Exploiting Multiple Guidance from 3DMM for Face Reenactment

Huayu Zhang[1], **Yurui Ren**[1], **Yuanqi Chen**[1], **Ge Li**[1], **Thomas H. Li**[1,2*]

[1] School of Electronic and Computer Engineering, Peking University
[2] Information Technology R&D Innovation Center of Peking University
zhanghy@stu.pku.edu.cn, {yrren, cyq373, thomas}@pku.edu.cn, geli@ece.pku.edu.cn

Abstract

Face reenactment aims to transfer the expression and pose from a driving face to a source face. Great progress has been made on this task with the recent success of deep generative models. However, it remains challenging when the two faces hold a significant pose discrepancy. Identity change and image distortion arise as pose discrepancy increases. To tackle these problems, we propose to exploit multiple guidance derived from the 3D morphable face model (3DMM). Firstly, a precomputed optical flow is utilized to guide the estimation of motion fields. Secondly, a precomputed occlusion map is utilized to guide the perception of occluded areas. Finally, a rendered image is utilized to guide the restoration of missing contents. We present a new reenactment framework to integrate the above guidance and generate high-quality results. Extensive experiments show the superior performance of our framework compared with several state-of-the-art methods. Ablation studies demonstrate the effectiveness of exploiting multiple guidance from the 3DMM.

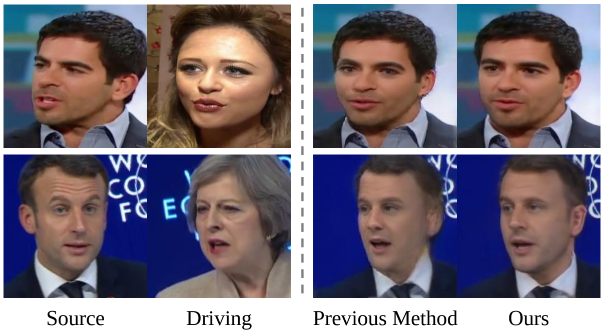

Figure 1: Challenging cases of face reenactment under a significant pose discrepancy. The first row and the second row show the problems of identity change and image distortion, respectively. While our method can alleviate these problems and achieve a clear improvement.

Introduction

Given a source face and a driving face, face reenactment refers to the task of generating a reenacted face with expression and pose from the driving face and identity from the source face. This task has a wide range of practical applications such as film production, next-generation communication, and role-playing video games.

Traditional solutions (Thies et al. 2015, 2016) to face reenactment mainly come from the graphics community. With advances in generative adversarial networks (GANs) (Goodfellow et al. 2014), learning-based methods have demonstrated impressive results. The commonly used strategy is to decouple appearance and motion information from input images and obtain motion descriptors such as facial landmarks. Based on this strategy, some methods (Wu et al. 2018; Zakharov et al. 2019; Zhang et al. 2019; Ha et al. 2020) generate reenacted results by translating motion descriptors into images. Another kind of method (Wiles, Koepke, and Zisserman 2018; Siarohin et al. 2019a,b; Zhao and Zhang 2022) predicts dense motion fields to deform source images in the pixel or feature domain. Nevertheless, it remains challenging for these methods to generate realistic results when there exists a significant pose discrepancy between the source face and the driving face. The large pose discrepancy not only makes it difficult to capture long-distance motion patterns, but the caused occlusions place high demands on restoring missing contents.

To deal with pose discrepancy, 3D information is critically important since the head motions happen in the 3D physical world. Compared with 2D photographic depictions, 3D representations contain richer spatial information that is beneficial for modeling large motions. Recently, several methods (Yao et al. 2020; Doukas, Zafeiriou, and Sharmanska 2021; Ren et al. 2021; Hong et al. 2022) have been proposed to utilize 3D information for more accurate motion estimation. However, proper handling of occlusions with 3D information is also important while rarely considered.

In this paper, we propose to exploit multiple guidance derived from the 3D morphable face model (3DMM) (Blanz and Vetter 1999) for both accurate motion estimation and proper handling of occlusions. Based on the 3D face representation, we first compute an optical flow for the inner face, which serves as a good starting point to guide the motion estimation for the entire image. To handle occlusions, we then compute an occlusion map for the inner face to indicate the occluded areas and guide the perception of occlu-

[*] Corresponding author.

sions. Finally, to enhance the restoration of missing contents caused by occlusions, we render the target 3D face into a 2D image to guide the generation of faithful facial structure. A new reenactment framework is presented to utilize the above guidance to generate high-quality reenacted results. We evaluate our framework qualitatively and quantitatively on two public datasets: VoxCeleb (Nagrani, Chung, and Zisserman 2017) and CelebV (Wu et al. 2018). Experimental results show the proposed multiple guidance can facilitate the generation of accurate facial motions and photo-realistic results, even under a significant pose discrepancy. Our main contributions can be summarized as follows:

- We propose to utilize multiple guidance provided by the 3DMM for face reenactment, including computed optical flow, computed occlusion map, and rendered mesh.
- We present a new end-to-end reenactment framework, which effectively incorporates the multiple guidance to estimate accurate motion fields, properly handle occlusions, and generate photo-realistic results.
- Experiments demonstrate that our framework can alleviate the problems caused by a significant pose discrepancy and achieve superior performance compared with several state-of-the-art methods.

Related Work

3D Morphable Face Model

3D Morphable Face Model (3DMM) (Blanz and Vetter 1999) is a parametric face model that represents 3D facial shapes and textures with a set of parameters. To achieve better facial modeling, several variants (Paysan et al. 2009; Cao et al. 2013) have been proposed to represent faces with fine-grained semantic parameters of identity, expression, pose, etc. These models produce 3D face representations from multiple face scans using principal component analysis (PCA). We use Basel Face Model (BFM) (Paysan et al. 2009) in our framework.

Face Reenactment

2D-based methods 2D-based methods directly decouple appearance and motion information from input 2D images, which can be roughly classified into translation-based methods and warping-based methods.

Translation-based methods model the task of face reenactment as an image translation problem (Isola et al. 2017; Zhu et al. 2017). For instance, ReenactGAN (Wu et al. 2018) translates the transformed facial landmarks of a specific source face into reenacted images with an encoder-decoder model. To generalize to unseen persons, Zakharov et al. (2019) propose to synthesize a neural talking head with several images of that person. Landmarks and source appearance are fused in the generator to produce reenacted faces. Nevertheless, fine-tuning is still required for unseen identities due to the use of meta-learning mechanism. Zhang et al. (2019) introduce a one-shot framework trained in an unsupervised manner, which only needs a single source image. Reenacted images are generated by fusing the source appearance with the target face parsing maps. However, the results are distorted when reenacting a different identity. Burkov et al. (2020) utilize pose augmentation to boost the performance of cross-identity reenactment. Bi-layer (Zakharov et al. 2020) generates two components from perspectives of low frequency and high frequency separately and fuses them to obtain final results. Since the used motion descriptors such as landmarks and parsing maps are subject-specific, the identity preservation problem arises when the driving face holds a clearly different facial shape from the source face. Ha et al. (2020) propose a landmark transformer to alleviate this problem, while the fidelity of reenacted images is still not high enough.

Warping-based methods model the relative motion between the source face and the driving face with dense motion fields such as optical flow, which are used to deform source images into reenacted ones. X2Face (Wiles, Koepke, and Zisserman 2018) predicts optical flow from driving images, pose vectors or audio. The predictions are then used to warp source face images. However, it lacks the ability to generate contents that do not exist in the source images. Monkey-Net (Siarohin et al. 2019a) applies the estimated optical flow to deform the source face in the feature domain and achieves better results. FOMM (Siarohin et al. 2019b) models local motions around keypoints with affine transformations and obtains more accurate motion fields. To boost the performance of motion estimation, One-shot Talking Head (Wang, Mallya, and Liu 2021) extends the dimensionality of keypoints and predicts several flow fields. MRAA (Siarohin et al. 2021) introduces a region-based motion field modeling approach. Zhao and Zhang (2022) propose to use more flexible TPS transformation to replace affine transformation. DaGAN (Hong et al. 2022) predicts a face depth map to guide motion estimation and image generation. Generally, warping-based methods achieve higher fidelity than translation-based methods since the high-frequency information can be better preserved with spatial deformation. Nevertheless, the motion fields are commonly learned from scratch in a self-supervised manner due to the lack of labels. The performance will drop significantly when large motions are observed. Our method is also warping-based, while the motion field is first computed and then completed instead of learned from scratch.

3D-based methods 3D-based methods are built upon the prior knowledge of 3D face models. Kim et al. (2018) render the reenacted meshes of a 3D face model and translate them into photo-realistic images. This method requires training on a large number of images of a specific person and therefore has to be retrained for new-coming identities. Yao et al. (2020) employ a graph neural network to learn optical flow from meshes of source and driving faces, while the areas that cannot be modeled by 3D models are less considered. HeadGAN (Doukas, Zafeiriou, and Sharmanska 2021) extracts PNCC representations from reconstructed face meshes to perform conditional synthesis and takes audio features as complementary information to boost performance. SAFA (Wang, Zhang, and Li 2021) combines FOMM with 3D models to estimate flexible expressional motions, while the subject-specific facial keypoints are still involved and lead

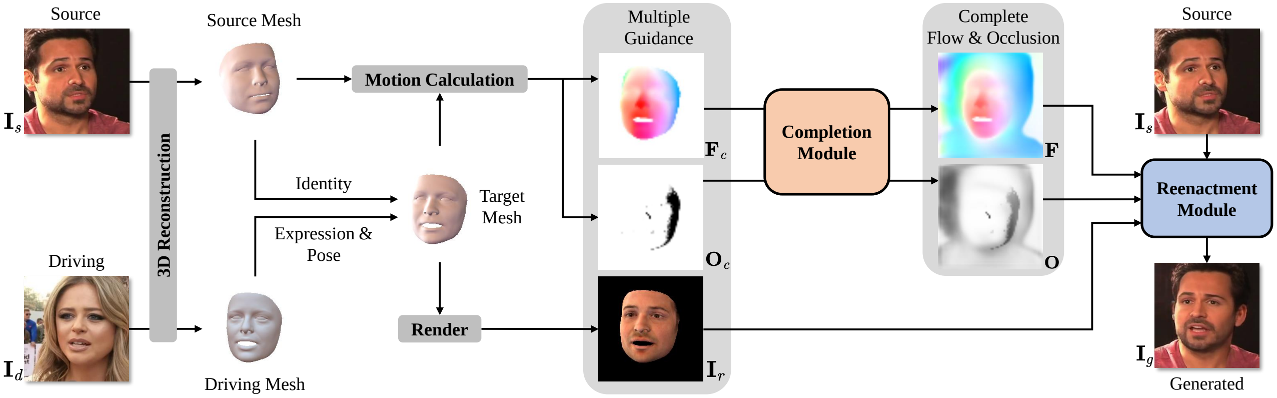

Figure 2: Overview of the proposed reenactment framework. Instead of estimating motion fields from scratch, we introduce a new scheme where the optical flow and occlusion map are first computed and then completed. The reenacted image is generated by deforming the source image with the complete optical flow and occlusion map, under the guidance of the rendered image.

to the identity preservation problem. StyleRig (Tewari et al. 2020b) uses 3DMM parameters to modulate the latent space of pre-trained StyleGAN (Karras, Laine, and Aila 2019) to control the expression and pose of generated face images. Based on this, Tewari et al. (2020a) propose an optimization-based method to compute embeddings of existing images to perform reenactment on real-world faces. More recent work PIRenderer (Ren et al. 2021) uses 3DMM parameters to predict an optical flow and control the generation process. However, the results become unrealistic when a significant identity or pose discrepancy occurs. Different from these methods, we extract multiple guidance from the 3DMM and apply them for both accurate motion estimation and proper handling of occlusions.

Methodology

Given a source image $\mathbf{I}_s$ and a driving image $\mathbf{I}_d$, we aim to generate an image $\mathbf{I}_g$ with the expression and pose of $\mathbf{I}_d$ and other attributes of $\mathbf{I}_s$ such as identity, illumination, and background. As illustrated in Fig. 2, we first perform 3D reconstruction on $(\mathbf{I}_s, \mathbf{I}_d)$ and extract multiple guidance including computed optical flow $\mathbf{F}_c$, computed occlusion map $\mathbf{O}_c$, and rendered image $\mathbf{I}_r$. Subsequently, we employ a completion module to estimate motions for the entire image based on the computed results $(\mathbf{F}_c, \mathbf{O}_c)$. Finally, we apply the complete optical flow $\mathbf{F}$ and occlusion map $\mathbf{O}$ to transform $\mathbf{I}_s$ into $\mathbf{I}_g$ in our reenactment module, which is also guided by the rendered image $\mathbf{I}_r$.

Extraction of Multiple Guidance

In this part, we perform 3D reconstruction on input images and obtain multiple guidance from reconstructed 3D faces, as shown in Fig. 2. A pre-trained face reconstruction model provided by (Deng et al. 2019) is used to extract 3DMM parameters from face images. Given a face image, the reconstruction model regresses a vector $\mathbf{v} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\gamma}, \mathbf{p}) \in \mathbb{R}^{257}$, where $\boldsymbol{\alpha} \in \mathbb{R}^{80}$, $\boldsymbol{\beta} \in \mathbb{R}^{64}$, and $\boldsymbol{\delta} \in \mathbb{R}^{80}$ represent the identity, expression, and texture parameters, respectively. $\boldsymbol{\gamma} \in \mathbb{R}^{27}$ and $\mathbf{p} \in \mathbb{R}^{6}$ refer to the illumination and face pose, respectively. The mesh topology we adopt contains $v = 35709$ vertices and $f = 70789$ triangles. The 3D coordinates $\mathbf{V} \in \mathbb{R}^{v \times 3}$ of mesh vertices can be computed with the regressed parameters.

$$\mathbf{V} = (\mathbf{V}_{mean} + \boldsymbol{\alpha}\mathbf{V}_{id} + \boldsymbol{\beta}\mathbf{V}_{exp})\mathbf{R}^{\top} + \mathbf{t} \tag{1}$$

where $\mathbf{V}_{mean}$ and $\mathbf{V}_{id}$ are the PCA bases of BFM (Paysan et al. 2009), and $\mathbf{V}_{exp}$ is built from (Cao et al. 2013). $\mathbf{R} \in \mathrm{SO}(3)$ and $\mathbf{t} \in \mathbb{R}^3$ are the rotation and translation derived from $\mathbf{p}$, respectively. With attributes $\mathbf{A} \in \mathbb{R}^{v \times d}$ assigned to the vertices, a 2D image can be rendered through rasterization.

$$\mathbf{I} = \mathcal{R}(\mathbf{V}, \mathbf{A}, \mathcal{C}) \tag{2}$$

where $\mathbf{I} \in \mathbb{R}^{h \times w \times d}$ is the rendered image. h and w denote the spatial size while d denotes the attribute dimension. $\mathcal{R}$ is the rendering function that rasterizes 3D vertex attributes into the 2D image plane. $\mathcal{C}$ is the perspective camera model. We use PyTorch3D (Ravi et al. 2020) to implement the rendering process.

Given $\mathbf{I}_s$ and $\mathbf{I}_d$, we extract their 3DMM parameters $\mathbf{v}_s$ and $\mathbf{v}_d$, which are then recombined to obtain target parameters $\mathbf{v}_t = (\boldsymbol{\alpha}_s, \boldsymbol{\beta}_d, \boldsymbol{\delta}_s, \boldsymbol{\gamma}_s, \mathbf{p}_d)$. The target mesh can be constructed with $\mathbf{v}_t$ and then rendered into an image $\mathbf{I}_r$ with computed color according to the texture and illumination of the source face. $\mathbf{I}_r$ serves as the guidance of facial structure to facilitate the restoration of missing contents caused by occlusions. The optical flow for the inner face can be computed with the constructed vertices as follows:

$$\mathbf{F}_c = \mathcal{R}(\mathbf{V}_t, \mathcal{P}(\mathbf{V}_s, \mathcal{C}), \mathcal{C}) - \mathcal{G}(h, w) \tag{3}$$

where $\mathcal{P}$ is the function that projects 3D vertices into the 2D coordinate space. $\mathcal{G}$ is used to generate a 2D coordinate grid of shape $\mathbb{R}^{h \times w \times 2}$. We additionally compute an occlusion map for the inner face to indicate the occluded areas and guide the perception of occlusions.

$$\mathbf{O}_c^{i,j} = \mathbb{1}(\mathbf{P}_t^{i,j} \in \mathbf{P}_s) \tag{4}$$

where $\mathbf{O}_c \in \mathbb{R}^{h \times w \times 1}$ is the computed occlusion map. i and j denote the indices of row and column. $\mathbf{P}_t$ and $\mathbf{P}_s$ are the

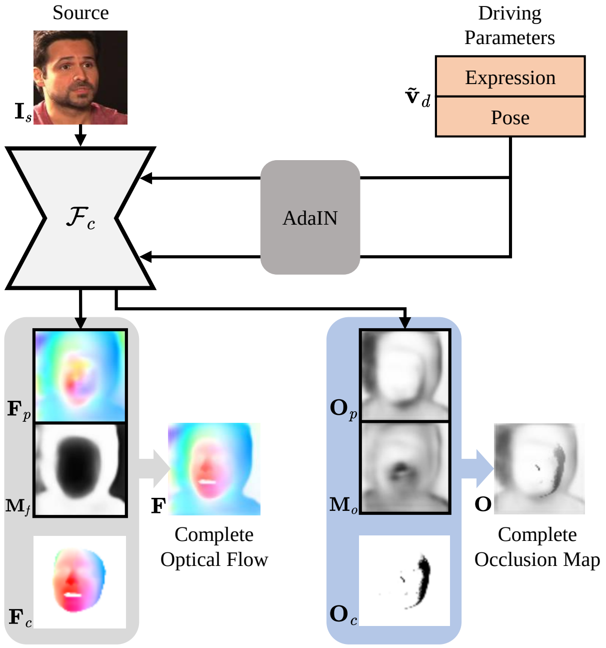

Figure 3: Illustration of the completion module.

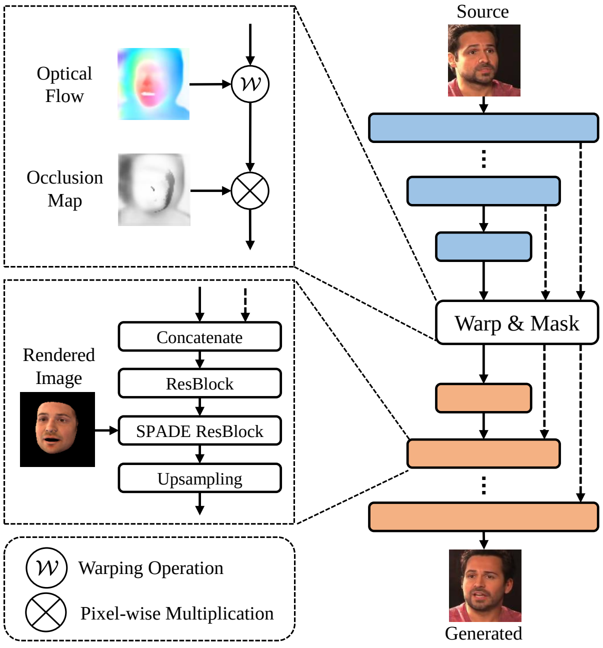

Figure 4: Implementation details of the reenactment module.

nearest triangles at each pixel of the inner facial region in $\mathbf{I}_g$ and $\mathbf{I}_s$, respectively. The indicator function $\mathbb{1}$ identifies such a rule that a target pixel is occluded if its corresponding triangle cannot be found in the set of source triangles.

Flow-and-Occlusion Completion

After obtaining the optical flow and occlusion map for the inner face, we employ a completion network $\mathcal{F}_c$ to estimate motions for the entire image based on the computed results, as shown in Fig. 3. The precomputed optical flow and occlusion map serve as good starting points to estimate final motion fields. $\mathcal{F}_c$ takes the source image $\mathbf{I}_s$ and a subset of driving parameters $\tilde{\mathbf{v}}_d = (\boldsymbol{\beta}_d, \mathbf{p}_d)$ as inputs and produces an optical flow, an occlusion map, and corresponding masks.

$$\{\mathbf{F}_p, \mathbf{O}_p, \mathbf{M}_f, \mathbf{M}_o\} = \mathcal{F}_c(\mathbf{I}_s, \tilde{\mathbf{v}}_d) \tag{5}$$

where $\mathbf{F}_p$ and $\mathbf{O}_p$ represent the predicted optical flow and occlusion map, respectively. $\mathbf{M}_f$ and $\mathbf{M}_o$ are the predicted masks for fusing the predicted and computed results. We adopt the adaptive instance normalization (AdaIN) (Huang and Belongie 2017) to inject $\tilde{\mathbf{v}}_d$ into $\mathcal{F}_c$. The final optical flow and occlusion map can be obtained as follows:

$$\mathbf{F} = \mathbf{M}_f \odot \mathbf{F}_p + (1 - \mathbf{M}_f) \odot \mathbf{F}_c \tag{6}$$

$$\mathbf{O} = \mathbf{M}_o \odot \mathbf{O}_p + (1 - \mathbf{M}_o) \odot \mathbf{O}_c \tag{7}$$

where $\odot$ is the element-wise multiplication.

With the guidance of the computed results, the completion network can pay more attention to the motion estimation of non-facial regions, which cannot be modeled by the 3DMM. The final results take advantage of the computed and predicted results to make a consistent and accurate estimation for both the facial and non-facial regions.

Reenacting

In this part, we employ a reenactment network $\mathcal{F}_r$ to transform the source image $\mathbf{I}_s$ into the reenacted image $\mathbf{I}_g$, as shown in Fig. 4. $\mathcal{F}_r$ is designed with an encoder-decoder architecture. To preserve source textures, we fuse features at multiple scales via skip connections. The encoder features are first warped and then masked before being ported to the corresponding decoder layers. The optical flow $\mathbf{F}$ is used to warp the source features to the desired regions, while the occlusion map $\mathbf{O}$ is used to mask out the occluded areas that should be inpainted. To guide the restoration of missing contents caused by occlusions and produce a faithful facial structure, we inject the rendered image $\mathbf{I}_r$ into the decoder part of $\mathcal{F}_r$ with spatially-adaptive normalization (SPADE) (Park et al. 2019).

Training

We train our framework end-to-end in a self-supervised manner. In the training stage, a pair of images $(\mathbf{I}_s, \mathbf{I}_d)$ are randomly chosen from each video to perform self-reenactment. While the identities of $\mathbf{I}_s$ and $\mathbf{I}_d$ can be different in the inference stage. The framework is trained with a reconstruction loss $\mathcal{L}_c$ and a style loss $\mathcal{L}_s$, which are based on the perceptual loss of (Johnson, Alahi, and Fei-Fei 2016). Similar to (Siarohin et al. 2019b), we downsample $\mathbf{I}_g$ and $\mathbf{I}_d$ several times and calculate $\mathcal{L}_c$ at multiple scales. $\mathcal{L}_c$ and $\mathcal{L}_s$ calculate ℓ_1 distance and statistic error between the activation maps of generated image $\mathbf{I}_g$ and ground-truth $\mathbf{I}_d$.

$$\mathcal{L}_c = \sum_{i,j} \left\| \phi_i(\mathbf{I}_g^j) - \phi_i(\mathbf{I}_d^j) \right\| \tag{8}$$

$$\mathcal{L}_s = \sum_{i} \left\| G_i^{\phi}(\mathbf{I}_g) - G_i^{\phi}(\mathbf{I}_d) \right\| \tag{9}$$

	Self-reenactment				Cross-reenactment			
	CSIM↑	AED↓	APD↓	FID↓	CSIM↑	AED↓	APD↓	FID↓
X2Face	0.613	2.679	0.2096	36.75	0.462	4.071	0.2898	50.09
FOMM	0.759	1.350	0.0379	9.80	0.521	3.329	0.0775	32.40
PIRenderer	0.760	1.322	0.0413	9.05	0.572	2.923	0.0744	21.73
DaGAN	**0.763**	1.300	**0.0366**	8.65	0.495	3.191	0.0678	36.88
Ours	0.761	**1.255**	0.0372	**8.51**	**0.625**	**2.919**	**0.0589**	**16.77**

Table 1: Quantitative comparisons on the VoxCeleb dataset. The up and down arrows represent higher and lower values for better performance. The best results are highlighted in bold.

where ϕ_i is the activation map of the i-th layer in the pretrained VGG-19 network, and j denotes downsampling j times. G_i^{ϕ} is the Gram matrix constructed from ϕ_i. The final loss used for training our framework is a weighted summation of the above losses.

$$\mathcal{L} = \mathcal{L}_c + \lambda_s \mathcal{L}_s \tag{10}$$

where we set $\lambda_s = 250$ in the experiments.

	CSIM↑	AED↓	APD↓	FID↓
X2Face	0.556	3.702	0.0761	55.72
FOMM	0.516	3.540	0.0854	39.21
PIRenderer	0.567	2.968	0.0611	26.57
DaGAN	0.500	3.442	0.0779	47.46
Ours	**0.620**	**2.944**	**0.0484**	**21.96**

Table 2: Quantitative comparisons of cross-reenactment on the CelebV dataset.

Experiments

Experimental Setup

Datasets We conduct experiments on two public datasets: VoxCeleb (Nagrani, Chung, and Zisserman 2017) and CelebV (Wu et al. 2018). The VoxCeleb dataset contains 22496 talking-head videos extracted from YouTube. Following the same pre-processing method described in (Siarohin et al. 2019b), we crop valid faces from the original videos. A total of 17913 and 514 videos with lengths varying from 64 to 1024 frames are obtained for the train and test splits, respectively. We use a similar test set sampling strategy of (Ha et al. 2020) and collect 10280 image pairs for both self-reenactment and cross-reenactment. For self-reenactment, 20 image pairs are randomly sampled from each video of the test split, while the two images in each pair are from different identities under the cross-reenactment setup. To further evaluate the performance of reenacting unseen identities, similar to the in-the-wild scenario, we construct another test set using the CelebV dataset that includes videos of five different celebrities. We randomly sample 1000 image pairs for each identity and obtain 5000 pairs in total.

Metrics We use multiple metrics to evaluate the quality of results. To measure the realism of generated images, Fréchet Inception Distance (**FID**) (Heusel et al. 2017) is utilized to estimate the difference between the distributions of the generated and real images. Besides, the quality of identity preservation is evaluated with the cosine similarity (**CSIM**) of embedding vectors from pretrained face recognition model. For the motion accuracy, following the previous work of Ren et al. (2021), we use the Average Expression Distance (**AED**) and Average Pose Distance (**APD**) to evaluate the quality of expression and pose transfer, respectively.

Training Details

We train our framework on the VoxCeleb dataset for 100 epochs using four TITAN V GPUs. The ADAM optimizer (Kingma and Ba 2014) is adopted with an initial learning rate of 1×10^{-4}. The learning rate is decreased to 1×10^{-5} after 60 epochs. The batch size is set to 8 in our experiments.

Comparisons

We evaluate the performance of our framework under two setups: self-reenactment and cross-reenactment. The evaluation results are compared with the following methods: X2Face (Wiles, Koepke, and Zisserman 2018), FOMM (Siarohin et al. 2019b), PIRenderer (Ren et al. 2021) and DaGAN (Hong et al. 2022). The absolute motions are used for all methods.

Self-reenactment We first compare the generated results under the self-reenactment setup, where the source image and the driving image are from the same person. The quantitative results are reported in Table 1. Our framework achieves the best results on AED and FID and competitive results on CSIM and APD. The improvement in AED and FID shows that our framework can capture more accurate facial motions and generate more realistic results.

Cross-reenactment We also perform comparisons under the cross-reenactment setup, where the source and driving faces are from different identities. Compared with self-reenactment, cross-reenactment is a more meaningful setup, which is required for the vast majority of practical applications. The quantitative results are provided in Table 1 and Table 2. Our framework achieves the best results on CSIM, which demonstrates that the identity information of the source face is well-preserved. The lower AED and APD indicate that our framework is able to capture facial motions more accurately with multiple guidance from the 3DMM. In addition, our framework produces more realistic results,

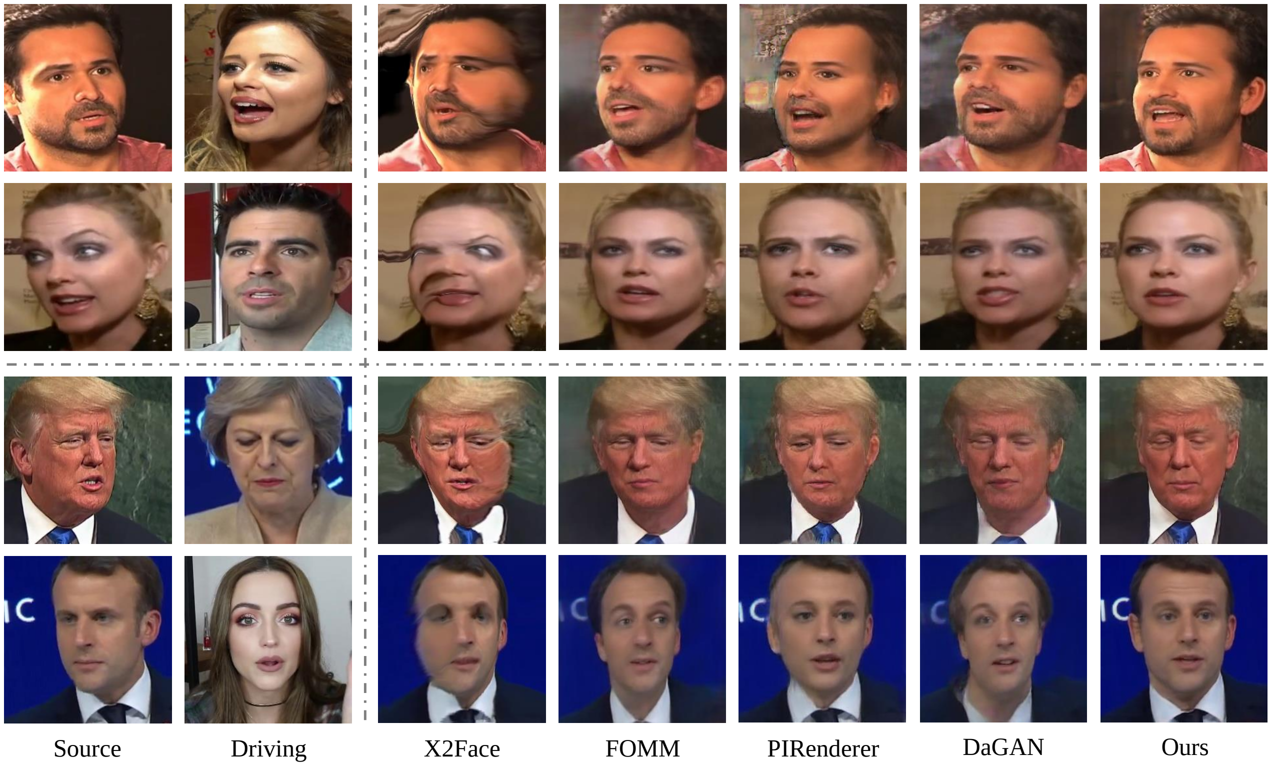

Figure 5: Qualitative comparisons of cross-reenactment on the VoxCeleb dataset (top two rows) and the CelebV dataset (bottom two rows).

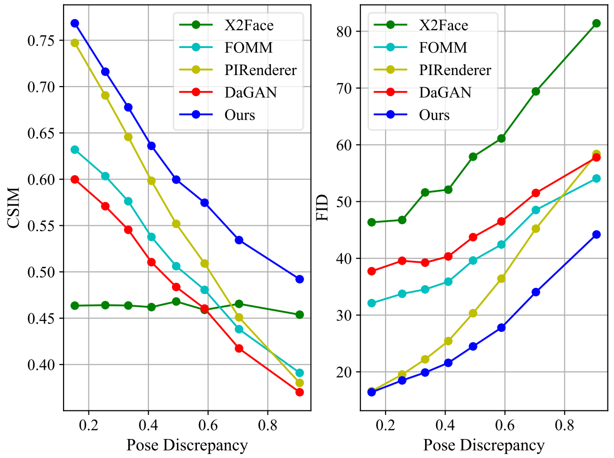

Figure 6: Quantitative comparisons of cross-reenactment on the VoxCeleb dataset under different pose discrepancies.

reaching a better FID score. To investigate the performance under different pose discrepancies between the source and driving faces, we analyze the quantitative results of cross-reenactment on the VoxCeleb dataset, as shown in Fig. 6. It can be seen that our framework can preserve source identities better and generate more realistic results under both small and large pose discrepancies.

The qualitative results are shown in Fig. 5. It can be seen that X2Face produces unrealistic distorted results. This is because it uses the predicted optical flow to deform the source image directly in the pixel domain and lacks the ability to generate reliable contents for the occluded areas. It is difficult to estimate accurate optical flow directly from the input images, especially under a large motion. By deforming the source image with the estimated optical flow in the feature domain, FOMM can produce reasonable results. However, since it uses subject-specific sparse keypoints to estimate facial motions, the facial structure is distorted when the source face and the driving face hold significantly different facial shapes. The facial expressions are not well-reenacted either. When the identities of the source face and the driving face are quite different (e.g. different genders), PIRenderer fails to faithfully preserve the source identity. There appear some leakages of identity-specific features (e.g. mustache in the second row of Fig. 5) from the driving face to the reenacted face. One possible explanation is that PIRenderer maps the driving parameters into a latent vector and injects it into the generation process. The model may overfit some specific identities during the training process, leading to the leakage of identity information. DaGAN generates more accurate facial expressions than FOMM by incorporating the learned depth map. However, the facial shapes are still inaccurate due to the use of subject-specific sparse keypoints. With the help of multiple guidance derived from the 3DMM, our method produces more realistic results with

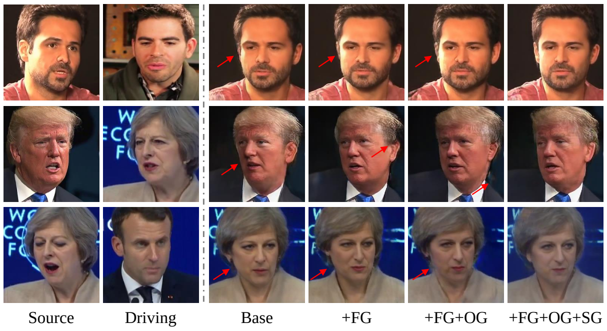

Figure 7: Qualitative results of the ablation study. The computed optical flow (flow guidance, FG), computed occlusion map (occlusion guidance, OG), and rendered image (synthesis guidance, SG) are added to the base model in turn.

	CSIM↑	AED↓	APD↓	FID↓
Base	0.595	3.195	0.0550	23.36
+FG	0.615	3.055	0.0542	24.03
+FG+OG	0.616	3.043	0.0544	22.39
+FG+OG+SG	**0.620**	**2.944**	**0.0484**	**21.96**

Table 3: Quantitative ablation results of cross-reenactment on the CelebV dataset. We add the multiple guidance in turn.

more accurate facial motions compared with the other methods, even under a large identity or pose discrepancy between the source face and the driving face. Notably, compared with PIRenderer and DaGAN, which also incorporate 3D information, our framework exploits 3D priors in a more comprehensive way. We extract multiple guidance from the 3D face meshes for both accurate motion estimation and proper handling of occlusions. By utilizing the rich priors contained in the 3DMM, our framework generates more realistic results and preserves source identities better.

Ablation Study

We perform ablation experiments on the CelebV dataset to investigate the effectiveness of the proposed multiple guidance. The optical flow and occlusion map used in the base model are directly predicted by the completion network, without guidance of the computed results. Besides, no rendered image is injected into the reenactment network. We add the proposed multiple guidance to the base model in turn and report the quantitative results in Table 3. The flow guidance (FG), occlusion guidance (OG), and synthesis guidance (SG) represent the computed optical flow, computed occlusion map, and rendered image, respectively. It can be seen that the flow guidance improves CSIM and AED, which means that it can facilitate more accurate motion estimation and benefit identity preservation. Comparing the second and third rows of Table 3, the occlusion guidance brings a clear improvement to FID, which indicates that it can facilitate the generation of realistic results. The synthesis guidance improves all four metrics by providing explicit clues for properly handling occlusions.

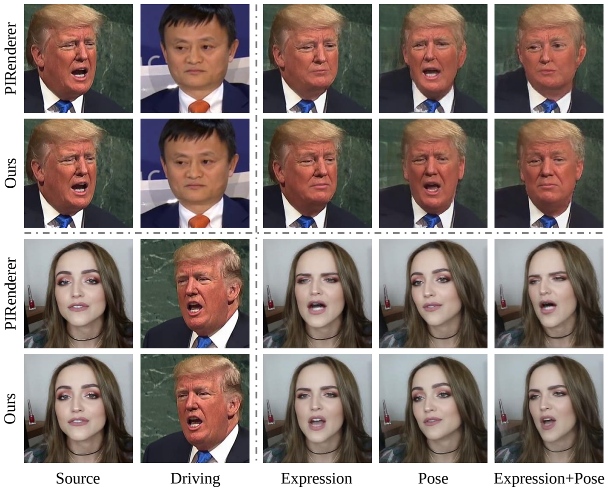

Figure 8: Qualitative results of disentangled reenactment.

The qualitative results are shown in Fig. 7. It can be seen that the base model suffers from distortions of facial shapes due to predicted inaccurate motion fields. By incorporating the flow guidance as a good starting point for motion estimation, the distortions of facial shapes are reduced. With explicit guidance for the perception of occlusions, the visual quality improves to a certain extent. The synthesis guidance provides semantic information such as facial structure and texture, further helping generate more accurate facial motions and realistic results.

Disentangled Reenactment

Since the 3DMM parameters are fully disentangled, it is possible to reenact one of the expression and pose alone while keeping the other one unchanged. The reenacted results are shown in Fig. 8. It can be seen that our framework can reenact expression or pose independently and generate more realistic images than PIRenderer (Ren et al. 2021).

Conclusion

In this paper, we propose to exploit multiple guidance derived from the 3DMM and present a new face reenactment framework. The multiple guidance not only helps with more accurate motion estimation but also facilitates proper handling of occlusions. Our framework can generate realistic results with accurate facial motions, even under a significant head pose discrepancy between the source face and the driving face. Experimental results show the superior performance of our framework compared with several state-of-the-art methods. Ablation studies demonstrate the effectiveness of the proposed multiple guidance.

Acknowledgments

This work is supported by National Natural Science Foundation of China (No. 62172021) and Shenzhen Fundamental Research Program (GXWD20201231165807007-20200806163656003). We thank all reviewers for their valuable feedback and helpful comments.

References

Blanz, V.; and Vetter, T. 1999. A morphable model for the synthesis of 3D faces. In *SIGGRAPH*.

Burkov, E.; Pasechnik, I.; Grigorev, A.; and Lempitsky, V. 2020. Neural head reenactment with latent pose descriptors. In *CVPR*.

Cao, C.; Weng, Y.; Zhou, S.; Tong, Y.; and Zhou, K. 2013. Facewarehouse: A 3d facial expression database for visual computing. *TVCG*.

Deng, Y.; Yang, J.; Xu, S.; Chen, D.; Jia, Y.; and Tong, X. 2019. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *CVPRW*.

Doukas, M. C.; Zafeiriou, S.; and Sharmanska, V. 2021. Headgan: One-shot neural head synthesis and editing. In *ICCV*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *NIPS*.

Ha, S.; Kersner, M.; Kim, B.; Seo, S.; and Kim, D. 2020. Marionette: Few-shot face reenactment preserving identity of unseen targets. In *AAAI*.

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 *NIPS*.

Hong, F.-T.; Zhang, L.; Shen, L.; and Xu, D. 2022. Depth-Aware Generative Adversarial Network for Talking Head Video Generation. In *CVPR*.

Huang, X.; and Belongie, S. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *CVPR*.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *CVPR*.

Kim, H.; Garrido, P.; Tewari, A.; Xu, W.; Thies, J.; Niessner, M.; Pérez, P.; Richardt, C.; Zollhöfer, M.; and Theobalt, C. 2018. Deep video portraits. *TOG*.

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

Nagrani, A.; Chung, J. S.; and Zisserman, A. 2017. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*.

Park, T.; Liu, M.-Y.; Wang, T.-C.; and Zhu, J.-Y. 2019. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*.

Paysan, P.; Knothe, R.; Amberg, B.; Romdhani, S.; and Vetter, T. 2009. A 3D face model for pose and illumination invariant face recognition. In *AVSS*.

Ravi, N.; Reizenstein, J.; Novotny, D.; Gordon, T.; Lo, W.-Y.; Johnson, J.; and Gkioxari, G. 2020. Accelerating 3d deep learning with pytorch3d. *arXiv preprint arXiv:2007.08501*.

Ren, Y.; Li, G.; Chen, Y.; Li, T. H.; and Liu, S. 2021. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *ICCV*.

Siarohin, A.; Lathuilière, S.; Tulyakov, S.; Ricci, E.; and Sebe, N. 2019a. Animating arbitrary objects via deep motion transfer. In *CVPR*.

Siarohin, A.; Lathuilière, S.; Tulyakov, S.; Ricci, E.; and Sebe, N. 2019b. First order motion model for image animation. In *NIPS*.

Siarohin, A.; Woodford, O. J.; Ren, J.; Chai, M.; and Tulyakov, S. 2021. Motion representations for articulated animation. In *CVPR*.

Tewari, A.; Elgharib, M.; Bernard, F.; Seidel, H.-P.; Pérez, P.; Zollhöfer, M.; and Theobalt, C. 2020a. Pie: Portrait image embedding for semantic control. *TOG*.

Tewari, A.; Elgharib, M.; Bharaj, G.; Bernard, F.; Seidel, H.-P.; Pérez, P.; Zollhofer, M.; and Theobalt, C. 2020b. Stylerig: Rigging stylegan for 3d control over portrait images. In *CVPR*.

Thies, J.; Zollhöfer, M.; Nießner, M.; Valgaerts, L.; Stamminger, M.; and Theobalt, C. 2015. Real-time expression transfer for facial reenactment. *TOG*.

Thies, J.; Zollhofer, M.; Stamminger, M.; Theobalt, C.; and Nießner, M. 2016. Face2face: Real-time face capture and reenactment of rgb videos. In *CVPR*.

Wang, Q.; Zhang, L.; and Li, B. 2021. SAFA: Structure Aware Face Animation. In *3DV*.

Wang, T.-C.; Mallya, A.; and Liu, M.-Y. 2021. One-shot free-view neural talking-head synthesis for video conferencing. In *CVPR*.

Wiles, O.; Koepke, A.; and Zisserman, A. 2018. X2face: A network for controlling face generation using images, audio, and pose codes. In *ECCV*.

Wu, W.; Zhang, Y.; Li, C.; Qian, C.; and Loy, C. C. 2018. Reenactgan: Learning to reenact faces via boundary transfer. In *ECCV*.

Yao, G.; Yuan, Y.; Shao, T.; and Zhou, K. 2020. Mesh guided one-shot face reenactment using graph convolutional networks. In *ACM MM*.

Zakharov, E.; Ivakhnenko, A.; Shysheya, A.; and Lempitsky, V. 2020. Fast bi-layer neural synthesis of one-shot realistic head avatars. In *ECCV*.

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

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

Zhao, J.; and Zhang, H. 2022. Thin-Plate Spline Motion Model for Image Animation. In *CVPR*.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*.