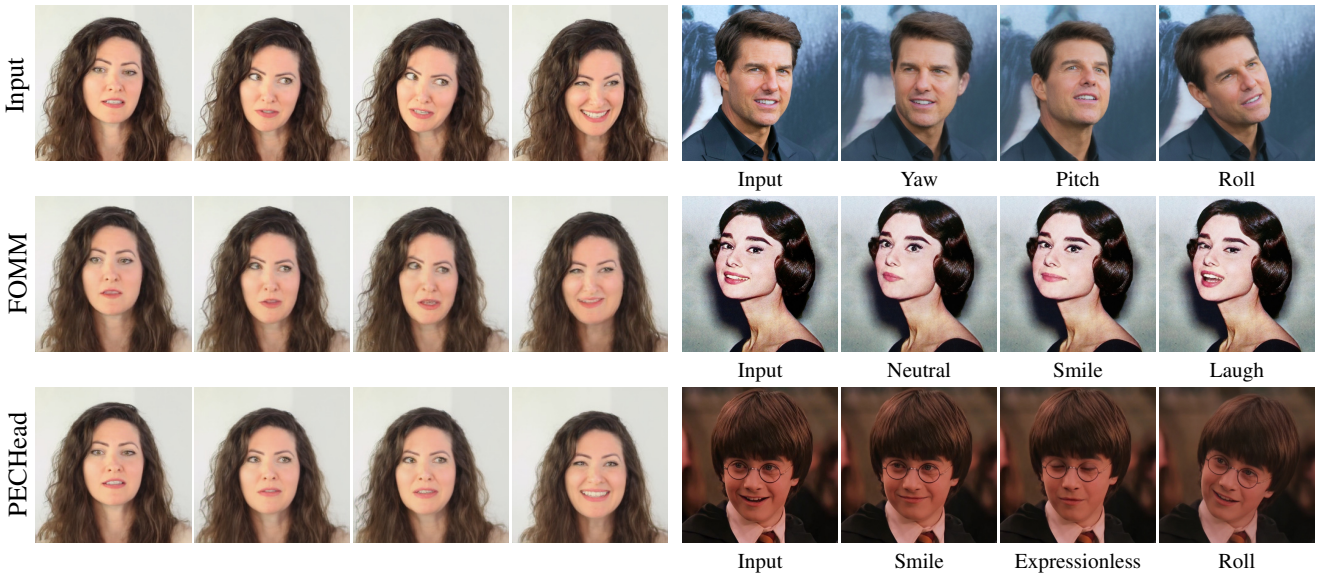
High-Fidelity and Freely Controllable Talking Head Video Generation

Yue Gao Yuan Zhou Jinglu Wang Xiao Li Xiang Ming Yan Lu
Microsoft Research
{yuegao, zhouyuan, jinglu.wang, li.xiao, xiangming, yanlu}@microsoft.com



(a). Frontalization (b). Free pose and expression editing

Figure 1. Presented herein are representative results showcasing the effectiveness of our proposed method in the tasks of frontalization, as well as free pose and expression editing. (a) FOMM [46] often produces face distortion issues, while our model **PECHead** generates high-fidelity results. (b) Our proposed framework empowers the generation of talking head videos that offer free control over the head pose and expression. More results can be found in the supplementary materials.

Abstract

Talking head generation is to generate video based on a given source identity and target motion. However, current methods face several challenges that limit the quality and controllability of the generated videos. First, the generated face often has unexpected deformation and severe distortions. Second, the driving image does not explicitly disentangle movement-relevant information, such as poses and expressions, which restricts the manipulation of different attributes during generation. Third, the generated videos tend to have flickering artifacts due to the inconsistency of the extracted landmarks between adjacent frames. In this paper, we propose a novel model that produces high-fidelity talking head videos with free control over head pose and expression. Our method leverages both self-supervised learned landmarks and 3D face model-based landmarks to model the motion. We also introduce a novel motion-aware multi-scale feature alignment module to effectively transfer the motion without face distortion. Furthermore, we enhance the smoothness of the synthesized talking head videos with a feature context adaptation and propagation module. We evaluate our model on challenging datasets and demonstrate its state-of-the-art performance. More information is available at https://yuegao.me/PECHead.

1. Introduction

Talking head video generation is a process of synthesizing a talking head video with a given source identity and target motion. This process is also called face reenactment when using a driving head to define the relative movement to the source identity [4]. This generation technique can

be used in various applications, including video conferencing [51], movie effects [39], and entertainment [6]. Due to the rapid development of deep learning [20] and generative adversarial networks (GAN) [17, 27], impressive works have been conducted on talking head generation [22, 51], face reenactment [2, 24, 40, 49, 57, 59, 61, 65, 66], and image animation [46, 47, 52, 67]. These works focus on animating objects beyond the face and head [45, 46].

Early works on talking head generation require multiple source and driving images to generate one result [4, 13]. Recent works focus on one-shot generation [51, 61, 66], *i.e.*, using only one source frame to generate the target by transferring the pose information of one driving frame. Currently, the mainstream works [45–47] follow a self-supervised learning pipeline. They mainly utilized the self-supervised learned landmarks to model the movement of the identity between the source and driving images. The learned landmarks pairs are first detected from both source and driving images, and then the dense flow field is estimated from the two sets of learned landmarks to transform the source features and guide the reconstruction of the driving image. To further improve the performance, recent approaches propose to utilize additional information, such as 3D learned landmarks [51] and depth map [22], or enhance the model structure, for example, adopting the flexible thin-plate spline transformation [14, 67], and representing the motion as a combination of latent vectors [52].

However, there are still many challenges with these methods. First, the generated face often has unexpected deformation and severe distortions. The learned landmarks-based approaches [46, 51, 67], such as FOMM [46], which only utilizes the 2D learned landmarks without face shape constraints, produces frontalization results with apparent face distortions (see Fig. 1a). The predefined landmarks-based methods [13, 24, 59, 63] model the movement between the source and driving images only based on the predefined facial landmarks, leading to the non-facial parts of the head (such as the hair and neck) are not well handled. Second, all the movement information needs to be obtained via one single driving image. It is rare and difficult to decouple and manipulate these movement-relevant information, including poses and expressions, when generating the new image. Third, in order to achieve smooth and natural movements in generated videos, prior methods [46, 47, 67] typically incorporate techniques to smoothen the extracted landmarks learned between adjacent frames. However, the sensitivity and inconsistency of the extracted landmarks pose a challenge in achieving smoothness, resulting in generated videos that are prone to flickering.

To address the above challenges, we propose the **P**ose and **E**xpression **C**ontrollable **Head** model (**PECHead**), which can generate high-fidelity video face reenactment results and enable talking head video generation with full control over head pose and expression. The proposed method first incorporates the learned landmarks and the predefined face landmarks to model the overall head movement and the detailed facial expression changing in parallel. We utilize the single image-based face reconstruction model [12] to obtain the face landmarks and project them into 2D image space. This approach constrains the face to a physically reasonable shape, thereby reducing distortion during motion transfer, as demonstrated in the last row of Fig. 1a. In this work, we introduce the use of learned sparse landmarks for global motion and predefined dense landmarks for local motion, with the Motion-Aware Multi-Scale Feature Alignment (MMFA) module serving to align these two groups of features. Then we use different coefficients as input conditions to control the estimation of both predefined and learned landmarks, so that we can realize the head pose and expression manipulation (Fig. 1b). Moreover, inspired by the recent low-level video processing works [8, 33], we propose the Context Adaptation and Propagation (CAP) module to further improve the smoothness of the generated video. Our proposed method is evaluated on multiple talking head datasets, and experimental results indicate that it achieves state-of-the-art performance, generating high-fidelity face reenactment results and talking head videos with the ability to control the desired head pose and facial expression.

Our contributions can be summarized as follows:

- We propose a novel method, **PECHead**, that generates high-fidelity face reenactment results and talking head videos. Our approach leverages head movements to control the estimation of learned and predefined landmarks, enabling free control over the head pose and expression in talking head generation.
- We incorporate the learned and predefined face landmarks for global and local motion estimation with the proposed Motion-Aware Multi-Scale Feature Alignment module, which substantially enhances the quality of synthesized images.
- We introduce a video-based pipeline with the Context Adaptation and Propagation module to further improve the smoothness and naturalness of the generated videos.
- Extensive qualitative and quantitative results across several datasets demonstrate the superiority of the proposed framework for high-fidelity video face reenactment and freely controllable talking head generation.

2. Related Works

Image Animation. Image animation is to transfer motion information from one domain to another. Traditional approaches often rely on strong priors such as face

meshes [5], human keypoints [6], or action units [15, 42]. In recent years, there has been a growing interest in self-supervised methods that only require videos. Monkey-Net [45] uses sparse learned landmarks to estimate optical flow for animating arbitrary objects. FOMM [46] extends Monkey-Net [45] by incorporating local affine transformation. MRAA [47] proposes region-based motion representations, while LIA [52] represents motion as a set of learned motion directions. These methods eliminate the requirement of explicit structure representations. Zhao et al. [67] leverage thin-plate spline transformation [14] for motion estimation, which is more flexible than traditional approaches.

Talking Head Generation. In recent years, a significant amount of research has been dedicated to the task of face reenactment or talking head generation. This is largely owing to the development of large-scale face data [25, 60], 3D morphable face model (3DMM) and 3D mesh [16, 34], neural radiance fields (NeRF) [38], and face landmark detectors [36, 56]. We will discuss these methods in details as follows.

3D face model-based methods. For example, Face2Face [49] employs a deformation transfer approach to track facial expressions of both source and driving videos, followed by re-rendering of the synthesized faces. Ma et al. [37] reconstructs an individual-specific face model with high-resolution facial geometry and appearance.

Direct synthesis-based models synthesize target faces by decoding latent appearance and motion representations. For instance, Zakharov et al. [63] introduce the first direct synthesis method for face reenactment. LPD [4] utilizes head pose augmentation, while DAE-GAN [64] disentangles identity and pose representations using the deforming autoencoder [44].

3D mesh-based methods utilize neural head models to synthesize realistic head avatars from videos. Grassal et al. [18] proposes a neural head model that provides a disentangled shape and appearance representation. ROME [32] uses a single image to estimate a person-specific head mesh and texture to synthesize neural head avatars.

NeRF-based methods use NeRF as a novel 3D proxy to represent the head geometry and appearance. For example, AD-NeRF [19] proposes using NeRF for audio-driving talking head video generation. Head-NeRF [23] uses NeRF to control the pose and various semantic attributes of the generated images. However, NeRF-based methods are not effective in generalizing across identities, and the models are relatively complex compared to the sparse landmark-based models.

Warping-based methods use learned landmarks/regions pairs to estimate motion fields [45–47], performs warping on the feature maps, and generates images. X2Face [55] uses latent vectors that are learned to be predictable of warping fields. Bi-layer [62] employs a bi-layer representation via summing two components, a coarse image directly predicted by a rendering network and a warped texture image. PIRender [43] controls the face motions directly with 3DMMs. OSFV [51] extracts 3D learned landmarks with 3D convolution nets for better modeling the head, and utilizes the rotation matrix to transform the overall viewpoint but not for free control of all head poses and expressions. HeadGAN [13] and Face2Face$^{\rho}$ [59] are two such methods that estimate motion information from input images using 3D meshes and landmarks, respectively. DaGAN [22] presents a self-supervised depth estimator and cross-modal attention to generate motion fields. While these methods have shown promising results, there is still room for improvement in terms of flexibility, physics-consistency, and video smoothness. To address these limitations, this paper proposes a novel approach that leverages both self-supervised learned landmarks and predefined landmarks for motion transfer while also improving the smoothness of the resulting videos.

3. Method

3.1. Overview

This section describes the proposed method **PECHead**, (see Fig. 2 for illustration), which mainly contains four parts: Generator G, Face Shape Reconstructor R, Head Pose-Aware Keypoint Estimator E, and Multi-Scale Discriminator D. Our framework follows the basic pipeline proposed by Siarohin et al. [45–47]. We first extract the face coefficients and predefined landmarks through R, and then estimate the learned landmarks through E with the head pose and expression as conditions. The generator G takes the predefined and learned landmarks pairs to estimate the dense flow and generates the results. During training, our model takes two sequences with the same subject and number of adjacent frames. We denote the frames in these two sequences as source frame $x_t^s \in \mathbb{R}^{3\times H\times W}$ and driving frame $x_t^d \in \mathbb{R}^{3\times H\times W}$, where $1\leq t\leq T$, T is the sequence length, and $H\times W$ is the spatial size. The model is learned to reconstruct the driving frame x_t^d, and the synthesized frame is denoted as $\hat{x}_t^d$. In the following sections, the frame index $t-1$ or t are omitted for brevity, except when necessary. At test time, we can extract the coefficients from the driving frames or modify the coefficients of the source frames, to get different landmarks pairs. This allows us to transfer the motion from the driving frames or edit the source frames.

Generator. Generator G mainly contains the encoder, bottleneck module, and decoder. The encoder extracts the raw feature f^r of the current source frame. The bottleneck module aligns the raw feature f^r to the driving frame and adapts it to the context information. The details will be discussed in Sec. 3.2 and Sec. 3.3. The decoder generates the recon-

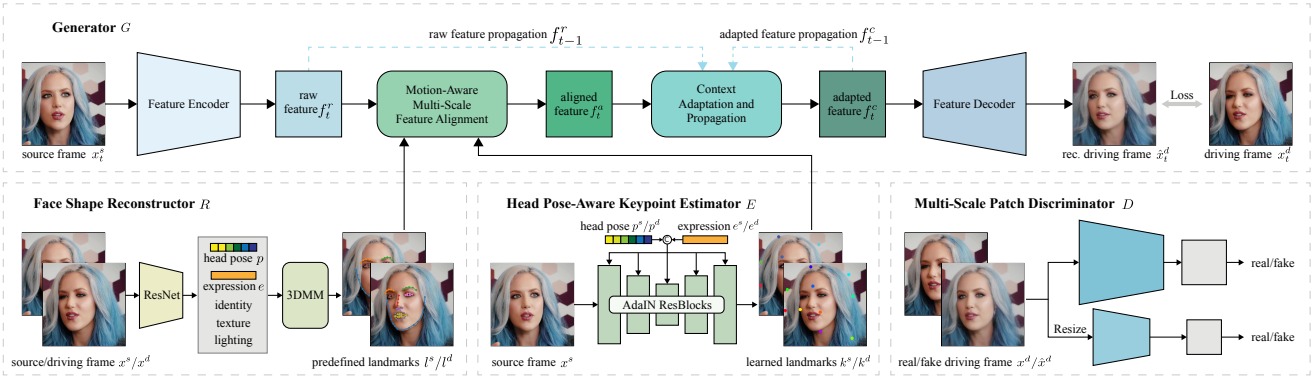


Figure 2. The overview of our method, which contains four parts: a) the Generator G; b) the Face Shape Reconstructor R; c) the Head Pose-Aware Keypoint Estimator E; and d) the Multi-Scale Discriminator D. The light blue dash arrows stand for the feature propagation.

structed frames $\hat{x}^d$ based on the adapted feature f^c.

Face Shape Reconstructor. As mentioned in Sec. 1, existing self-supervised learned landmarks-based models can not freely control the head poses and facial expressions. To solve this problem, we incorporate the predefined face landmarks [3] with the learned landmarks. Specifically, a single image-based Face Shape Reconstructor R is adopted to extract the landmarks l^s, l^d, head pose p^s, p^d and expression e^s, e^d from the source and driving frames, respectively,

$$l^z, p^z, e^z = R(x^z); z \in \{s, d\}. \tag{1}$$

Our Reconstructor R is derived from the state-of-the-art face reconstruction model [12], which uses a ResNet [20] to obtain a set of coefficients and fits a Basel Face Model (BFM) [1, 16]. We can further compute the 3D landmarks and project them to the 2D space. When calculating the driving landmarks l^d, the other coefficients (*i.e.*, identity, texture, and lighting) extracted from the source frame are used to preserve the identity.

Head Pose-Aware Keypoint Estimator. Existing keypoint-based models [22, 46, 51, 67] directly feed the source and driving frames to keypoint detector to obtain the learned landmarks pairs. Instead, we use the source frame x^s conditioned with corresponding head pose p and expression e to obtain source learned landmarks k^s and driving learned landmarks k^d, facilitating manually specified head pose and expression editing. This process is formulated as,

$$k^z = E(x^s, p^z, e^z); z \in \{s, d\}, k \in \mathbb{R}^{K\times 2}, \tag{2}$$

where K is the number of learned landmarks, we set K=10. The E is trained to detect the learned landmarks based on the appearance provided from the source frame x^s obeying the head pose p and expression e. The head pose p and expression e are injected with AdaIN [26] module. As the coefficients of the face model are decoupled by definition, the different learned and predefined landmarks k', l' can be obtained by modifying the head pose p or expression e, and manipulated frame $\hat{x}'$ can be generated correspondingly.

Multi-Scale Discriminator. Following the existing generative models [22, 27, 29, 30, 41, 46, 51], we utilize a Multi-scale Patch Discriminator D to encourage the generator G produce more realistic frames.

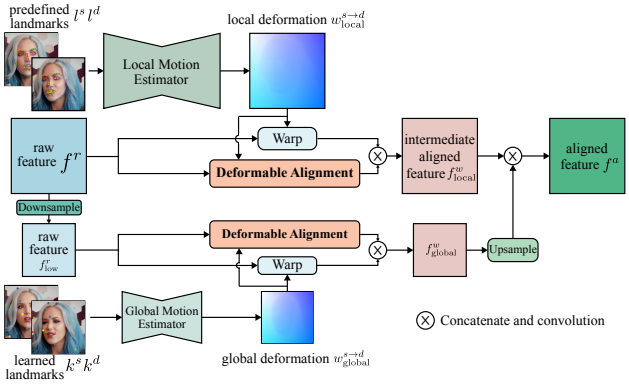


Figure 3. Motion-Aware Multi-Scale Feature Alignment module.

3.2. Motion-Aware Multi-Scale Feature Alignment

Although both the learned and predefined landmarks are represented in 2D image space, our experimental results demonstrate that directly merging these points in series does not give satisfactory results. As the learned ones are freely learned by the model, while the predefined ones are artificially defined, their physical meanings of them are different. Therefore, we propose the Motion-Aware Multi-Scale Feature Alignment (MMFA) module to incorporate the learned sparse and predefined dense landmarks.

As shown in Fig. 3, the MMFA correlates the predefined landmarks l^s, l^d and the learned landmarks k^s, k^d for deforming the raw feature extracted by the encoder. The sparse learned landmarks detected from the whole frame can provide more global motion information, *i.e.*, the overall head movement. And the landmarks can be used for modeling more details of the motion, such as expression changing, as they are estimated from the face shape model. We use two motion estimators Φ to estimate the global and local motion information based on the learned and prede-

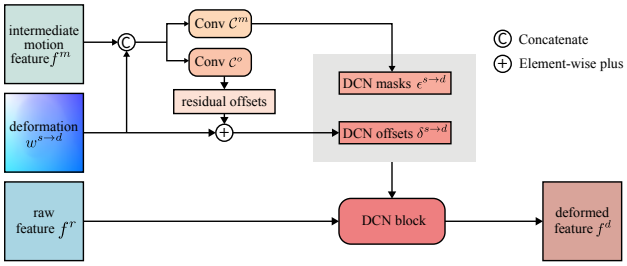


Figure 4. Deformable Alignment module.

fined landmarks, respectively. Following FOMM [46], the Gaussian heatmap-based motion estimators are employed. The global motion is applied on the downsampled raw feature $f^r_{\text{low}} \in \mathbb{R}^{C^1\times h/2\times w/2}$, which contains higher level information. And the local motion is applied on the raw feature $f^r \in \mathbb{R}^{C^2\times h\times w}$ for more details, where C^1 and C^2 are the channel dimension, $h\times w$ is the spatial size of the feature map. The motion information contains two parts, deformation $w^{s\to d} \in \mathbb{R}^{2\times h\times w}$, and occlusion $o^{s\to d}\in\mathbb{R}^{h\times w}$ (omitted in Fig. 3 for brevity). With the deformation and occlusion map, the raw feature can be warped as $f^w = \mathcal{W}(f^r, w^{s\to d})\cdot o^{s\to d} + (1-o^{s\to d})\cdot f^r$, where $\mathcal{W}$ is the warping operation.

According to Chan et al. [7, 8], the deformable alignment [31] demonstrates significant improvements over the flow-based alignment. As shown in Fig. 4, the deformable alignment takes the last feature map f^m in the motion estimator and deformation map $w^{s\to d}$ to compute offsets $\delta^{s\to d}$ and masks $\epsilon^{s\to d}$ of the deformable convolution (DCN) [11], and then a DCN is applied,

$$\begin{aligned}\delta^{s\to d} &= w^{s\to d} + \mathcal{C}^o(c(f^m, w^{s\to d})),\\ \epsilon^{s\to d} &= \sigma(\mathcal{C}^m(c(f^m, w^{s\to d})),\\ f^d &= \mathcal{D}(f^r; \delta^{s\to d}, \epsilon^{s\to d}),\end{aligned} \tag{3}$$

where c is the concatenation operation, $\mathcal{C}^{\{o,m\}}$ denotes convolution blocks, σ is the sigmoid and $\mathcal{D}$ stands for the DCN. We apply the alignment on both global and local levels.

3.3. Context Adaptation and Propagation

We introduce the Context Adaptation and Propagation (CAP) module to make the model produce smooth videos. The illustration of CAP is shown in Fig. 5. First, the raw feature f^r_{t-1} of the previous source frame is sequentially warped with the frame flow $w_{t-1\to t}$ and the current local deformation $w^{s\to d}_{\text{local}}$ (Two-Step Warping). The frame flow is computed on x^s_{t-1} and x^s_t using image-based flow estimator [48]. Second, the warped previous raw feature $\tilde{f}^r_{t-1}$, hidden feature h_{t-1}, adapted feature f^c_{t-1} and the current aligned feature f^a_t are concatenated. We further feed the concatenated feature to the Context Adaptation module, which is composed of several convolution blocks, to get the feature in the same spatial and channel size with f^a_t. Then, the feature is further refined with the Feature Refinement

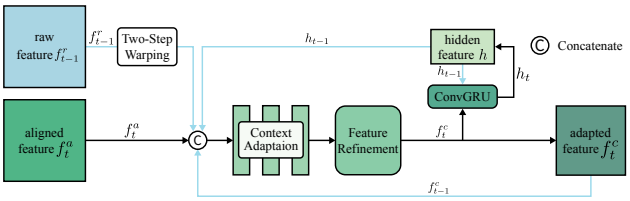


Figure 5. Context Adaptation and Propagation module.

module. After that, we get the adapted feature f^c_t for the current frame, and the hidden feature is updated with the adapted feature f^c_t using a ConvGRU block [9]. The proposed CAP module can be formed as,

$$\begin{aligned}h_t &= \text{ConvGRU}(f^c_{t-1}, h_{t-1}),\\ \tilde{f}^r_{t-1} &= \mathcal{W}(\mathcal{W}(f^r_{t-1}, w_{t-1\to t}), w^{s\to d}_{\text{local}}),\\ f^c_t &= \text{FR}(\text{CA}(c(f^a_t, \tilde{f}^r_{t-1}, f^c_{t-1}, h_{t-1}))),\end{aligned} \tag{4}$$

where CA, FR represent the Context Adaptation and Feature Refinement submodules, respectively. The hidden feature h and the previous adapted feature f^c are initialized to zeros for the first frame.

3.4. Objective Function

Following existing works [46, 47], our model is trained with the reconstruction task. We briefly discuss these losses and leave the details in the supplementary material.

Pixel-wise Loss $\mathcal{L}_p$. The pixel-wise loss is employed to ensure the synthesis frames are similar to the driving frames.

Perceptual Loss $\mathcal{L}_v$. Similar to existing methods [46, 47, 51, 67], we use a pre-trained VGG [28] to guarantee consistency of high level characteristics between driving frame x^d and generated frame $\hat{x}^d$.

Learned Landmarks Loss $\mathcal{L}_k$. The learned landmarks loss [51] is used to encourage the estimated learned landmarks k to spread out across the whole frame.

Equivariance Loss $\mathcal{L}_e$. The equivariance loss [46, 47] is applied to constrain the consistency of Head Pose-Aware Keypoint Estimator E.

Warping Loss $\mathcal{L}_w$. This loss is designed to ensure the motion estimators to predict the deformations reasonably, making the warped source frame closer to the driving frame.

GAN Loss $\mathcal{L}_G, \mathcal{L}_D$. We adopt the hinge loss as our adversarial loss [35], and two different scale patch discriminator is used for better performance [27].

Full Objective Function. The total loss of the generation step is formulated as,

$$L_G = \lambda_p\mathcal{L}_p + \lambda_v\mathcal{L}_v + \lambda_k\mathcal{L}_k + \lambda_e\mathcal{L}_e + \lambda_w\mathcal{L}_w + \lambda_G\mathcal{L}_G, \tag{5}$$

where $\lambda_p, \lambda_v, \lambda_k, \lambda_e, \lambda_w$ and λ_G are the weights of loss functions. And the loss of the discrimination step is formulated as $L_D = \mathcal{L}_D$. We follow the standard GAN practice [27] to train the model.

Figure 6. Comparison of same-identity video reconstruction results obtained by the proposed method and other state-of-the-art approaches.

Table 1. Quantitative results of different methods on four datasets for the same-identity video reconstruction.

Methods	VoxCeleb2					TalkHead-1KH					CelebV-HQ					VFHQ				
	L_1	MS-SSIM	PSNR	FID	AKD	L_1	MS-SSIM	PSNR	FID	AKD	L_1	MS-SSIM	PSNR	FID	AKD	L_1	MS-SSIM	PSNR	FID	AKD
FOMM [46]	0.0481	0.838	23.02	25.90	1.219	0.0431	0.821	23.28	33.22	2.905	0.0602	0.769	21.85	62.84	3.453	0.0526	0.780	21.76	47.82	2.868
MRAA [47]	0.0353	0.881	25.94	26.23	0.929	0.0361	0.882	25.50	32.57	1.057	0.0568	0.777	22.33	64.23	2.863	0.0454	0.812	22.60	40.17	2.123
OSFV [51]	0.0403	0.865	25.66	30.21	1.279	0.0432	0.837	23.59	35.12	3.100	0.0589	0.746	21.56	67.40	2.432	0.0491	0.804	21.79	41.95	1.730
TPSMM [67]	0.0318	0.902	26.88	24.39	0.709	0.0359	0.886	25.53	32.77	0.983	0.0615	0.757	22.05	64.89	3.714	0.0516	0.780	22.10	40.84	2.254
LIA [52]	0.0538	0.846	22.29	30.23	1.049	0.0477	0.879	24.43	38.89	0.932	0.0654	0.754	20.75	65.15	2.287	0.0537	0.815	21.47	42.27	1.502
DaGAN [22]	0.0359	0.881	25.64	24.92	0.844	0.0413	0.846	23.95	34.35	2.405	0.0637	0.739	21.32	68.04	4.800	0.0453	0.826	22.56	37.36	1.523
Face2Face$^{\rho}$ [59]	0.0507	0.816	20.83	31.71	1.332	0.0466	0.832	22.45	37.64	1.772	0.0709	0.710	19.94	71.87	3.754	0.0649	0.764	19.55	84.57	1.863
PECHead	**0.0304**	**0.905**	**26.96**	**23.05**	**0.626**	**0.0357**	**0.903**	**26.76**	**30.10**	**0.746**	**0.0552**	**0.803**	**24.29**	**56.68**	**1.215**	**0.0435**	**0.859**	**23.03**	**31.20**	**0.839**

4. Experiments

Datasets. We evaluate our model on VoxCeleb2 [10], TalkingHead-1KH [51], CelebV-HQ [68], and VFHQ [58].

Implementation Details. In the generator G, the encoder and decoder are composed of two downsample and upsample ResBlocks [20]. The Estimator E consists of four downsample and upsample AdaIN [26] based ResBlocks. The Reconstructor R is separately trained and the landmarks obtained by the widely used framework [3]. More details about the datasets, network structures, and settings are provided in the supplementary material.

Baselines. We compare our approach with the recently proposed representative methods, FOMM [46], MRAA [47], OSFV [51], TPSMM [67], LIA [52], Face2Face$^{\rho}$ [59] and DaGAN [22]. Our re-implementation of OSFV [51] is used with all settings followed by the original paper, while all other methods use the official implementation.

Metrics. We use L_1, MS-SSIM [53, 54], and PSNR to evaluate the low-level similarity between the synthesis and the driving images. We also leverage FID [21] and FVD [50] to assess the image and video quality. The average keypoint distance (AKD) [45, 46] is adopted to measure the semantic consistency. The cross-identity similarity (CSIM) [51] is used to evaluate the identity preservation for cross-identity video face reenactment. The average rotation distance (ARD) [13] and the facial action unit hamming distance (AUH) [13] are to measure errors of head pose angles and facial expressions. For MS-SSIM, PSNR, and CSIM, larger values indicate better results, others the opposite.

Table 2. Quantitative results for the cross-identity reenactment.

Methods	CelebV-HQ				VFHQ			
	CSIM	ARD	AUH	FVD	CSIM	ARD	AUH	FVD
FOMM [46]	0.687	2.76	0.174	202.5	0.675	2.18	0.174	211.7
MRAA [47]	0.670	2.65	0.145	219.1	0.662	2.07	0.159	205.9
OSFV [51]	0.706	3.21	0.171	207.3	0.754	4.11	0.205	213.4
TPSMM [67]	0.673	1.85	0.125	220.2	0.674	1.84	0.143	207.8
LIA [52]	0.713	2.68	0.143	199.5	0.712	2.48	0.170	213.8
DaGAN [22]	0.716	2.66	0.154	205.9	0.684	1.91	0.143	217.6
Face2Face$^{\rho}$ [59]	0.535	9.91	0.251	232.5	0.673	2.13	0.170	206.4
PECHead	**0.733**	**0.85**	**0.118**	**192.2**	**0.789**	**0.81**	**0.104**	**201.6**

4.1. Same-identity Video Reconstruction

We compare our models with state-of-the-art techniques for self-reenactment, where the source and driving frames depict the same individual. Quantitative results are presented in Tab. 1, and qualitative results are shown in Fig. 6.

Figure 7. Comparison of cross-identity face reenactment results obtained by the proposed method and other approaches.

Table 3. Quantitative results of pose and expression editing.

Methods	TalkHead-1KH			VFHQ		
	ARE	FID	AUH	ARE	FID	AUH
OSFV [51]	4.89	**40.96**	0.136	3.46	**53.21**	0.158
Face2Face$^{\rho}$ [59]	2.44	88.71	0.121	2.11	125.72	0.141
PECHead	**1.15**	42.04	**0.075**	**0.93**	56.16	**0.080**

Our models demonstrate significant improvements across all metrics. Most existing methods can generate satisfactory results for small pose movements, but for scenarios with significant pose variations, keypoint-based methods (FOMM [46], MRAA [47], TPSMM [67]) may produce distorted faces due to a lack of 3D facial constraints. The OSFV method [51], which uses 3D keypoints, can produce consistent facial shapes, but the image quality is still unsatisfactory. The depth-based method DaGAN [22] has face distortion issues, indicating that self-supervised depth estimation is insufficient. Latent vector-based models (LIA [52]) perform poorly in capturing facial details and may entangle appearance information in the latent code. The Face2Face$^{\rho}$ method [59] performs poorly due to inaccurate motion estimation. Our method excels in preserving facial shape while accurately transferring facial expressions compared to existing techniques.

4.2. Cross-identity Video Face Reenactment

We conducted experiments on the TalkingHead-1KH [51] and VFHQ [58] datasets to explore cross-identity motion transfer, where the source and driving frames depict different individuals. As shown in Fig. 7, our method

Figure 8. Head pose and expression editing results.

can produce convincing cross-identity face reenactment results that are more realistic, particularly in terms of facial expressions, compared to other techniques. Keypoint-based methods (FOMM [46], MRAA [47], TPSMM [67]) struggle to produce convincing results with noticeable face distortion for samples with large pose variations. Face2Face$^{\rho}$ [59] performs poorly since the landmarks only represent facial parts, making it challenging to handle non-facial characteristics like hair. Quantitative results are presented in Tab. 2. Our method outperforms other techniques with the highest identity preservation ability and video quality, as well as the lowest pose angle and expression error. Supplementary materials provide additional results, as well as subsequent experiments.

Figure 9. Head pose and expression freely controllable face reenactment.

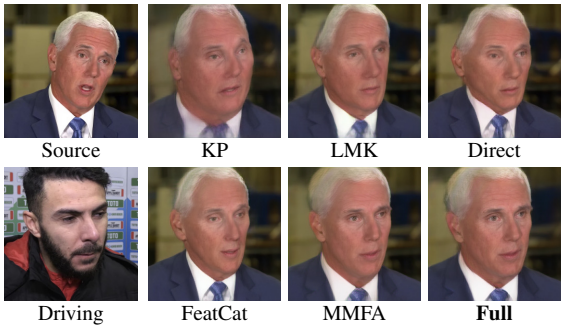


Figure 10. Ablation studies of our proposed model.

4.3. Head Pose and Expression Editing

For head pose editing, we compare the performance of frontalization. The metric average rotation error (ARE) [13] is adopted to measure the ability to control head pose, and FID is used to measure image quality. For expression editing, we compare the performance of expression transfer, and AUH is used to measure the expression error. Among the baselines, only the OSFV [51] and Face2Face$^{\rho}$ can manipulate the source frame without an explicit driving frame. The results are shown in Tab. 3 and Fig. 8. Although the OSFV [51] has slightly better FID scores, it can not manipulate the pose and expression well. The Face2Face$^{\rho}$ [51] fails to estimate the flow field, causing poor results.

4.4. Free Editing on Wild Identities

Finally, we demonstrate the strong capability of PECHead by editing wild face images downloaded from the Internet. The results are shown in Fig. 9. Our method can generate the face with desired head poses and expressions by changing the values of head pose p and expression e.

4.5. Ablation Studies

To validate the effectiveness of each component, we first evaluate the performance of using both self-supervised learned and predefined facial landmarks. We then assess the performance of the proposed MMFA module. Finally, we evaluate the performance of the proposed video-based framework involving the CAP module. Hence, we have the following settings: (1) Self-supervised learned landmarks only (KP); (2) Predefined landmarks only (LMK); (3) Directly merge the learned and predefined landmarks (Direct); (4) Concatenate the feature maps at a single level (FeatCat); (5) The Motion-Aware Multi-Scale Feature Alignment (MMFA); (6) The Full model. The results in Fig. 10 and Tab. 4 reveal three important conclusions. Firstly, utilizing both self-supervised learned landmarks and predefined landmarks is crucial to avoid face distortion and obtain high-quality results. Secondly, the motion-aware multi-scale feature alignment (MMFA) module effectively aligns features from different scales, resulting in high-quality outcomes. Lastly, the context adaptation and propagation (CAP) module propagates context information across frames, improving the smoothness of video synthesis. Notably, only our full model produces high-fidelity results.

Table 4. Quantitative results for ablation studies.

Settings	TalkHead-1KH					VFHQ				
	L_1	FID	CSIM	ARD	FVD	L_1	FID	CSIM	ARD	FVD
KP	0.0446	35.82	0.726	1.41	215.8	0.0491	37.8	0.712	1.40	218.5
LMK	0.0426	37.30	0.717	1.29	213.9	0.0485	36.6	0.709	1.37	217.9
Direct	0.0439	35.58	0.730	1.37	212.7	0.0474	32.9	0.724	1.33	217.8
FeatCat	0.0430	34.96	0.732	1.34	208.2	0.0462	32.0	0.733	1.09	213.9
MMFA	0.0375	31.27	0.764	0.81	206.8	0.0448	**31.0**	0.782	0.85	209.9
Full	**0.0357**	**30.10**	**0.779**	**0.79**	**199.6**	**0.0435**	31.2	**0.789**	**0.84**	**201.6**

5. Conclusion

In this work, we present a novel method called PECHead, which generates high-fidelity talking head videos with free control over head pose and expression. Leveraging both learned and predefined landmarks, we introduce a motion-aware multi-scale feature alignment module to model global and local movements simultaneously. Furthermore, to improve the smoothness and naturalness of video synthesis, we introduce a context adaptation and propagation module that adapts the context of previous frames. Our method outperforms existing approaches in face reenactment and controllable talking head generation, achieving state-of-the-art results.

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*, pages 187–194, 1999. 4

[2] Stella Bounareli, Vasileios Argyriou, and Georgios Tzimiropoulos. Finding directions in gan's latent space for neural face reenactment. *arXiv preprint arXiv:2202.00046*, 2022. 2

[3] Adrian Bulat and Georgios Tzimiropoulos. 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 IEEE International Conference on Computer Vision*, pages 1021–1030, 2017. 4, 6

[4] 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, 2, 3

[5] Chen Cao, Qiming Hou, and Kun Zhou. Displaced dynamic expression regression for real-time facial tracking and animation. *ACM Transactions on graphics (TOG)*, 33(4):1–10, 2014. 3

[6] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5933–5942, 2019. 2, 3

[7] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Understanding deformable alignment in video super-resolution. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 973–981, 2021. 5

[8] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5972–5981, 2022. 2, 5

[9] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014. 5

[10] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018. 6

[11] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. 5

[12] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2, 4

[13] Michail Christos Doukas, Stefanos Zafeiriou, and Viktoriia Sharmanska. Headgan: One-shot neural head synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14398–14407, 2021. 2, 3, 6, 8

[14] Jean Duchon. Splines minimizing rotation-invariant seminorms in sobolev spaces. In *Constructive theory of functions of several variables*, pages 85–100. Springer, 1977. 2, 3

[15] E Friesen and Paul Ekman. Facial action coding system: a technique for the measurement of facial movement. *Palo Alto*, 3(2):5, 1978. 3

[16] Thomas Gerig, Andreas Morel-Forster, Clemens Blumer, Bernhard Egger, Marcel Luthi, Sandro Schönborn, and Thomas Vetter. Morphable face models-an open framework. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 75–82. IEEE, 2018. 3, 4

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[18] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural head avatars from monocular rgb videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18653–18664, 2022. 3

[19] Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5784–5794, 2021. 3

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 4, 6

[21] 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. *Advances in neural information processing systems*, 30, 2017. 6

[22] Fa-Ting Hong, Longhao Zhang, Li Shen, and Dan Xu. Depth-aware generative adversarial network for talking head video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3397–3406, 2022. 2, 3, 4, 6, 7

[23] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. Headnerf: A real-time nerf-based parametric head model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20374–20384, 2022. 3

[24] 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*, pages 642–650, 2022. 2

[25] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008. 3

[26] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. 4, 6

[27] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 2, 4, 5

[28] 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. 5

[29] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 4

[30] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 4

[31] Nan Rosemary Ke, Anirudh Goyal ALIAS PARTH GOYAL, Olexa Bilaniuk, Jonathan Binas, Michael C Mozer, Chris Pal, and Yoshua Bengio. Sparse attentive backtracking: Temporal credit assignment through reminding. *Advances in neural information processing systems*, 31, 2018. 5

[32] Taras Khakhulin, Vanessa Sklyarova, Victor Lempitsky, and Egor Zakharov. Realistic one-shot mesh-based head avatars. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II*, pages 345–362. Springer, 2022. 3

[33] Jiahao Li, Bin Li, and Yan Lu. Deep contextual video compression. *Advances in Neural Information Processing Systems*, 34:18114–18125, 2021. 2

[34] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017. 3

[35] Jae Hyun Lim and Jong Chul Ye. Geometric gan. *arXiv preprint arXiv:1705.02894*, 2017. 5

[36] Jiangjing Lv, Xiaohu Shao, Junliang Xing, Cheng Cheng, and Xi Zhou. A deep regression architecture with two-stage re-initialization for high performance facial landmark detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3317–3326, 2017. 3

[37] Luming Ma and Zhigang Deng. Real-time hierarchical facial performance capture. In *Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, pages 1–10, 2019. 3

[38] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 3

[39] Jacek Naruniec, Leonhard Helminger, Christopher Schroers, and Romann M Weber. High-resolution neural face swapping for visual effects. In *Computer Graphics Forum*, volume 39, pages 173–184. Wiley Online Library, 2020. 2

[40] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsganv2: Improved subject agnostic face swapping and reenactment. *arXiv preprint arXiv:2202.12972*, 2022. 2

[41] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346, 2019. 4

[42] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European conference on computer vision (ECCV)*, pages 818–833, 2018. 3

[43] Yurui Ren, Ge Li, Yuanqi Chen, Thomas H Li, and Shan Liu. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13759–13768, 2021. 3

[44] Zhixin Shu, Mihir Sahasrabudhe, Riza Alp Guler, Dimitris Samaras, Nikos Paragios, and Iasonas Kokkinos. Deforming autoencoders: Unsupervised disentangling of shape and appearance. In *Proceedings of the European conference on computer vision (ECCV)*, pages 650–665, 2018. 3

[45] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Animating arbitrary objects via deep motion transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2377–2386, 2019. 2, 3, 6

[46] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 2, 3, 4, 5, 6, 7

[47] Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13653–13662, 2021. 2, 3, 5, 6, 7

[48] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020. 5

[49] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016. 2, 3

[50] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 6

[51] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10039–10049, 2021. 2, 3, 4, 5, 6, 7, 8

[52] Yaohui Wang, Di Yang, Francois Bremond, and Antitza Dantcheva. Latent image animator: Learning to animate images via latent space navigation. *arXiv preprint arXiv:2203.09043*, 2022. 2, 3, 6, 7

[53] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[54] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee, 2003. 6

[55] Olivia Wiles, A Koepke, and Andrew Zisserman. X2face: A network for controlling face generation using images, audio, and pose codes. In *Proceedings of the European conference on computer vision (ECCV)*, pages 670–686, 2018. 3

[56] Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. Look at boundary: A boundary-aware face alignment algorithm. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2129–2138, 2018. 3

[57] Wayne Wu, Yunxuan Zhang, Cheng Li, Chen Qian, and Chen Change Loy. Reenactgan: Learning to reenact faces via boundary transfer. In *Proceedings of the European conference on computer vision (ECCV)*, pages 603–619, 2018. 2

[58] Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 657–666, 2022. 6, 7

[59] Kewei Yang, Kang Chen, Yuan-Chen Guo, Daoliang Guo, Song-Hai Zhang, and Weidong Zhang. Face2face$^{\rho}$: Real-time high-resolution one-shot face reenactment. In *European conference on computer vision*. Springer, 2022. 2, 3, 6, 7

[60] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5525–5533, 2016. 3

[61] Guangming Yao, Yi Yuan, Tianjia Shao, Shuang Li, Shanqi Liu, Yong Liu, Mengmeng Wang, and Kun Zhou. One-shot face reenactment using appearance adaptive normalization. *arXiv preprint arXiv:2102.03984*, 2021. 2

[62] Egor Zakharov, Aleksei Ivakhnenko, Aliaksandra Shysheya, and Victor Lempitsky. Fast bi-layer neural synthesis of one-shot realistic head avatars. In *European Conference on Computer Vision*, pages 524–540. Springer, 2020. 3

[63] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9459–9468, 2019. 2, 3

[64] Xianfang Zeng, Yusu Pan, Mengmeng Wang, Jiangning Zhang, and Yong Liu. Realistic face reenactment via self-supervised disentangling of identity and pose. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12757–12764, 2020. 3

[65] Jiangning Zhang, Xianfang Zeng, Mengmeng Wang, Yusu Pan, Liang Liu, Yong Liu, Yu Ding, and Changjie Fan. Freenet: Multi-identity face reenactment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5326–5335, 2020. 2

[66] Yunxuan Zhang, Siwei Zhang, Yue He, Cheng Li, Chen Change Loy, and Ziwei Liu. One-shot face reenactment. *arXiv preprint arXiv:1908.03251*, 2019. 2

[67] Jian Zhao and Hui Zhang. Thin-plate spline motion model for image animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3657–3666, 2022. 2, 3, 4, 5, 6, 7

[68] Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. Celebv-hq: A large-scale video facial attributes dataset. *arXiv preprint arXiv:2207.12393*, 2022. 6