Pose Adapted Shape Learning for Large-Pose Face Reenactment

Gee-Sern Jison Hsu Jie-Ying Zhang Huang Yu Hsiang Wei-Jie Hong
National Taiwan University of Science and Technology, Taipei, Taiwan
{jison, m11151021, m11103421 ,m11003424}@mail.ntust.edu.tw

Abstract

We propose the Pose Adapted Shape Learning (PASL) for large-pose face reenactment. The PASL framework consists of three modules, namely the Pose-Adapted face Encoder (PAE), the Cycle-consistent Shape Generator (CSG), and the Attention-Embedded Generator (AEG). Different from previous approaches that use a single face encoder for identity preservation, we propose multiple Pose-Adapted face Encodes (PAEs) to better preserve facial identity across large poses. Given a source face and a reference face, the CSG generates a recomposed shape that fuses the source identity and reference action in the shape space and meets the cycle consistency requirement. Taking the shape code and the source as inputs, the AEG learns the attention within the shape code and between the shape code and source style to enhance the generation of the desired target face. As existing benchmark datasets are inappropriate for evaluating large-pose face reenactment, we propose a scheme to compose large-pose face pairs and introduce the MPIE-LP (Large Pose) and VoxCeleb2-LP datasets as the new large-pose benchmarks. We compared our approach with state-of-the-art methods on MPIE-LP and VoxCeleb2-LP for large-pose performance and on VoxCeleb1 for the common scope of pose variation.

1. Introduction

Face reenactment refers to the transformation of the action of a reference face to a source face, and the action includes facial pose and expression. It is required that the reference action can be duplicated by the source and the source identity can be preserved after the transformation. The requirements can be concisely stated as the reference action transformation and source identity preservation. As face reenactment has a broad scope of applications in fields such as animation, film making, virtual reality, and others, it received increasing attention with many approaches proposed in recent years [2, 9, 14, 24, 26, 29, 31–33].

Many approaches rely on a face encoder to warrant the source identity preservation. VGGFace2 [4] and ArcFace [7] are two popular choices to many reenactment approaches [2, 3, 14, 20, 28]. Most, if not all, face encoders are made by pose-biased training, i.e., the training set is imbalanced in pose. More data are close to frontal pose and fewer data are with large or extreme pose. When encoding faces within median poses, i.e., yaw$< 45°$, these face encoders serve well as the encoding confidence is high. However, when encoding faces of large or extreme poses, the encoding confidence drops as the encoders are trained on less data of such poses. We propose the Pose-Adapted face Encoders (PAEs) to address this issue and better handle the identity preservation for large-pose face reenactment. Each PAE learns the identity features for a limited pose range. We designed 6 PAEs to cover all pose differences between the source and reference. The Pose-Adapted Encoding can be considered as a piece-wise approximate solution to encoding the nonlinear facial pose variation.

We propose the Cycle-consistent Shape Generator (CSG) and the Attention-Embedded Generator (AEG) in the framework to work with the PAEs for handling large-pose reenactment. Given a source face and a reference face as inputs, the CSG generates a shape code that fuses the source identity and reference action in the shape space and meets the cycle-consistency requirement. Taking the shape code and the source as inputs, the AEG learns the self-attention within the shape code and cross-attention between the shape code and source style to enhance the generation of the desired target face. Due to the fact that current benchmark datasets are inappropriate for the evaluation of large-pose face reenactment, we propose a scheme to select large-pose face pairs and construct the MPIE-LP (Large Pose) and VoxCeleb2-LP datasets from their originals.

We summarize the contributions of this work as follows:

- The PAEs (Pose-Adapted face Encoders) are verified effective for measuring and preserving facial identity across large-pose variation, and can be applied to other work where large-pose identity preservation is a concern.
- The proposed framework with cycle-consistent shape learning and attention-embedded generation is novel in architecture and outperforms state-of-the-art approaches for large-pose face reenactment.

- The scheme proposed to make the large-pose benchmark datasets, the MPIE-LP and VoxCeleb2-LP, is simple but effective in addressing the issue that existing datasets are inappropriate for evaluating large-pose performance.

The code and pretrained model are available on https://github.com/AvLab-CV/PASL.

In the following, we first review previous work in Sec. 2, then present our approach in Sec. 3, then the experiments in Sec. 4, and a conclusion to this work in Sec. 5.

2. Related Work

Most approaches can be generally divided into three categories, the warping-based, the landmark-based and the hybrid of both. For warping-based approaches, the search for the description of the motion field good for action transformation and identity preservation is considered [24, 30]. A first-order approximation to the motion field described by the keypoint-based optical flow is proposed in the First Order Motion Model (FOMM) [24]. It consists of a keypoint detector, a motion network, and a generator. The motion network takes the keypoint motion representation to describe the optical flow from the reference to the source. The generator takes the optical flow and an occlusion map to couple with the source image and the reference action for target face generation. The Mesh Guided One-Shot (MGOS) [30] learns the optical flow from 3D meshes to provide the shape and pose to reconstruct the reference action on the source.

For landmark-based approaches, the facial landmarks are explored to combine the reference action and source identity and guide the synthesis of the target (reenacted) face [14, 31, 33]. Few-Shot Talking Head [31] trains an embedder to encode the source landmarks, and a pair of generator and discriminator to transfer the reference action to the source. FReeNet [33] trains a landmark converter to transfer the reference landmarks to the source, and a generator to make the target face show the reference expression. Although FReeNet shows good performance in transferring facial expression, it cannot handle pose transformation. The Dual Generator (DG) [14] fuses reference landmark code and source facial code to generate the identity-preserving landmarks that keep the source identity while transferring the reference action. 3D landmarks are employed to deal with large-pose reenactment. The DG can be a pioneer in addressing large pose issues. However, the way to evaluate the performance must be modified as most test data do not reveal sufficiently large poses. This is the reason that we propose a scheme to make new benchmark datasets, the MPIE-LP and VoxCeleb2-LP.

The hybrid approaches consider both the warping motion fields and facial landmarks [9, 29, 32]. The HeadGAN [9] uses the 3DMM [35] to decompose a face into expression and identity parameters, and combines source identity and reference expression to synthesize target face. However, identity preservation that solely depends on source identity parameter cannot preserve sufficient identity characteristics. The Bi-layer [32] consists of a pose-dependent coarse layer and a pose-independent texture layer. Facial keypoints are used to predict the coarse component of target face with a warping field for merging with the texture. The Face2Face$^{\rho}$ [29] has a motion network for predicting the motion field and a rendering network driven by pose and motion field. During training, the 3DMM parameters of the source and reference are regressed and used to reconstruct three landmark images. The landmark images and the source images are fed into the motion network with the source images sent to the rendering module for generating the reenacted face. The HyperReenact [2] converts faces to the StyleGAN2 [16] latent space and uses a hypernetwork to refine source identity characteristics and conduct facial pose re-targeting, eliminating external editing that may cause artifacts. The proposed Pose Adapted Shape Learning (PASL) belongs to the hybrid category.

3. Proposed Approach

The proposed PASL framework is composed of three modules, the Pose-Adapted face Encoder (PAE), the Cycle-consistent Shape Generator (CSG) and the Attention-Embedded Generator (AEG). The configuration is shown in Figure 1. The PAE is proposed to better preserve facial identity across large-pose reenactment. Given a source I_s and a reference I_r as inputs, the CSG generates a recomposed shape s_{rc} that fuses the source identity and reference action in the shape space and meets the cycle-consistency requirement. Taking the recomposed shape s_{rc} and the source style code c_s as inputs, the AEG is trained to learn the attention within the shape code and the cross-attention between the shape code and source style to better generate the desired target face. Details of these modules are presented in the following sections.

3.1. Pose-Adapted face Encoder (PAE)

As mentioned in Sec. 1, most face encoders suffer from biased training, i.e., training on pose-imbalanced datasets. The consequence is that these encoders encode nearly frontal faces with high confidence, and the confidence decreases as the pose changes to profile. When using such frontal-biased encoders for identity-preserving training, the facial identity will not be well preserved if the source or reference pose is close to profile. The proposed 6 PAEs aim to address this issue.

We divide facial pose into three sets according to the absolute yaw angle θ_y, and label them as 'f' for frontal with $\theta_y < 30°$, 's' for side with $30° \leq \theta_y < 60°$, and 'p' for profile with $60° \leq \theta_y$. We consider the following six pose pairs: frontal-vs-frontal (ff), side-vs-side (ss), profile-

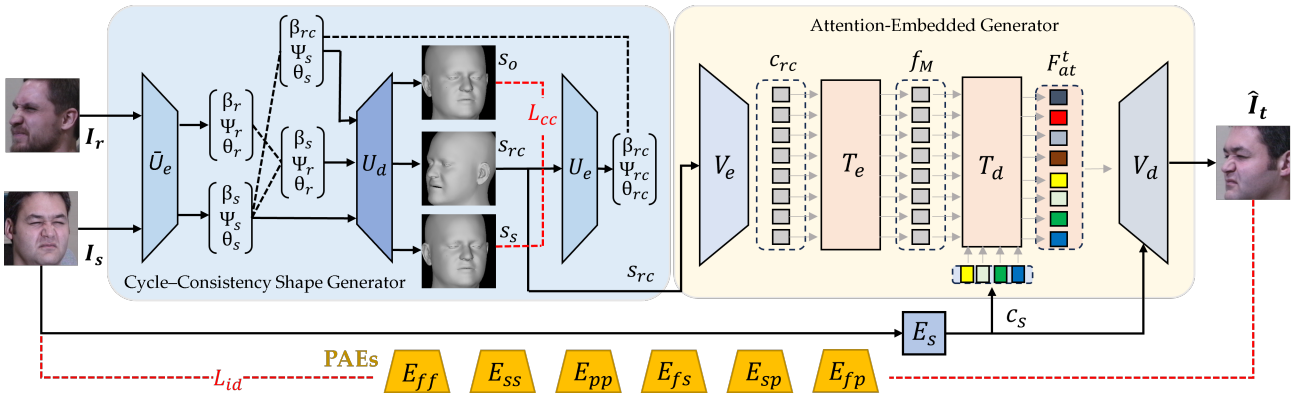

Figure 1. PASL model composed of 6 Pose-Adapted Encoders (PAEs), Cycle-consistent Shape Generator (CSG), Attention-Embedded Generator (AEG). Red dashed lines indicate losses, black dashed lines indicate recomposed shape parameters

vs-profile (pp), frontal-vs-side (fs), side-vs-profile (sp) and frontal-vs-profile (fp) when measuring the identity similarity between the source and the reenacted output. Six Pose-Adapted Encoders (PAEs), denoted as E_{ff}, E_{ss}, E_{pp}, E_{fs}, E_{sp} and E_{fp}, are trained on the corresponding pose-conditioned datasets and explored for pose-adapted identity preservation. The encoder E_{ff} is trained on nearly frontal faces with yaw angle $\theta_y < 30°$. E_{ss} is trained on faces with yaw angle $30° \leq \theta_y < 50°$. E_{pp} is trained on faces with yaw angle $\theta_y \geq 50°$. E_{fs}, E_{sp}, and E_{fp} are trained on faces with poses uniformly distributed within specific orientation ranges. E_{fs} is trained on faces with poses from frontal to side ($0° \sim 50°$). E_{sp} is trained on faces with poses from side to profile ($30° \sim 90°$). E_{fp} is trained on faces with poses in frontal and profile ranges $0° \sim 30°$ and $50° \sim 90°$, i.e., the union of the two sets.

Assuming that face is symmetry in yaw, we flip the faces with $\theta_y < 0°$ to the other side, i.e., $|\theta_y|$. Therefore, the similarity between, for example, $\theta_y^1 < 0°$ and $\theta_y^2 > 0°$ is considered the same as between $|\theta_y^1|$ and θ_y^2. See Sec. 4 for more details about implementation and experiments.

3.2. Cycle-consistent Shape Generator

Several recent approaches exploit the decomposition of the identity and expression shapes from a 3D model, and combine the source identity and reference expression shapes to synthesize target face [9, 29]. The most popular 3D model considered is the 3DMM [1]. The Cycle-consistent Shape Generator (CSG) in our framework was also designed for the recomposition of source identity and reference expression shapes. However, two properties make CSG a different design. The first is that CSG is modified from the DECA model [10], which offers a better decomposition of identity and expression shapes than the 3DMM. The second is the cycle consistency that we built into the CSG to make the output more consistent with the source input in the identity shape space than the original DECA.

DECA [10] is an extension of the FLAME model [19] which reconstructs a 3D face by using identity shape parameter β, expression shape parameter ψ, and pose parameter θ. DECA revised FLAME by using two encoders and three decoders. One encoder converts a 2D face into a latent code that includes the FLAME parameters $\{\beta, \psi, \theta\}$. The other encoder converts the face into another latent code that captures expression details. $\{\beta, \psi, \theta\}$ are decoded to a reconstructed shape by a decoder. As DECA and FLAME aim for animation, they only consider the expression and pose variation for the same subject. For our framework, we do not just consider the expression and pose transfer across different subjects, we also impose cycle-consistent shape generation that requires the reconstructed source shape to be encoded and re-decoded back to the input source shape with the original expression and pose.

Our CSG redesigns the workflow and builds in the cycle consistency that makes the reconstructed output close to the source input. Figure 1 shows the configuration of the CSG. The encoder $\bar{U}_e$ encodes the source I_s into the latent code $g_s = [\beta_s, \psi_s, \theta_s]$ and the reference I_r into $g_r = [\beta_r, \psi_r, \theta_r]$. The latent codes g_s and g_r can be decoded to shapes s_s and s_r by the decoder U_d. For recomposition, U_d decodes the concatenation of source identity code and reference action code, $[\beta_s, \psi_r, \theta_r]$, to a recomposed shape s_{rc}. To the best of our knowledge, the previous work that takes advantage of the identity and action decomposition of a 3D model all ends up using this recomposed shape s_{rc} to make the target face [9, 29]. We instead consider s_{rc} an intermediate shape which can be encoded by U_e to $[\beta_{rc}, \psi_{rc}, \theta_{rc}]$. The recomposed identity code β_{rc} and the original source action code form another recomposed code $[\beta_{rc}, \psi_s, \theta_s]$, which can be decoded by U_d to an output shape s_o. Note that $\bar{U}_e$ and U_e are the same encoders, but the former takes face image as input and the latter takes shape map and an add-in texture as input (see [10] for details). In practice, U_e and U_d are initialized by the DECA encoder and decoder, respectively,

and retrained to minimize the cycle-consistent shape loss defined in (4).

3.3. Attention-Embedded Generator

The AEG (Attention-Embedded Generator) is designed to learn the self-attention features of the recomposed shape s_{rc} and the cross-attention characteristics between the recomposed shape s_{rc} and the source style for improving the target face generation. It is made of an encoder V_e, a decoder V_d, a transformer T, a style encoder E_s, and a discriminator D_f. The configuration is shown in Figure 1. The transformer T works between the encoder V_e and decoder V_d so that the input to T is a latent code and the output is an attention feature code. The encoder V_e, made of four res-blocks, converts the recomposed shape s_{rc} to a latent code c_{rc}. The transformer T takes the latent code c_{rc} and the source style code c_s as inputs to generate an attention feature code F^t_{at}. F^t_{at} is then decoded by the decoder V_d to make the target face $\hat{I}_t$. The source style code is obtained by entering the source I_s into the style encoder E_s. The discriminator D_f is built by using the same structure as the style encoder E_s but with a 1D output to distinguish the generated $\hat{I}_t$ from the real I_s.

The transformer T consists of a multi-headed transformer encoder (t-encoder) T_e and a multi-headed decoder (t-decoder) T_d. The t-encoder T_e converts the recomposed shape code c_{rc} to a self-attention feature sequence f_M. f_M is decoded by the t-decoder T_d to an attention feature code F^t_{at}. The t-encoder T_e is made of M multi-headed self-attention layers. Each layer consists of a multi-headed self-attention module and an MLP (Multilayer Perceptron) to capture the global and local attention across the entities of the feature sequence from the last layer. To better preserve the ordering information of the feature sequence, the position encoding in [25] is added to the input feature sequence at each layer. Denote the feature sequences at layer-m and $m-1$ as f_m and f_{m-1}, respectively. The multi-headed self-attention module at layer-m first maps f_{m-1} into a triplet representation in terms of query Q_m, key K_m and value V_m as shown below.

$$Q_m = f_{m-1}\boldsymbol{W}_q, K_m = f_{m-1}\boldsymbol{W}_k, V_m = f_{m-1}\boldsymbol{W}_v \quad (1)$$

where $\boldsymbol{W}_q$, $\boldsymbol{W}_k$, and $\boldsymbol{W}_v$ are the mapping weights to learn during training for Q_m, K_m and V_m, respectively. The output $\boldsymbol{o}_{m,j}$ from the j-th head is computed as follows:

$$\boldsymbol{o}_{m,j} = \text{softmax}\left(\frac{Q_m K_m^T}{\sqrt{d_k}}\right) V_m, \quad j \in \{1, ..., N_h\} \quad (2)$$

where d_k is the dimension of K_m, and N_h is the number of heads. The outputs $\boldsymbol{o}_{m,j}$ from all N_h heads are concatenated and processed by the MLP to produce the feature sequence f_m. The above learning process is repeated for all M layers to obtain the self-attention feature sequence f_M.

Similar in structure to the t-encoder T_e, the t-decoder T_d comprises N multi-headed attention decoding layers. It performs multi-headed self-attention and encoder-decoder attention operations. The encoder-decoder attention derives the key and value vectors from the encoder output with query vectors given by the decoding layer. We take the recomposed shape s_{rc} to make the query embeddings Q. Q is considered as a learned positional encoding to each attention layer. Each query embedding learns from the attention features to infer the attributes of the source style code. The query embedding is processed in parallel at each decoding layer. The output embeddings will collect the features of the style attributes across N decoding layers and yield an attention feature code F^t_{at}. F^t_{at} is entered to the decoder V_d, which is made of 4 upsampling res-blocks. AdaIN [16] is applied to enter the style code c_s into all res-blocks of V_d to enhance style preservation at the target $\hat{I}_t$.

When training the AEG, we first trained it for self-reenactment, where the source and reference were the same subject and the ground-truth target I_t was available; and then retrained it for cross-reenactment, where the source and reference were different subjects. The following loss functions are considered:

Pose-Adapted Identity Loss The proposed 6 PAEs are exploited to compute the following identity loss that depends on the poses of the source I_s and generated target face $\hat{I}_t$.

$$\mathcal{L}_{id} = 1 - \cos(E_{id}(I_s), E_{id}(\hat{I}_t)) \quad (3)$$

where E_{id} is chosen from E_{ff}, E_{ss}, E_{pp}, E_{fs}, E_{sp} and E_{fp}, depending on the yaw angles of I_s and I_r.

Cycle-consistent Shape Loss As discussed in Sec. 3.2, the encoder U_e and decoder U_d are initialized by the DECA-pretrained encoder and decoder, respectively, and retrained to minimize the following cycle-consistent shape loss,

$$\mathcal{L}_{cc} = |U_d(\beta_s, \psi_s, \theta_s) - U_d(\beta_{rc}, \psi_s, \theta_s)|_1 \quad (4)$$

Style Loss To ensure that the generated face keeps the same style as the source, the following loss is needed to minimize the difference between the style features of $\hat{I}_t$ and I_s.

$$\mathcal{L}_{sty} = |E_s(\hat{I}_t) - E_s(I_s)|_2 \quad (5)$$

Perceptual Loss To enhance the perceptual similarity between $\hat{I}_t$ and ground truth I_t, we consider the following loss [15] computed by using the multi-layer VGG-19 features during self-reenactment training,

$$\mathcal{L}_{per} = \sum_i^N |VGG_i(\hat{I}_t) - VGG_i(I_t)|_1 \quad (6)$$

where $VGG_i()$ is the i_{th} layer feature map, and $N = 5$ determined by experiments.

Adversarial Loss To make the generated target face $\hat{I}_t$ appear as a photo-quality face, the following adversarial losses for the AEG generator, denoted as G, and the discriminator D_f are needed,

$$\mathcal{L}_G^{adv} = -\mathbb{E}_{s_{rc}\sim p(s_{rc}),c_s\sim p(c_s)} \log\left[1 - D_f(\hat{I}_t)\right] \quad (7)$$

$$\mathcal{L}_{D_f}^{adv} = \mathbb{E}_{I_s\sim p(I_s)} \log\left[D_f\left(I_s\right)\right] + \mathbb{E}_{s_{rc}\sim p(s_{rc}),c_s\sim p(c_s)} \log\left[1 - D_f(\hat{I}_t)\right] \quad (8)$$

The full objective function for training the AEG is a weighted sum of the above loss functions:

$$\mathcal{L}_{AEG} = \mathcal{L}_G^{adv} + \lambda_{id}\mathcal{L}_{id} + \lambda_{cc}\mathcal{L}_{cc} + \lambda_{sty}\mathcal{L}_{sty} + \lambda_{per}\mathcal{L}_{per} \quad (9)$$

where $\lambda_{id}, \lambda_{cc}, \lambda_{sty}, \lambda_{per}$ are the weights determined in the experiments.

4. Experiment

We first introduce the datasets and the new specifications and protocols good for evaluating large-pose performance, then an ablation study on the performance of PAEs and different settings of our model, and then a comparison with state-of-the-art approaches.

4.1. Datasets and Protocols

The cleaned MS1M-V3 dataset (5.1M images of 93k subjects) [8] was selected to train and validate the 6 Pose-Adapted face Encoders (PAEs). 85% subjects were selected for training, and the rest 15% for validation. The IJB-C dataset [27] was used to test the performance. The 3DDFA2 [12] was used to annotate the facial landmarks and head pose for each face. Six pose subsets were made from the training set to train the 6 PAEs (E_{ff}, E_{ss}, E_{pp}, E_{fs}, E_{sp}, E_{fp}), which were tested on the six pose subsets formed from the testing set. The network and feature embedding settings of each PAE followed those for making the VGGFace2 [4] with a modified ResNet-50 as the backbone. See Supplementary document for more details about the PAEs.

We selected MPIE [11], VoxCeleb1 [22]and VoxCeleb2 [6] datasets for evaluating the proposed PASL model. To better evaluate large-pose performance, we made MPIE-LP (Large Pose) and VoxCeleb2-LP subsets from MPIE and VoxCeleb2. The VoxCeleb1 was used the same way as previous approaches as this dataset did not provide enough large-pose data.

MPIE-LP is made from MPIE [11] as it provides close-to-ground-truth images for performance comparison. Due to the requirement that each subject must have 4 facial expressions, the MPIE-LP is made with 127 subjects with 13 poses across 9 illumination conditions. As the pose in MPIE is labeled by a specific yaw angle θ_y, we define $\theta_y = 0°, \pm 15°$ as frontal pose, $\theta_y = \pm 30°, \pm 45°$ as side pose, and $\theta_y = \pm 60°, \pm 75°, \pm 90°$ as profile pose. Only 4 pose pair sets (ss, pp, fp and sp), which exhibit sufficient large-pose differences between the source and reference, are included in the MPIE-LP. This results in minimum pose difference 45° (in fp set) and maximum 180° (in pp set). We split 127 subjects into 80 for training and 47 for testing. As MPIE offers near ground-truth images of large poses, it was selected for ablation study.

VoxCeleb1 offers over 100k videos of 1,251 celebrities and is divided into training and testing sets. In our experiments, images were extracted from the videos sampled at 1 fps, resized to 256^2 pixels, and each face with landmarks detected by 3DDFA2 [12]. We followed the protocol same as previous work and trained the model on the training set. Note that the experiments on VoxCeleb1 cannot reveal large-pose performance as too few data of large pose in this dataset. However, it is a fair benchmark to compare with other approaches as many reported performance on it.

VoxCeleb2-LP is made from the VoxCeleb2, which is an extension of the VoxCeleb1. VoxCeleb2 contains over 1 million utterances of 6,112 celebrities, and is divided into training and testing sets. We extracted images from each selected video at 5 fps to capture fast pose change, and processed the images the same way as we performed for VoxCeleb1. It offers more faces in large or extreme poses than other datasets, and thus is good to benchmark large-pose performance. To make VoxCeleb2-LP, we selected 1259 celebrities from the training set that contained large poses to form 4.5 million large-pose pairs for training, and 196 celebrities from the testing set to form 700k large-pose pairs for performance testing. We also segmented the data into 6 pose pair sets, same as we did when using the MS1M-V3 [8] to make the 6 PAEs. The minimum pose difference is 20° in ff pose set, and maximum 180° in pp pose set. We again selected the subsets ss, pp, fp, and sp for evaluating large-pose performance. See Supplementary document for more dataset specifications.

Evaluation Metrics were selected to test the source identity preservation, reference action transformation and image quality of the generated target faces, including the Frechet-Inception Distance (FID) [13], Cosine Similarity (CSIM), Average Rotation Distance (ARD), Learned Perceptual Image Patch Similarity (LPIPS) [34]. As these are common metrics, see Supplementary document for details.

Training began for self-reenactment with minimum 2 images per identity and then for cross-reenactment, starting from scratch to minimize the losses in $\mathcal{L}_{AEG}$. A comparison study determined the weights for the losses as $\lambda_{id} = 10$, $\lambda_{cc} = 10$, $\lambda_{sty} = 1$ and $\lambda_{per} = 1$. Our programs were written in the Pytorch deep learning framework [23]. All experiments were run with batch size 8 on a Ubuntu 20.04 with NVIDIA 3090 GPU. We used the Adam [18] optimizer

Pose	Accuracy (%)					
	FF	FS	SS	FP	SP	PP
ArcFace[7]	82.1	49.2	59.3	31.2	38.7	28.1
MagFace[21]	80.7	48.1	58.5	28.6	35.5	26.4
VGGFace2[4]	86.7	69.8	79.8	58.6	63.3	60.7
AdaFace[17]	90.6	74.2	77.3	43.8	49.5	43.2
PAE	**92.2**	**80.1**	**84.7**	**67.2**	**71.4**	**73.8**

Table 1. Face verification performance of face encoders on six pose test sets.

Pose	Accuracy (%)					
	FF	FS	SS	FP	SP	PP
ArcFace[7]	83.1	58.7	63.4	41.6	49.5	43.2
MagFace[21]	81.6	59.8	68.4	43.8	55.4	55.5
VGGFace2[4]	89.7	**84.9**	83.6	65.6	69.2	63.9
AdaFace[17]	91.3	79.5	83.2	66.8	68.5	70.2
PAE	**92.2**	80.1	**84.7**	**67.2**	**71.4**	**73.8**

Table 2. Face verification performance of face encoders fine tuned on PAE training sets.

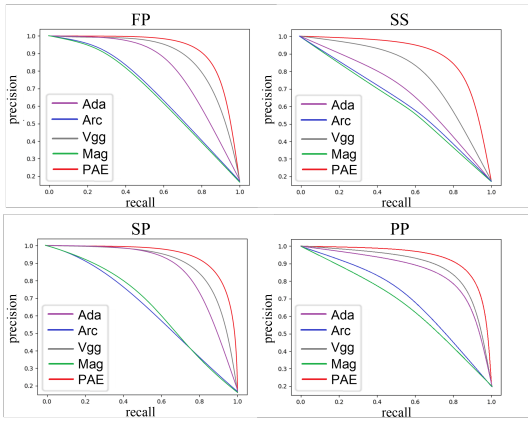

Figure 2. Precision and recall of PAE, Magface (Mag), Arcface (Arc), Adaface (Ada), and VGGFace2 (Vgg) encoders.

Metrics	FID↓	CSIM↑	ARD↓	LPIPS↓
	Recomposed Shape Encoder/Decoder			
w/o $\mathcal{L}_{cc}$	23.9	0.38	2.84	0.28
CSG	**18.1**	**0.46**	**2.24**	**0.21**
	Attention Mechanism			
StarGAN2[5]	28.4	0.37	2.41	0.26
AEG	**18.1**	**0.46**	**2.24**	**0.21**

Table 3. Self-reenactment performance with and without cycle-consistent shape loss L_{cc}, and with and without attention mechanism in the generator.

with $\beta 1 = 0.01$, $\beta 2 = 0.99$ and learning rate $1e^{-4}$.

4.2. Ablation Study

In the first part of our ablation study, we conducted a comparison experiment to verify the performance of PAEs in face verification. Several face encoders released in recent years were selected, including the Magface [21], Arcface [7], Adaface [17], and VGGFace2 [4]. Table 1 shows the comparison with the off-the-shelf pretrained versions of the selected encoders. The PAEs outperform all in 6 subsets with clear margins and the margin increases with pose difference. Table 2 shows the performance that we fine-tuned those encoders on our training set. Only VGGFace2 outperforms PAEs on the FS (frontal-side) subset, indicating that pose-adapted training may improve verification performance within the specified pose scope. However, the PAEs maintain satisfactory performance even for subsets of small pose differences. Figure 2 shows the Precision-Recall curves of PAEs and off-the-shelf encoders on 4 subsets of larger pose differences, offering a more comprehensive comparison. This study verifies that PAEs are better encoders, especially for faces of large poses.

In the second part of ablation study, we compared the performance with and without the cycle-consistent shape loss L_{cc}, and with and without the attention mechanism in the generator. The model without L_{cc} uses the DECA encoder and decoder as the encoder U_e and decoder U_d in the CSG. As for the generator without attention mechanism, we selected StarGAN2 [5] and trained it the same way as we did for the AEG. Table 3 shows the quantitative comparison for self-reenactment on the MPIE-LP testing set, and Figures 3 and 4 illustrate the differences in visual quality. It is verified that both cycle-consistent shape and attention-embedded generator improve the performance.

Figure 3. Qualitative comparison of with and without cycle-consistent shape loss (quantitative comparison in Table 3).

Figure 4. Qualitative comparison of with and without attention mechanism in the generator (quantitative comparison in Table 3).

	side-vs-side (ss)				side-vs-profile (sp)				frontal-vs-profile (fp)				profile-vs-profile (pp)			
Method(N)	FID↓	CSIM↑	ARD↓	LPIPS↓	FID↓	CSIM↑	ARD↓	LPIPS↓	FID↓	CSIM↑	ARD↓	LPIPS↓	FID↓	CSIM↑	ARD↓	LPIPS↓
MPIE-LP																
MagFace[21]	21.57	0.431	1.88	0.252	22.56	0.433	1.92	0.259	23.39	0.391	2.09	0.264	27.21	0.354	2.23	0.275
Arcface[7]	20.11	0.457	1.63	0.234	21.27	0.451	1.74	0.243	22.97	0.408	1.98	0.248	24.61	0.377	2.13	0.261
VGGFace2[4]	19.58	0.466	1.38	0.227	20.85	0.469	1.48	0.236	21.56	0.411	1.71	0.241	22.95	0.381	1.89	0.248
PAE	**16.96**	**0.491**	**1.14**	**0.192**	**17.48**	**0.483**	**1.19**	**0.198**	**18.84**	**0.427**	**1.38**	**0.214**	**21.60**	**0.401**	**1.47**	**0.227**
VoxCeleb2-LP																
MagFace[21]	41.02	0.483	3.26	-	42.46	0.471	3.34	-	43.95	0.459	3.58	-	45.62	0.442	3.77	-
Arcface[7]	39.51	0.498	3.13	-	41.53	0.487	3.22	-	42.76	0.473	3.42	-	44.39	0.465	3.53	-
VGGFace2[4]	37.61	0.526	2.88	-	40.92	0.513	2.99	-	41.76	0.497	3.18	-	42.93	0.480	3.35	-
PAE	**36.45**	**0.541**	**2.67**	-	**39.54**	**0.535**	**2.81**	-	**39.86**	**0.510**	**3.02**	-	**40.63**	**0.498**	**3.22**	-

Table 4. Cross-reenactment performance using different face encoders for computing identity loss $\mathcal{L}_{id}$

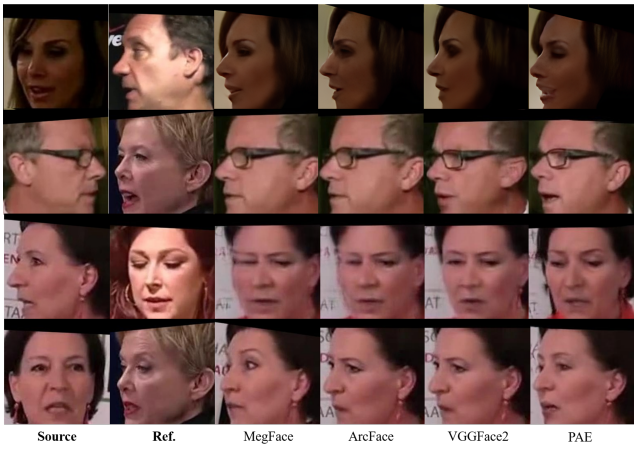

Figure 5. Samples from cross-reenactment on VoxCeleb2-LP with identity loss made by different face encoders, performance in Table 4.

Metrics	FID↓	CSIM↑	ARD↓	LPIPS↓
MPIE-LP				
Bi-layer[32]	106.8	0.12 / 0.03 / 0.22	3.22	0.54
FOMM[24]	62.7	0.06 / 0.08 / 0.08	10.86	0.37
DG[14]	23.4	0.23 / 0.26 /0.33	2.15	0.24
HyperReenact[2]	83.9	0.22 / 0.26 / 0.35	6.71	0.47
PASL	**17.5**	0.25 / 0.21 / **0.47**	**1.21**	**0.18**
VoxCeleb1				
Bi-layer[32]	52.8	0.64	2.19	0.23
FOMM[24]	35.6	0.65	4.63	0.27
HeadGAN[9]	58.0	0.68	1.35	0.24
Face2Face$^{\rho}$[29]	-	-	-	0.21
DG[14]	41.1	0.65	1.53	0.18
HyperReenact[2]	38.2	0.68	1.99	0.32
PASL	**32.4**	**0.71**	**1.32**	**0.15**
VoxCeleb2-LP				
Bi-layer[32]	108.3	0.2 / 0.11 / 0.28	3.18	0.53
FOMM[24]	49.1	0.26 / 0.22 / 0.33	3.67	0.35
DG[14]	36.1	0.28 / 0.29 / 0.46	2.94	0.24
HyperReenact[2]	78.5	0.39 / 0.38 / 0.52	3.03	0.23
PASL	**32.8**	0.42 / 0.39 / **0.58**	**2.71**	**0.21**

Table 5. Self-reenactment performance on MPIE-LP, VoxCeleb1 and VoxCeleb2-LP (CSIM by ArcFace/VggFace2/PAE)

In the third part of ablation study, we compared the reenacted faces synthesized by the proposed approach but using different face encoders for source identity preservation. Figure 5 shows the cross-reenacted faces associated with the quantitative performance comparison in Table 4. The PAEs consistently outperform other encoders in all metrics and exhibit superior performance in faithfully capturing facial features.The PAEs demonstrate enhanced fidelity, resulting in a more accurate reconstruction of these facial components. The difficulty level of identity preservation can be seen from the CSIM of each pose set. The MPIE-LP-pp with CSIM 0.401 appears the most difficult as many pairs are 180° apart, as required in the controlled settings. The VoxCeleb2-LP-ss with CSIM 0.541 seems the easiest as many pairs collected in the wild have smaller pose differences. Other metrics also reveal some interesting characteristics of the pose sets.

4.3. Comparison with State of the Art

Table 5 and 6 show the performance for self- and cross-reenactment compared with state-of-the-art approaches. As CSIM is a common metric for measuring identity preservation, we followed the common practice of using ArcFace as the feature extractor for experiments on VoxCeleb1. For experiments on MPIE-LP and VoxCeleb2-LP, we computed CIM by using ArcFace, VggFace2, and PAEs. Based on the performance comparison study in Sec. 4.2, the PAEs are more reliable than ArcFace and VggFace2 in measuring the identity similarity between faces of large poses.

Several selected methods do not offer code or models for testing, except for the FOMM [24], Bi-layer [32], DG [14] and HyperReenact[2]. In that case, we duplicated the reported performance and image samples directly from their papers. Note that LPIPS can only be computed for self-reenactment or when the ground-truth target face is available; therefore, it is not available for VoxCeleb1 and

Metrics	FID↓	CSIM↑	ARD↓	LPIPS↓
MPIE-LP				
Bi-layer[32]	106.3	0.13 / 0.03 / 0.21	3.22	0.54
FOMM[24]	68.9	0.19 / 0.21 / 0.25	10.91	0.39
DG[14]	24.2	0.21 / 0.23 / 0.28	2.35	0.24
HyperReenact[2]	85.2	0.2 / 0.26 / 0.31	6.65	0.48
PASL	**18.1**	0.27 / 0.24 / **0.46**	**2.24**	**0.21**
VoxCeleb1				
Bi-layer[32]	52.9	0.56	2.18	-
FOMM[24]	52.9	0.53	10.9	-
HeadGAN[9]	48.1	0.62	2.35	-
Face2Face$^{\rho}$[29]	44.5	0.61	3.68	-
DG[14]	45.1	0.57	2.11	-
HyperReenact[2]	40.3	0.61	2.2	-
PASL	**38.4**	**0.68**	**2.05**	-
VoxCeleb2-LP				
Bi-layer[32]	110.8	0.16 / 0.01 / 0.24	3.22	-
FOMM[24]	62.5	0.25 / 0.29 / 0.39	5.82	-
DG[14]	44.2	0.25 / 0.26 / 0.39	3.21	-
HyperReenact[2]	79.2	0.33 / 0.34 / 0.45	4.63	-
PASL	**37.9**	0.35 / 0.39 / **0.52**	**2.93**	-

Table 6. Cross-reenactment performance on MPIE-LP, VoxCeleb1 and VoxCeleb2-LP (CSIM by ArcFace/VggFace2/PAE)

VoxCeleb2-LP. The Face2Face$^{\rho}$[29] uses LPIPS and other metrics for evaluating self-reenactment, and FID, CSIM and ARD for evaluating cross-reenactment. For the approaches with code and models available, we trained and tested their models the same way as we performed for the PASL. The performance comparison is shown in Table 5 for self-reenactment and Table 6 for cross-reenactment. The PASL outperforms the selected SOTA approaches in all metrics on the three benchmark datasets. For the general scope of pose and expression variation, the PASL demonstrates the performance on VoxCeleb1. For handling large pose scenarios, the PASL shows the performance on the MPIE-LP and VoxCeleb2-LP datasets.

Figures 6, 7 and 8 show the reenacted face samples for the comparison with state-of-the-art methods reported in Table 5 and 6. The Bi-layer [32] can transform the pose and expression for minor pose variation, but cannot preserve the source identity, as shown in Figure 7. The identity preservation is worsened when handling large pose, as shown in Figure 6, 8 and in Table 5, 6. The FOMM [24] cannot handle large pose, either, as shown in Figure 6, 7 and 8. The DG [14] demonstrates good performance in handling large poses, although it exhibits a slight deficiency in identity preservation, as evidenced by the quantitative results presented in Tables 5 and 6, as well as qualitative analysis illustrated in the figure. The HyperReenact [2] emerges as the second-best method in terms of identity preservation, but displays comparatively lower efficacy across other evaluation metrics.

5. Conclusion

Different from most previous approaches that deal with common pose variation scope, we propose the Pose-Adapted Shape Learning for handling large-pose face reenactment. We develop a set of Pose Adapted Encoders (PAEs) to better preserve the source identity across large pose and serve as a strong pose-robust face encoder. We design the Cycle-consistent Shape Generator (CSG) and the Attention-Embedded Generator (AEG) to better generate the desired target face. To better evaluate the performance, we propose a scheme to make MPIE-LP and VoxCeleb2-LP datasets with sufficient large-pose data. The proposed approach is successfully verified on MPIE-LP, VoxCeleb1 and VoxCeleb2-LP datasets.

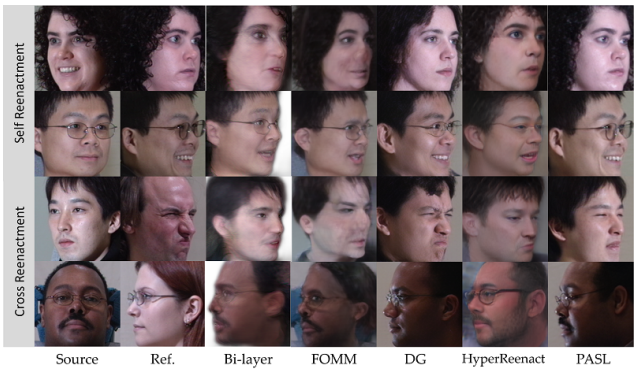

Figure 6. Samples from comparison with state of the art on MPIE-LP, with performance in Table 5 and 6.

Figure 7. Samples from comparison with state of the art on VoxCeleb1, with performance in Table 5 and 6.

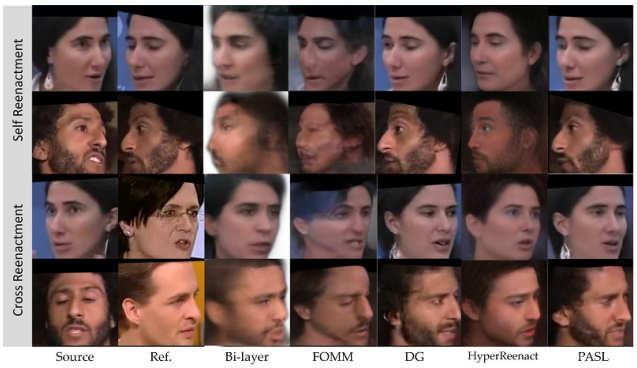

Figure 8. Samples from comparison with state of the art on Vox2-LP, with performance in Table 5 and 6.

References

[1] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 1999. 3

[2] Stella Bounareli, Christos Tzelepis, Vasileios Argyriou, Ioannis Patras, and Georgios Tzimiropoulos. Hyperreenact: One-shot reenactment via jointly learning to refine and retarget faces. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 1, 2, 7, 8

[3] Egor Burkov, Igor Pasechnik, Artur Grigorev, and Victor Lempitsky. Neural head reenactment with latent pose descriptors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13786–13795, 2020. 1

[4] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018. 1, 5, 6, 7

[5] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8188–8197, 2020. 6

[6] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*, 2018. 5

[7] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 1, 6, 7

[8] Jiankang Deng, Jia Guo, Debing Zhang, Yafeng Deng, Xiangju Lu, and Song Shi. Lightweight face recognition challenge. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 5

[9] Michail Christos Doukas, Stefanos Zafeiriou, and Viktoriia Sharmanska. Headgan: One-shot neural head synthesis and editing. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1, 2, 3, 7, 8

[10] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. 2021. 3

[11] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and Vision Computing*, 2010. 5

[12] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3d dense face alignment. In *European Conference on Computer Vision*, pages 152–168. Springer, 2020. 5

[13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*, 2017. 5

[14] Gee-Sern Hsu, Chun-Hung Tsai, and Hung-Yi Wu. Dual-generator face reenactment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 642–650, 2022. 1, 2, 7, 8

[15] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 4

[16] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 2, 4

[17] Minchul Kim, Anil K Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 6

[18] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[19] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), 2017. 3

[20] Jin Liu, Peng Chen, Tao Liang, Zhaoxing Li, Cai Yu, Shuqiao Zou, Jiao Dai, and Jizhong Han. Li-net: Large-pose identity-preserving face reenactment network. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021. 1

[21] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14225–14234, 2021. 6, 7

[22] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*, 2017. 5

[23] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 5

[24] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *NIPS*, 2019. 1, 2, 7, 8

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[26] Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Jan Kautz, and Bryan Catanzaro. Few-shot video-to-video synthesis. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019. 1

[27] Cameron Whitelam, Emma Taborsky, Austin Blanton, Brianna Maze, Jocelyn C. Adams, Tim Miller, Nathan D. Kalka, Anil K. Jain, James A. Duncan, Kristen E Allen, Jordan Cheney, and Patrick Grother. Iarpa janus benchmark-b face dataset. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 592–600, 2017. 5

[28] Chao Xu, Jiangning Zhang, Yue Han, Guanzhong Tian, Xianfang Zeng, Ying Tai, Yabiao Wang, Chengjie Wang, and Yong Liu. Designing one unified framework for high-fidelity face reenactment and swapping. In *European Conference on Computer Vision*, pages 54–71. Springer, 2022. 1

[29] Kewei Yang, Kang Chen, Daoliang Guo, Song-Hai Zhang, Yuan-Chen Guo, and Weidong Zhang. Face2face ρ: Real-time high-resolution one-shot face reenactment. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIII*, pages 55–71. Springer, 2022. 1, 2, 3, 7, 8

[30] Guangming Yao, Yi Yian, Tianjia Shao, and Kun Zhou. Mesh guided one-shot face reenactment using graph convolutional networks. *arXiv preprint arXiv:2008.07783*, 2020. 2

[31] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *ICCV*, 2019. 1, 2

[32] Egor Zakharov, Aleksei Ivakhnenko, Aliaksandra Shysheya, and Victor Lempitsky. Fast bi-layer neural synthesis of one-shot realistic head avatars. In *European Conference of Computer vision (ECCV)*, 2020. 2, 7, 8

[33] Jiangning Zhang, Xianfang Zeng, Mengmeng Wang, Yusu Pan, Liang Liu, Yong Liu, Yu Ding, and Changjie Fan. Freenet: Multi-identity face reenactment. In *CVPR*, 2020. 1, 2

[34] Richard Zhang, Alan C. Bovik, and Lei Zhang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018. 5

[35] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016. 2