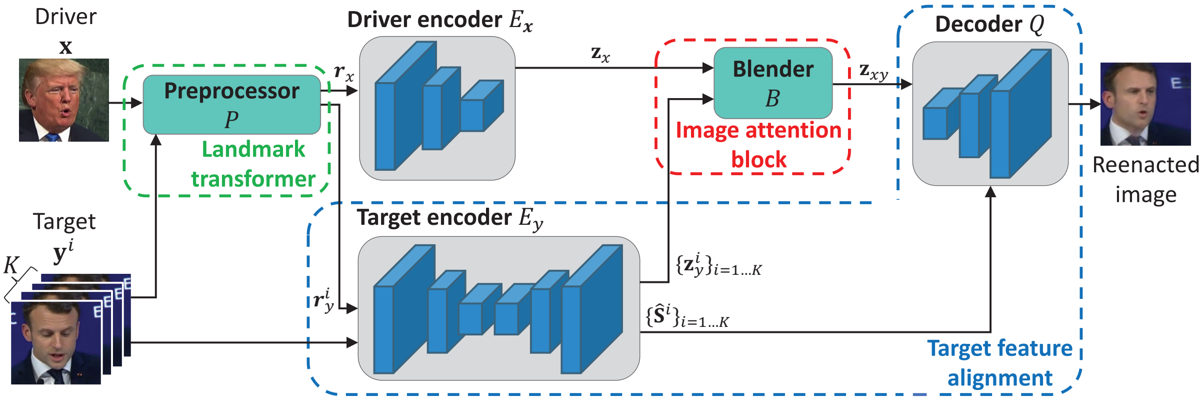

Figure 2: The overall architecture of MarioNETte.

2. Insufficient capacity of the compressed vector representation (e.g., AdaIN layer) to preserve the information of the target identity may lead the produced face to lose the detailed characteristics (Figure 1b).

3. Warping operation incurs a defect when dealing with large poses (Figure 1c).

We propose a framework called *MarioNETte*, which aims to reenact the face of unseen targets in a few-shot manner while preserving the identity without any fine-tuning. We adopt *image attention block* and *target feature alignment*, which allow MarioNETte to directly inject features from the target when generating image. In addition, we propose a novel *landmark transformer* which further mitigates the identity preservation problem by adjusting for the identity mismatch in an unsupervised fashion. Our contributions are as follows:

- We propose a few-shot face reenactment framework called MarioNETte, which preserves the target identity even in situations where the facial characteristics of the driver differs widely from those of the target. Utilizing image attention block, which allows the model to attend to relevant positions of the target feature map, together with target feature alignment, which includes multiple feature-level warping operations, proposed method improves the quality of the face reenactment under different identities.

- We introduce a novel method of landmark transformation which copes with varying facial characteristics of different people. The proposed method adapts the landmark of a driver to that of the target in an unsupervised manner, thereby mitigating the identity preservation problem without any additional labeled data.

- We compare the state-of-the-art methods when the target and the driver identities coincide and differ using VoxCeleb1 (Nagrani, Chung, and Zisserman 2017) and CelebV (Wu et al. 2018) dataset, respectively. Our experiments including user studies show that the proposed method outperforms the state-of-the-art methods.

MarioNETte Architecture

Figure 2 illustrates the overall architecture of the proposed model. A conditional *generator* G generates the reenacted face given the driver $\mathbf{x}$ and the target images $\{\mathbf{y}^i\}_{i=1...K}$, and the *discriminator* D predicts whether the image is real or not. The generator consists of following components:

- The **preprocessor** P utilizes a 3D landmark detector (Bulat and Tzimiropoulos 2017) to extract facial keypoints and renders them to landmark image, yielding $\mathbf{r}_x = P(\mathbf{x})$ and $\mathbf{r}_y^i = P(\mathbf{y}^i)$, corresponding to the driver and the target input respectively. Note that proposed landmark transformer is included in the preprocessor. Since we normalize the scale, translation and rotation of landmarks before using them in a landmark transformer, we utilize 3D landmarks instead of 2D ones.

- The **driver encoder** $E_x(\mathbf{r}_x)$ extracts pose and expression information from the driver input and produces driver feature map $\mathbf{z}_x$.

- The **target encoder** $E_y(\mathbf{y}, \mathbf{r}_y)$ adopts a U-Net architecture to extract style information from the target input and generates target feature map $\mathbf{z}_y$ along with the warped target feature maps $\hat{\mathbf{S}}$.

- The **blender** $B(\mathbf{z}_x, \{\mathbf{z}_y^i\}_{i=1...K})$ receives driver feature map $\mathbf{z}_x$ and target feature maps $\mathbf{Z}_y = [\mathbf{z}_y^1, \ldots, \mathbf{z}_y^K]$ to produce mixed feature map $\mathbf{z}_{xy}$. Proposed image attention block is basic building block of the blender.

- The **decoder** $Q(\mathbf{z}_{xy}, \{\hat{\mathbf{S}}^i\}_{i=1...K})$ utilizes warped target feature maps $\hat{\mathbf{S}}$ and mixed feature map $\mathbf{z}_{xy}$ to synthesize reenacted image. The decoder improves quality of reenacted image exploiting proposed target feature alignment.

For further details, refer to Supplementary Material A1.

Image attention block

To transfer style information of targets to the driver, previous studies encoded target information as a vector and mixed it with driver feature by concatenation or AdaIN layers (Liu et al. 2019; Zakharov et al. 2019). However, encoding targets

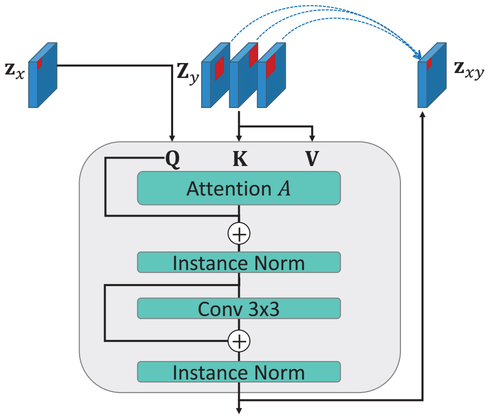

Figure 3: Architecture of the image attention block. Red boxes conceptually visualize how each position of $\mathbf{z}_x$ and $\mathbf{Z}_y$ are associated. Our attention can attend different position of each target feature maps with different importance.

as a spatial-agnostic vector leads to losing spatial information of targets. In addition, these methods are absent of innate design for multiple target images, and thus, summary statistics (e.g. mean or max) are used to deal with multiple targets which might cause losing details of the target.

We suggest image attention block (Figure 3) to alleviate aforementioned problem. The proposed attention block is inspired by the encoder-decoder attention of transformer (Vaswani et al. 2017), where the driver feature map acts as an attention query and the target feature maps act as attention memory. The proposed attention block attends to proper positions of each feature (red boxes in Figure 3) while handling multiple target feature maps (i.e., $\mathbf{Z}_y$).

Given driver feature map $\mathbf{z}_x \in \mathbb{R}^{h_x \times w_x \times c_x}$ and target feature maps $\mathbf{Z}_y = [\mathbf{z}_y^1, \ldots, \mathbf{z}_y^K] \in \mathbb{R}^{K \times h_y \times w_y \times c_y}$, the attention is calculated as follows:

$$\begin{aligned} \mathbf{Q} &= \mathbf{z}_x \mathbf{W}_q + \mathbf{P}_x \mathbf{W}_{qp} && \in \mathbb{R}^{h_x \times w_x \times c_a} \\ \mathbf{K} &= \mathbf{Z}_y \mathbf{W}_k + \mathbf{P}_y \mathbf{W}_{kp} && \in \mathbb{R}^{K \times h_y \times w_y \times c_a} \\ \mathbf{V} &= \mathbf{Z}_y \mathbf{W}_v && \in \mathbb{R}^{K \times h_y \times w_y \times c_x} \end{aligned} \tag{1}$$

$$A(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}\left(\frac{f(\mathbf{Q}) f(\mathbf{K})^T}{\sqrt{c_a}}\right) f(\mathbf{V}), \tag{2}$$

where $f : \mathbb{R}^{d_1 \times \ldots \times d_k \times c} \rightarrow \mathbb{R}^{(d_1 \times \ldots \times d_k) \times c}$ is a flattening function, all $\mathbf{W}$ are linear projection matrices that map to proper number of channels at the last dimension, and $\mathbf{P}_x$ and $\mathbf{P}_y$ are sinusoidal positional encodings which encode the coordinate of feature maps (further details of sinusoidal positional encodings we used are described in Supplementary Material A2). Finally, the output $A(\mathbf{Q}, \mathbf{K}, \mathbf{V}) \in \mathbb{R}^{(h_x \times w_x) \times c_x}$ is reshaped to $\mathbb{R}^{h_x \times w_x \times c_x}$.

Instance normalization, residual connection, and convolution layer follow the attention layer to generate output feature map $\mathbf{z}_{xy}$. The image attention block offers a direct mechanism of transferring information from multiple target images to the pose of driver.

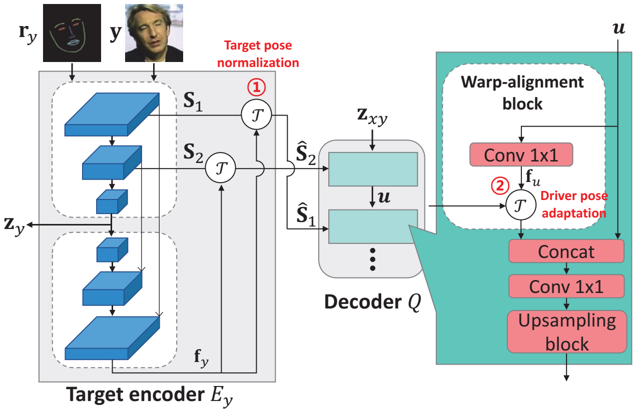

Figure 4: Architecture of target feature alignment.

Target feature alignment

The fine-grained details of the target identity can be preserved through the warping of low-level features (Siarohin et al. 2019). Unlike previous approaches that estimate a warping flow map or an affine transform matrix by computing the difference between keypoints of the target and the driver (Balakrishnan et al. 2018; Siarohin et al. 2018; 2019), we propose a target feature alignment (Figure 4) which warps the target feature maps in two stages: (1) *target pose normalization* generates pose normalized target feature maps and (2) *driver pose adaptation* aligns normalized target feature maps to the pose of the driver. The two-stage process allows the model to better handle the structural disparities of different identities. The details are as follows:

1. **Target pose normalization.** In the target encoder E_y, encoded feature maps $\{\mathbf{S}_j\}_{j=1\ldots n_y}$ are processed into $\hat{\mathbf{S}} = \{\mathcal{T}(\mathbf{S}_1; \mathbf{f}_y), \ldots, \mathcal{T}(\mathbf{S}_{n_y}; \mathbf{f}_y)\}$ by estimated normalization flow map $\mathbf{f}_y$ of target and warping function $\mathcal{T}$ (① in Figure 4). The following *warp-alignment block* at decoder treats $\hat{\mathbf{S}}$ in a target pose-agnostic manner.

2. **Driver pose adaptation.** The warp-alignment block in the decoder receives $\{\hat{\mathbf{S}}^i\}_{i=1\ldots K}$ and the output $\mathbf{u}$ of the previous block of the decoder. In a few-shot setting, we average resolution-compatible feature maps from different target images (i.e., $\hat{\mathbf{S}}_j = \sum_i \hat{\mathbf{S}}_j^i / K$). To adapt pose-normalized feature maps to the pose of the driver, we generate an estimated flow map of the driver $\mathbf{f}_u$ using 1×1 convolution that takes $\mathbf{u}$ as the input. Alignment by $\mathcal{T}(\hat{\mathbf{S}}_j; \mathbf{f}_u)$ follows (② in Figure 4). Then, the result is concatenated to $\mathbf{u}$ and fed into the following residual upsampling block.

Landmark Transformer

Large structural differences between two facial landmarks may lead to severe degradation of the quality of the reenactment. The usual approach to such a problem has been to learn a transformation for every identity (Wu et al. 2018) or by preparing a paired landmark data with the same expressions (Zhang et al. 2019). However, these methods are unnatural in a few-shot setting where we handle unseen identities, and moreover, the labeled data is hard to be acquired. To overcome this difficulty, we propose a novel *landmark transformer* which transfers the facial expression of the driver to an arbitrary target identity. The landmark transformer utilizes multiple videos of unlabeled human faces and is trained in an unsupervised manner.

Landmark decomposition

Given video footages of different identities, we denote $\mathbf{x}(c,t)$ as the t-th frame of the c-th video, and $\mathbf{l}(c,t)$ as a 3D facial landmark. We first transform every landmark into a normalized landmark $\bar{\mathbf{l}}(c,t)$ by normalizing the scale, translation, and rotation. Inspired by 3D morphable models of face (Blanz and Vetter 1999), we assume that normalized landmarks can be decomposed as follows:

$$\bar{\mathbf{l}}(c,t) = \bar{\mathbf{l}}_m + \bar{\mathbf{l}}_{id}(c) + \bar{\mathbf{l}}_{exp}(c,t), \tag{3}$$

where $\bar{\mathbf{l}}_m$ is the average facial landmark geometry computed by taking the mean over all landmarks, $\bar{\mathbf{l}}_{id}(c)$ denotes the landmark geometry of identity c, computed by $\bar{\mathbf{l}}_{id}(c) = \sum_t \bar{\mathbf{l}}(c,t)/T_c - \bar{\mathbf{l}}_m$ where T_c is the number of frames of c-th video, and $\bar{\mathbf{l}}_{exp}(c,t)$ corresponds to the expression geometry of t-th frame. The decomposition leads to $\bar{\mathbf{l}}_{exp}(c,t) = \bar{\mathbf{l}}(c,t) - \bar{\mathbf{l}}_m - \bar{\mathbf{l}}_{id}(c)$.

Given a target landmark $\bar{\mathbf{l}}(c_y, t_y)$ and a driver landmark $\bar{\mathbf{l}}(c_x, t_x)$ we wish to generate the following landmark:

$$\bar{\mathbf{l}}(c_x \to c_y, t_x) = \bar{\mathbf{l}}_m + \bar{\mathbf{l}}_{id}(c_y) + \bar{\mathbf{l}}_{exp}(c_x, t_x), \tag{4}$$

i.e., a landmark with the identity of the target and the expression of the driver. Computing $\bar{\mathbf{l}}_{id}(c_y)$ and $\bar{\mathbf{l}}_{exp}$ is possible if enough images of c_y are given, but in a few-shot setting, it is difficult to disentangle landmark of unseen identity into two terms.

Landmark disentanglement

To decouple the identity and the expression geometry in a few-shot setting, we introduce a neural network to regress the coefficients for linear bases. Previously, such an approach has been widely used in modeling complex face geometries (Blanz and Vetter 1999). We separate expression landmarks into semantic groups of the face (e.g., mouth, nose and eyes) and perform PCA on each group to extract the expression bases from the training data:

$$\bar{\mathbf{l}}_{exp}(c,t) = \sum_{k=1}^{n_{exp}} \alpha_k(c,t)\mathbf{b}_{exp,k} = \mathbf{b}_{exp}^T \boldsymbol{\alpha}(c,t), \tag{5}$$

where $\mathbf{b}_{exp,k}$ and α_k represent the basis and the corresponding coefficient, respectively.

The proposed neural network, a *landmark disentangler* M, estimates $\boldsymbol{\alpha}(c,t)$ given an image $\mathbf{x}(c,t)$ and a landmark $\bar{\mathbf{l}}(c,t)$. Figure 5 illustrates the architecture of the landmark disentangler. Once the model is trained, the identity and the expression geometry can be computed as follows:

$$\begin{aligned} \hat{\boldsymbol{\alpha}}(c,t) &= M\left(\mathbf{x}(c,t),\, \bar{\mathbf{l}}(c,t)\right) \\ \hat{\mathbf{l}}_{exp}(c,t) &= \lambda_{exp}\mathbf{b}_{exp}^T \hat{\boldsymbol{\alpha}}(c,t) \\ \hat{\mathbf{l}}_{id}(c) &= \bar{\mathbf{l}}(c,t) - \bar{\mathbf{l}}_m - \hat{\mathbf{l}}_{exp}(c,t), \end{aligned} \tag{6}$$

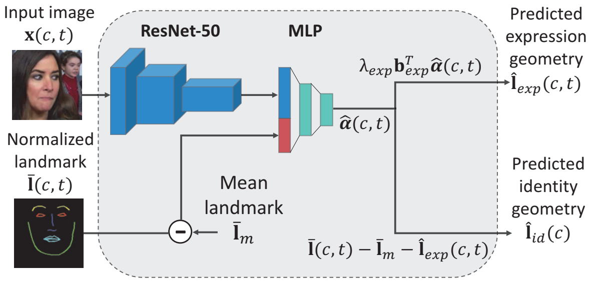
Input image
$\mathbf{x}(c,t)$
ResNet-50
MLP
$\lambda_{exp}\mathbf{b}_{exp}^T\hat{\boldsymbol{\alpha}}(c,t)$
Predicted expression geometry
$\hat{\mathbf{l}}_{exp}(c,t)$
$\hat{\boldsymbol{\alpha}}(c,t)$
Normalized landmark
$\bar{\mathbf{l}}(c,t)$
Mean landmark
$\bar{\mathbf{l}}_m$
Predicted identity geometry
$\hat{\mathbf{l}}_{id}(c)$
$\bar{\mathbf{l}}(c,t) - \bar{\mathbf{l}}_m - \hat{\mathbf{l}}_{exp}(c,t)$

Figure 5: Architecture of landmark disentangler. Note that $\bar{\mathbf{l}}(c,t)$ is a set of landmark points but visualized as an image in the figure.

where λ_{exp} is a hyperparameter that controls the intensity of the predicted expressions from the network. Image feature extracted by a ResNet-50 and the landmark, $\bar{\mathbf{l}}(c,t) - \bar{\mathbf{l}}_m$, are fed into a 2-layer MLP to predict $\hat{\boldsymbol{\alpha}}(c,t)$.

During the inference, the target and the driver landmarks are processed according to Equation 6. When multiple target images are given, we take the mean value over all $\hat{\mathbf{l}}_{id}(c_y)$. Finally, landmark transformer converts landmark as:

$$\hat{\mathbf{l}}(c_x \to c_y, t_x) = \bar{\mathbf{l}}_m + \hat{\mathbf{l}}_{id}(c_y) + \hat{\mathbf{l}}_{exp}(c_x, t_x). \tag{7}$$

Denormalization to recover the original scale, translation, and rotation is followed by the rasterization that generates a landmark adequate for the generator to consume. Further details of landmark transformer are described in Supplementary Material B.

Experimental Setup

Datasets We trained our model and the baselines using VoxCeleb1 (Nagrani, Chung, and Zisserman 2017), which contains 256×256 size videos of 1,251 different identities. We utilized the test split of VoxCeleb1 and CelebV (Wu et al. 2018) for evaluating self-reenactment and reenactment under a different identity, respectively. We created the test set by sampling 2,083 image sets from randomly selected 100 videos of VoxCeleb1 test split, and uniformly sampled 2,000 image sets from every identity from CelebV. The CelebV data includes the videos of five different celebrities of widely varying characteristics, which we utilize to evaluate the performance of the models reenacting unseen targets, similar to in-the-wild scenario. Further details of the loss function and the training method can be found at Supplementary Material A3 and A4.

Baselines MarioNETte variants, with and without the landmark transformer (*MarioNETte+LT* and *MarioNETte*, respectively), are compared with state-of-the-art models for few-shot face reenactment. Details of each baseline are as follows:

- **X2Face** (Wiles, Koepke, and Zisserman 2018). X2face utilizes direct image warping. We used the pre-trained model provided by the authors, trained on VoxCeleb1.

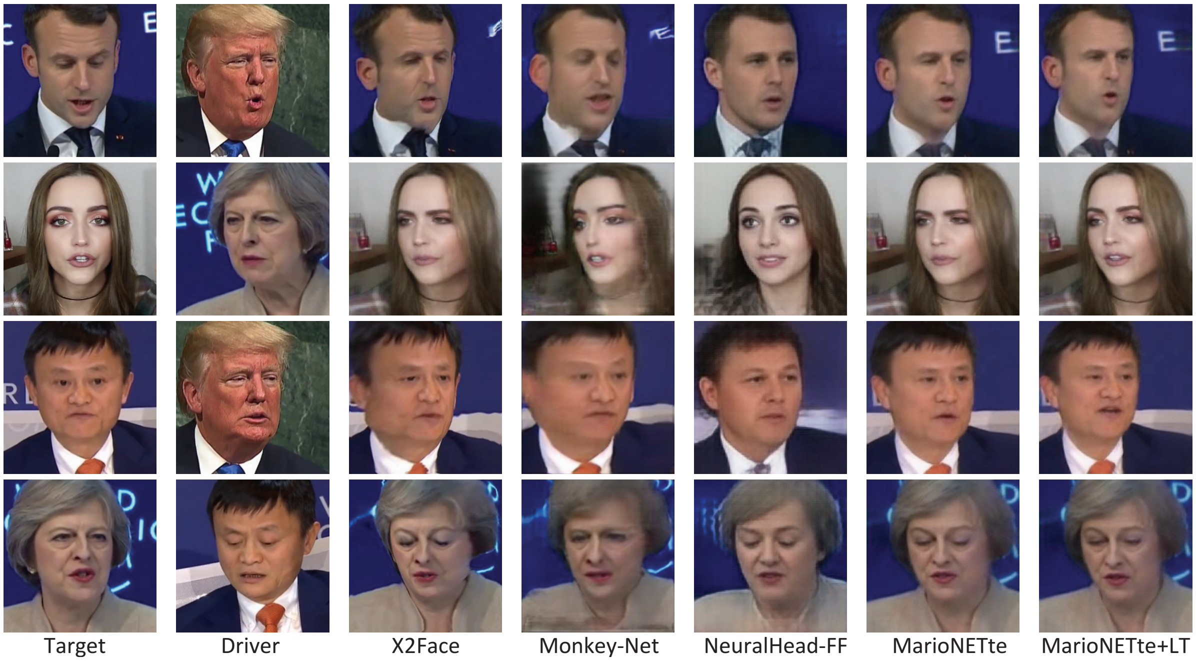

Figure 6: Images generated by the proposed method and baselines, reenacting different identity on CelebV in one-shot setting.

- **Monkey-Net** (Siarohin et al. 2019). Monkey-Net adopts feature-level warping. We used the implementation provided by the authors. Due to the structure of the method, Monkey-Net can only receive a single target image.
- **NeuralHead** (Zakharov et al. 2019). NeuralHead exploits AdaIN layers. Since a reference implementation is absent, we made an honest attempt to reproduce the results. Our implementation is a feed-forward version of their model (*NeuralHead-FF*) where we omit the meta-learning as well as fine-tuning phase, because we are interested in using a single model to deal with multiple identities.

Metrics We compare the models based on the following metrics to evaluate the quality of the generated images. Structured similarity (**SSIM**) (Wang et al. 2004) and peak signal-to-noise ratio (**PSNR**) evaluate the low-level similarity between the generated image and the ground-truth image. We also report the masked-SSIM (**M-SSIM**) and masked-PSNR (**M-PSNR**) where the measurements are restricted to the facial region.

In the absence of the ground truth image where different identity drives the target face, the following metrics are more relevant. Cosine similarity (**CSIM**) of embedding vectors generated by pre-trained face recognition model (Deng et al. 2019) is used to evaluate the quality of identity preservation. To inspect the capability of the model to properly reenact the pose and the expression of the driver, we compute **PRMSE**, the root mean square error of the head pose angles, and **AUCON**, the ratio of identical facial action unit values, between the generated images and the driving images. OpenFace (Baltrusaitis et al. 2018) is utilized to compute pose angles and action unit values.

Experimental Results

Models were compared under self-reenactment and reenactment of different identities, including a user study. Ablation tests were conducted as well. All experiments were conducted under two different settings: one-shot and few-shot, where one or eight target images were used respectively.

Self-reenactment

Table 1 illustrates the evaluation results of the models under self-reenactment settings on VoxCeleb1. *MarioNETte* surpasses other models in every metric under few-shot setting and outperforms other models in every metric except for PSNR under the one-shot setting. However, *MarioNETte* shows the best performance in M-PSNR which implies that it performs better on facial region compared to baselines. The low CSIM yielded from *NeuralHead-FF* is an indirect evidence of the lack of capacity in AdaIN-based methods.

Reenacting Different Identity

Table 2 displays the evaluation result of reenacting a different identity on CelebV, and Figure 6 shows generated images from proposed method and baselines. *MarioNETte* and *MarioNETte+LT* preserve target identity adequately, thereby outperforming other models in CSIM. The proposed method alleviates the identity preservation problem regardless of the driver being of the same identity or not. While *NeuralHead-FF* exhibits slightly better performance in terms of PRMSE and AUCON compared to *MarioNETte*, the low CSIM of *NeuralHead-FF* portrays the failure to preserve the target identity. The landmark transformer significantly boosts identity preservation at the cost of a slight decrease in PRMSE and AUCON. The decrease may be due to the PCA bases for

Model (# target)	CSIM↑	SSIM↑	M-SSIM↑	PSNR↑	M-PSNR↑	PRMSE↓	AUCON↑
X2face (1)	0.689	0.719	0.941	22.537	31.529	3.26	0.813
Monkey-Net (1)	0.697	0.734	0.934	**23.472**	30.580	3.46	0.770
NeuralHead-FF (1)	0.229	0.635	0.923	20.818	29.599	3.76	0.791
MarioNETte (1)	**0.755**	**0.744**	**0.948**	23.244	**32.380**	**3.13**	**0.825**
X2face (8)	0.762	0.776	0.956	24.326	33.328	3.21	0.826
NeuralHead-FF (8)	0.239	0.645	0.925	21.362	29.952	3.69	0.795
MarioNETte (8)	**0.828**	**0.786**	**0.958**	**24.905**	**33.645**	**2.57**	**0.850**

Table 1: Evaluation result of self-reenactment setting on VoxCeleb1. Upward/downward pointing arrows correspond to metrics that are better when the values are higher/lower.

Model (# target)	CSIM↑	PRMSE↓	AUCON↑
X2face (1)	0.450	3.62	0.679
Monkey-Net (1)	0.451	4.81	0.584
NeuralHead-FF (1)	0.108	**3.30**	**0.722**
MarioNETte (1)	0.520	3.41	0.710
MarioNETte+LT (1)	**0.568**	3.70	0.684
X2face (8)	0.484	**3.15**	0.709
NeuralHead-FF (8)	0.120	3.26	**0.723**
MarioNETte (8)	0.608	3.26	0.717
MarioNETte+LT (8)	**0.661**	3.57	0.691

Table 2: Evaluation result of reenacting a different identity on CelebV. **Bold** and underlined values correspond to the best and the second-best value of each metric, respectively.

Model (# target)	vs. Ours	vs. Ours+LT	Realism ↑
X2Face (1)	0.07	0.09	0.093
Monkey-Net (1)	0.05	0.09	0.100
NeuralHead-FF (1)	0.17	0.17	0.087
MarioNETte (1)	-	0.51	0.140
MarioNETte+LT (1)	-	-	**0.187**
X2Face (8)	0.09	0.07	0.047
NeuralHead-FF (8)	0.15	0.16	0.080
MarioNETte (8)	-	0.52	0.147
MarioNETte+LT (8)	-	-	**0.280**

Table 3: User study results of reenacting different identity on CelebV. *Ours* stands for our proposed model, *MarioNETte*, and *Ours+LT* stands for *MarioNETte+LT*.

the expression disentanglement not being diverse enough to span the whole space of expressions. Moreover, the disentanglement of identity and expression itself is a non-trivial problem, especially in a one-shot setting.

User Study

Two types of user studies are conducted to assess the performance of the proposed model:

- **Comparative analysis.** Given three example images of the target and a driver image, we displayed two images generated by different models and asked human evaluators to select an image with higher quality. The users were asked to assess the quality of an image in terms of (1) identity preservation, (2) reenactment of driver's pose and expression, and (3) photo-realism. We report the winning ratio of baseline models compared to our proposed models. We believe that user reported score better reflects the quality of different models than other indirect metrics.
- **Realism analysis.** Similar to the user study protocol of Zakharov et al. (2019), three images of the same person, where two of the photos were taken from a video and the remaining generated by the model, were presented to human evaluators. Users were instructed to choose an image that differs from the other two in terms of the identity under a three-second time limit. We report the ratio of deception, which demonstrates the identity preservation and the photo-realism of each model.

For both studies, 150 examples were sampled from CelebV, which were evenly distributed to 100 different human evaluators.

Table 3 illustrates that our models are preferred over existing methods achieving realism scores with a large margin. The result demonstrates the capability of *MarioNETte* in creating photo-realistic reenactments while preserving the target identity in terms of human perception. We see a slight preference of *MarioNETte* over *MarioNETte+LT*, which agrees with the Table 2, as *MarioNETte+LT* has better identity preservation capability at the expense of slight degradation in expression transfer. Since the identity preservation capability of *MarioNETte+LT* surpasses all other models in realism score, almost twice the score of even *MarioNETte* on few-shot settings, we consider the minor decline in expression transfer a good compromise.

Ablation Test

We performed ablation test to investigate the effectiveness of the proposed components. While keeping all other things the same, we compare the following configurations reenacting different identities: (1) **MarioNETte** is the proposed method where both image attention block and target feature alignment are applied. (2) **AdaIN** corresponds to the same model as MarioNETte, where the image attention block is replaced with AdaIN residual block while the target feature alignment

Model (# target)	CSIM↑	PRMSE↓	AUCON↑
AdaIN (1)	0.063	3.47	0.724
+Attention (1)	0.333	**3.17**	**0.729**
+Alignment (1)	**0.530**	3.44	0.700
MarioNETte (1)	0.520	3.41	0.710
AdaIN (8)	0.069	3.40	0.723
+Attention (8)	0.472	**3.22**	**0.727**
+Alignment (8)	0.605	3.27	0.709
MarioNETte (8)	**0.608**	3.26	0.717

Table 4: Comparison of ablation models for reenacting different identity on CelebV.

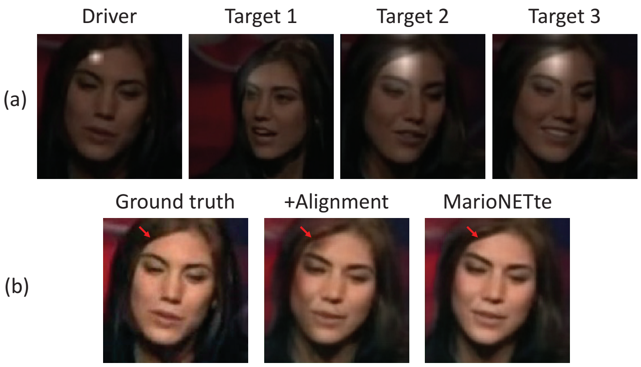

Figure 7: (a) Driver and target images overlapped with attention map. Brightness signifies the intensity of the attention. (b) Failure case of *+Alignment* and improved result generated by *MarioNETte*.

is omitted. (3) **+Attention** is a MarioNETte where only the image attention block is applied. (4) **+Alignment** only employs the target feature alignment.

Table 4 shows result of ablation test. For identity preservation (i.e., CSIM), *AdaIN* has a hard time combining style features depending solely on AdaIN residual blocks. *+Attention* alleviates the problem immensely in both one-shot and few-shot settings by attending to proper coordinates. While *+Alignment* exhibits a higher CSIM compared to *+Attention*, it struggles in generating plausible images for unseen poses and expressions leading to worse PRMSE and AUCON. Taking advantage of both attention and target feature alignment, *MarioNETte* outperforms *+Alignment* in every metric under consideration.

Entirely relying on target feature alignment for reenactment, *+Alignment* is vulnerable to failures due to large differences in pose between target and driver that *MarioNETte* can overcome. Given a single driver image along with three target images (Figure 7a), *+Alignment* has defects on the forehead (denoted by arrows in Figure 7b). This is due to (1) warping low-level features from a large-pose input and (2) aggregating features from multiple targets with diverse poses. *MarioNETte*, on the other hand, gracefully handles the situation by attending to proper image among several target images as well as adequate spatial coordinates in the target image. The attention map, highlighting the area where the image attention block is focusing on, is illustrated with white in Figure 7a. Note that *MarioNETte* attends to the forehead and adequate target images (Target 2 and 3 in Figure 7a) which has similar pose with driver.

Related Works

The classical approach to face reenactment commonly involves the use of explicit 3D modeling of human faces (Blanz and Vetter 1999) where the 3DMM parameters of the driver and the target are computed from a single image, and blended eventually (Thies et al. 2015; 2016). Image warping is another popular approach where the target image is modified using the estimated flow obtained form 3D models (Cao et al. 2013) or sparse landmarks (Averbuch-Elor et al. 2017). Face reenactment studies have embraced the recent success of neural networks exploring different image-to-image translation architectures (Isola et al. 2017) such as the works of Xu et al. (2017) and that of Wu et al. (2018), which combined the cycle consistency loss (Zhu et al. 2017). A hybrid of two approaches has been studied as well. Kim et al. (2018) trained an image translation network which maps reenacted render of a 3D face model into a photo-realistic output.

Architectures, capable of blending the style information of the target with the spatial information of the driver, have been proposed recently. AdaIN (Huang and Belongie 2017; Huang et al. 2018; Liu et al. 2019) layer, attention mechanism (Zhu et al. 2019; Lathuilière et al. 2019; Park and Lee 2019), deformation operation (Siarohin et al. 2018; Dong et al. 2018), and GAN-based method (Bao et al. 2018) have all seen a wide adoption. Similar idea has been applied to few-shot face reenactment settings such as the use of image-level (Wiles, Koepke, and Zisserman 2018) and feature-level (Siarohin et al. 2019) warping, and AdaIN layer in conjuction with a meta-learning (Zakharov et al. 2019). The identity mismatch problem has been studied through methods such as CycleGAN-based landmark transformers (Wu et al. 2018) and landmark swappers (Zhang et al. 2019). While effective, these methods either require an independent model per person or a dataset with image pairs that may be hard to acquire.

Conclusions

In this paper, we have proposed a framework for few-shot face reenactment. Our proposed image attention block and target feature alignment, together with the landmark transformer, allow us to handle the identity mismatch caused by using the landmarks of a different person. Proposed method do not need additional fine-tuning phase for identity adaptation, which significantly increases the usefulness of the model when deployed in-the-wild. Our experiments including human evaluation suggest the excellence of the proposed method.

One exciting avenue for future work is to improve the landmark transformer to better handle the landmark disentanglement to make the reenactment even more convincing.

References

Averbuch-Elor, H.; Cohen-Or, D.; Kopf, J.; and Cohen, M. F. 2017. Bringing portraits to life. *ACM Transactions on Graphics* 36(6):196.

Balakrishnan, G.; Zhao, A.; Dalca, A. V.; Durand, F.; and Guttag, J. 2018. Synthesizing images of humans in unseen poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8340–8348.

Baltrusaitis, T.; Zadeh, A.; Lim, Y. C.; and Morency, L.-P. 2018. Openface 2.0: Facial behavior analysis toolkit. In *Proceedings of the International Conference on Automatic Face and Gesture Recognition*, 59–66.

Bao, J.; Chen, D.; Wen, F.; Li, H.; and Hua, G. 2018. Towards open-set identity preserving face synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6713–6722.

Blanz, V., and Vetter, T. 1999. A morphable model for the synthesis of 3d faces. In *Proceedings of the ACM SIGGRAPH Conference on Computer Graphics*, volume 99, 187–194.

Bulat, A., and Tzimiropoulos, G. 2017. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *Proceedings of the International Conference on Computer Vision*.

Cao, C.; Weng, Y.; Zhou, S.; Tong, Y.; and Zhou, K. 2013. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics* 20(3):413–425.

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

Dong, H.; Liang, X.; Gong, K.; Lai, H.; Zhu, J.; and Yin, J. 2018. Soft-gated warping-gan for pose-guided person image synthesis. In *Advances in Neural Information Processing Systems*, 474–484.

Huang, X., and Belongie, S. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the International Conference on Computer Vision*, 1501–1510.

Huang, X.; Liu, M.-Y.; Belongie, S.; and Kautz, J. 2018. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision*, 172–189.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1125–1134.

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* 37(4):163.

Lathuilière, S.; Sangineto, E.; Siarohin, A.; and Sebe, N. 2019. Attention-based fusion for multi-source human image generation. *arXiv preprint arXiv:1905.02655*.

Liu, M.-Y.; Huang, X.; Mallya, A.; Karras, T.; Aila, T.; Lehtinen, J.; and Kautz, J. 2019. Few-shot unsupervised image-to-image translation. *arXiv preprint arXiv:1905.01723*.

Nagrani, A.; Chung, J. S.; and Zisserman, A. 2017. Voxceleb: a large-scale speaker identification dataset. *Proceedings of the Annual Conference of the International Speech Communication Association*.

Park, D. Y., and Lee, K. H. 2019. Arbitrary style transfer with style-attentional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5880–5888.

Siarohin, A.; Sangineto, E.; Lathuilière, S.; and Sebe, N. 2018. Deformable gans for pose-based human image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3408–3416.

Siarohin, A.; Lathuilière, S.; Tulyakov, S.; Ricci, E.; and Sebe, N. 2019. Animating arbitrary objects via deep motion transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Thies, J.; Zollhöfer, M.; Nießner, M.; Valgaerts, L.; Stamminger, M.; and Theobalt, C. 2015. Real-time expression transfer for facial reenactment. *ACM Transactions on Graphics* 34(6):183–1.

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/CVF Conference on Computer Vision and Pattern Recognition*, 2387–2395.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, 5998–6008.

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. In *Proceedings of the European Conference on Computer Vision*.

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*, 603–619.

Xu, R.; Zhou, Z.; Zhang, W.; and Yu, Y. 2017. Face transfer with generative adversarial network. *arXiv preprint arXiv:1710.06090*.

Zakharov, E.; Shysheya, A.; Burkov, E.; and Lempitsky, V. 2019. Few-shot adversarial learning of realistic neural talking head models. *arXiv preprint arXiv:1905.08233*.

Zhang, J.; Zeng, X.; Pan, Y.; Liu, Y.; Ding, Y.; and Fan, C. 2019. Faceswapnet: Landmark guided many-to-many face reenactment. *arXiv preprint arXiv:1905.11805*.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2223–2232.

Zhu, Z.; Huang, T.; Shi, B.; Yu, M.; Wang, B.; and Bai, X. 2019. Progressive pose attention transfer for person image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2347–2356.