

NOFA: NeRF-based One-shot Facial Avatar Reconstruction

WANGBO YU, Tencent AI Lab, China

YANBO FAN*, Tencent AI Lab, China

YONG ZHANG*, Tencent AI Lab, China

XUAN WANG*, Ant Group, China

FEI YIN, Tsinghua University, China

YUNPENG BAI, Tsinghua University, China

YAN-PEI CAO, Tencent AI Lab, China

YING SHAN, Tencent AI Lab, China

YANG WU, Tencent AI Lab, China

ZHONGQIAN SUN, Tencent AI Lab, China

BAOYUAN WU, The Chinese University of Hong Kong, Shenzhen, China

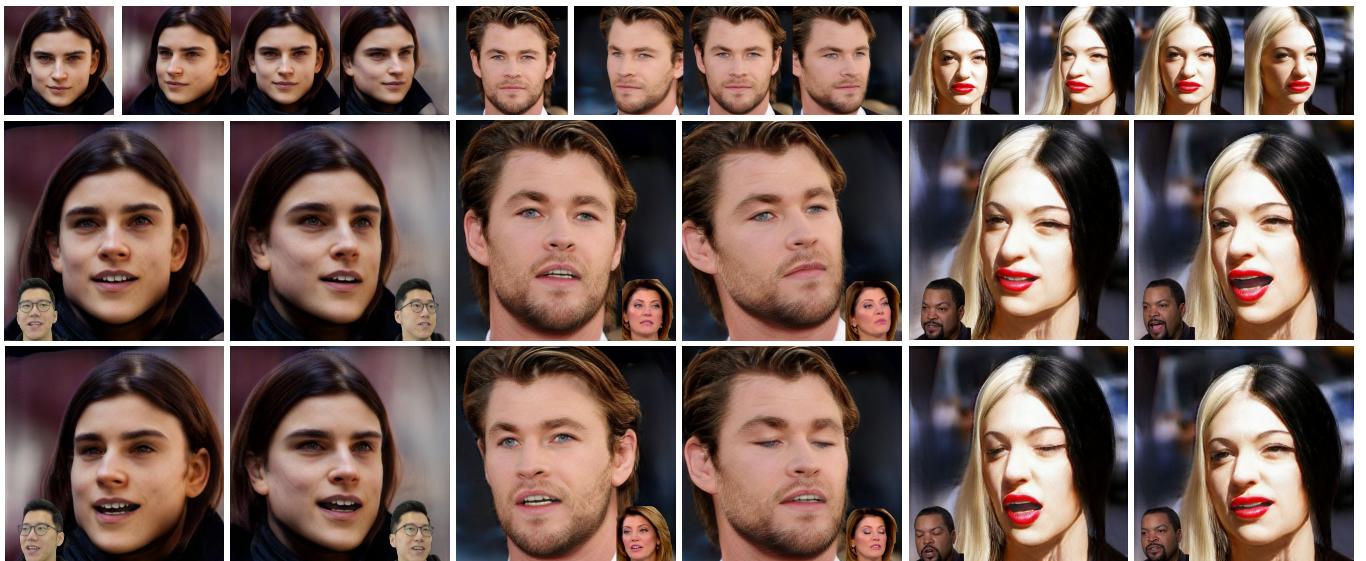


Fig. 1. Our method enables high-fidelity facial avatar reconstruction and reenactment given a single input image. The first row shows the input image and novel view synthesis results, the following rows show the facial reenactment results, where the facial motion of the avatars are controlled by the driving faces.

3D facial avatar reconstruction has been a significant research topic in computer graphics and computer vision, where photo-realistic rendering and flexible controls over poses and expressions are necessary for many related applications. Recently, its performance has been greatly improved with the development of neural radiance fields (NeRF). However, most existing NeRF-based facial avatars focus on subject-specific reconstruction and reenactment, requiring multi-shot images containing different views of the specific subject for training, and the learned model cannot generalize to new identities, limiting its further applications. In this work, we propose a one-shot 3D facial avatar reconstruction framework that only requires a single source image to reconstruct a high-fidelity 3D facial avatar. For the challenges of lacking generalization ability and missing multi-view information, we leverage the generative prior of 3D GAN and develop an efficient encoder-decoder network to reconstruct the canonical neural volume of the source image, and further propose a compensation network to complement facial details. To enable fine-grained control over facial dynamics, we propose

a deformation field to warp the canonical volume into driven expressions. Through extensive experimental comparisons, we achieve superior synthesis results compared to several state-of-the-art methods.

CCS Concepts: • Computing methodologies → Animation.

Additional Key Words and Phrases: Facial avatar, Video synthesis, NeRF

1 INTRODUCTION

Facial avatar reconstruction has been an important research topic in the field of computer graphics and computer vision, due to its phenomenal applications in virtual reality (VR), augmented reality (AR), movie industry, and teleconferencing. High-fidelity face reconstruction and fine-grained face reenactment are foundations for those applications.

To animate facial images, several 2D approaches have been proposed by utilizing flow-based warping in image or feature spaces to transfer motion, and encoder-decoder networks to synthesize

*Yanbo Fan, Yong Zhang and Xuan Wang are the corresponding authors.

appearance [10, 14, 40, 41, 49, 51, 56]. By training on large-scale face video datasets [11, 49, 62] with a large number of identities, these methods are generalizable to new identities and can produce vivid reenactment results given just a single facial image of the source identity. However, they have no constraints on the underlying 3D facial geometry and can hardly generate multi-view consistent images. Besides, they suffer from artifacts under large driven poses or expressions. Meanwhile, conventional parametric face model [6, 33]-based methods [13, 15, 18, 38, 45, 55] model 3D faces with template mesh and 3DMM parameters [6]. They support the flexible controls over poses and expressions. However, these mesh-based methods are memory-inefficient and less effective in modeling the non-face region, such as teeth, hair, and accessories. The accuracy of reconstruction and animation is also limited by the number of blendshapes or templates.

Recently, the photo-realistic and multi-view consistent rendering ability of Neural Radiance Fields (NeRF) [30] has sparked several works for NeRF-based facial avatar reconstruction [4, 16, 19]. They perform facial reenactment by learning deformation field or rendering function conditioned on control signals, where the pose and expression coefficients of 3DMMs are most commonly used. There exist two limitations of those methods. First, they require a large number of images containing different poses and expressions of the target face for training, which are not always available in real scenarios. Second, they are subject-dependent that can only be used to generate images of the training identity, *i.e.*, they are not generalizable to new identities. The lack of generalization ability and the demand on extensive multi-shot training data limit their further applications.

To address the aforementioned limitations of the existing methods, we aim to propose a generalizable NeRF-based one-shot facial avatar reconstruction method, which can be applicable to any new identity given only a single facial image. There are three main challenges for the considered objective: **1)** due to the complex facial dynamics and missing 3D information, it is challenging to learn a faithful reconstruction from a single input image, **2)** it is difficult to endow generalization ability for personalized NeRF, and **3)** fine-grained control over facial expression remains an challenging task in NeRF, especially for the one-shot situation. To tackle the challenges, we take advantage of the generative prior of a NeRF-based 3D GAN [8]. The latent space of 3D GAN encodes rich 3D-consistent generative prior, which helps synthesize neural volumes of diverse human faces. To align the unconditional latent space with real images, we develop a model with the encoder-decoder framework, and train it on a large-scale video dataset [62] in an end-to-end manner. Specifically, the encoder projects the input image to the latent space while the decoder maps the latent code to a neural volume, which is then used for image synthesis via neural rendering. Different from GAN inversion that learns a latent code that can accurately reconstruct the input image, here we aim to project the input image into a shared canonical space with an aligned expression, which is crucial for modeling facial dynamics. Due to the inevitable information loss of the encoder, the reconstructed volume lacks image-specific details, leading to poor identity preservation. To supplement identity information leakage, we design a compensation network to learn

a compensatory neural volume based on the input image and the intermediate feature from the decoder.

For the sake of fine-grained face reenactment, we exploit a deformation field to model facial dynamics, which learns a conditional mapping to deform each sampled point in the target space into the canonical space according to the driven signals. We extend the personalized deformation field in existing works to a generalized deformation field to handle different test identities and driven signals, by conditioning it on both identity and expression coefficients of 3DMMs and training it on the large-scale video dataset [62]. Thanks to the rich training expressions offered by the dataset, our deformation field can better model large and extreme motions compared with the personalized methods. We conduct extensive experiments and compare to both 2D and NeRF-based face reconstruction and reenactment methods, and demonstrate our superior performance in novel view synthesis and reenactment tasks.

Our main contributions are in three-fold: **1)** we propose a NeRF-based one-shot facial avatar reconstruction method that supports high-fidelity 3D face reconstruction and vivid face reenactment from a single face image. **2)** Once trained, our method can be generalized to new identities, which is more efficient and practical than personalized methods. **3)** We compare our method with several state-of-the-art methods and obtain superior synthesis performance.

2 RELATED WORK

2.1 Neural Scene Representation

Neural radiance fields (NeRF) [30] represents 3D scenes using MLP-based implicit function and achieve compelling rendering quality in 3D reconstruction tasks. Benefiting from its inherently differentiable rendering process, NeRF can be trained simply using multi-view images and their corresponding camera labels, and has been widely used in the field of 3D modeling and novel view synthesis. However, the conventional NeRF cannot handle dynamic subjects. Several approaches have been devoted to work around this limitation [4, 16, 19, 20, 31, 34, 42, 47, 52, 54]. The solutions can be roughly categorized into two categories: a train of thought is to condition the radiance field on control signals, which will change the density and color of the observed points. Another train of thought is to additionally learn a deformation field that accepts control signals and coordinates as input and predicts coordinate offsets from the deformed space into canonical space. For NeRF-based facial avatar synthesis [4, 16, 17, 19, 20, 58, 61], the pose and expression coefficients of 3D Morphable Face Models (3DMMs) [6] are employed as control signals to model facial deformations. These works study subject-dependent reconstruction and cannot generalize to different identities. What's more, a large set of facial images of the given identity are needed for training. Differently, we study the subject-agnostic problem, where only a single portrait image is given for reconstruction, and propose a generalizable model that can handle different testing faces.

2.2 3D-aware Generative Networks

Inspired by the breakthroughs achieved by 2D Generative Adversarial Networks (GANs) [25–27], recent researches [9, 39] have extended 2D image generation into 3D settings by combining GANs

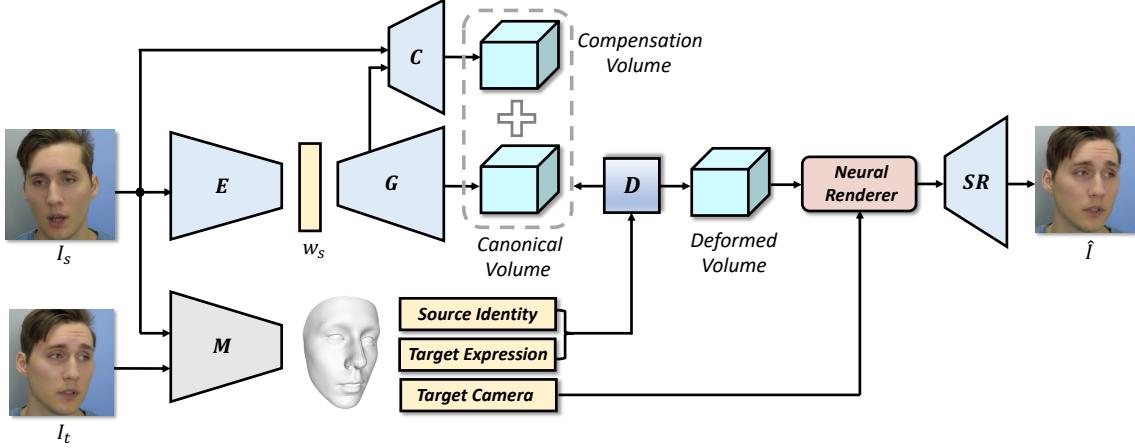


Fig. 2. Pipeline of NOFA. Given a source image I_s , an encoder E is adopted to embed the image into the latent space of volume generator G , which will produce the canonical volume V_c with an aligned expression while preserve the identity of I_s . The compensation network C is used to supplement image-specific details for V_c . To achieve explicit motion control, we employ pretrained M to extract 3DMM parameters from target image I_t and source image I_s , and use the combination of source identity and target expression parameters as control signal of the deformation field D , to deform the canonical volume with source identity into target expression. Finally, a hybrid neural renderer consists of volume rendering and super-resolution (SR) modules is adopted to render the final output \hat{I} given target camera parameters.

with the Implicit Neural Representations (INRs). These unconditional 3D GANs can generate photo-realistic rendering and enable controls over views. However, they do not support fine-grained and explicit expression controls. Recently [48] proposed a generative NeRF that overfits multiple identities at the same time, by learning subject-specific identity codes as the condition of NeRF MLPs. This method can be used for one-shot head avatar synthesis by fine-tuning the latent code and MLP parameters on a single source image. However, its training data are captured in studio conditions, and the one-shot synthesis results are of low quality due to its sparse latent space (only 15 identities are encoded). Some concurrent works [5, 43, 44, 53] further utilize the parameters of 3D Morphable Face Models (3DMMs) to explicitly control expression of the rendered faces. Yet, they are designed for unconditionally generating fake images and cannot be directly used in real applications. Besides, it is challenging for them to generate fine-grained motion because they are trained on discrete face image datasets. Different from these methods, we make use of the prior of 3D GAN and jointly train an encoder-decoder network on large-scale video datasets, achieving real image 3D reconstruction and vivid motion reenactment.

2.3 GAN Inversion

GAN inversion techniques act as a bridge for bringing GANs to real world applications such as image editing and reenactment. Existing GAN inversion approaches can be roughly divided into three categories: the optimization-based methods [1, 2], which optimize the latent codes by minimizing the distance between the ground truth image and the generated one, achieving promising reconstruction results yet limited by its low efficiency. The learning-based methods utilize an encoder network to directly encode the input images into latent codes [36] [3], equipping with high efficiency and generalization ability while the reconstruction results often lacks

fine details due to the information loss in the encoding process. The hybrid GAN inversion approaches utilize a learned encoder to predict an initial latent code and further refine it in the optimization process [64] [57], the generator parameters are also optimized in [37] to achieve better results. Despite their success in real image editing, GAN inversion is usually applied for editing global facial attributes such as age, makeup and gender. It is non-trivial to be used for fine-grained face reenactment.

2.4 One-shot Talking Head Synthesis

One-shot talking head synthesis aims to generate talking face videos from a given source image and a driving video. The generated videos should maintain the facial characteristics of the source image and the facial movements in the driving video. A large amount of works study these in the 2D image or feature space, where the key idea is to learn two separated networks to control motion and model appearance. For example, the works of [40, 41] predict warping flow from key-points to warp features of source images into target motion. The work of [35] uses 3DMM parameters to modulate flow generator and a refine network to supplement fine details. [56] further leverages the prior of StyleGAN [26] to enhance appearance. These methods are trained on the large-scale face video datasets [11, 49, 62] containing rich identities and expressions, thus can be generalized to unseen motion and identity given just a single input image. However, they cannot handle large pose changes due to the artifacts brought by feature warping. Some methods [14, 49] have devoted to address this problem by introducing 3D CNNs to produce 3D feature representation of the input image and apply 3D feature warping. Nevertheless, the learned representation doesn't model the underlying 3D facial geometry and the warping process lack explicit 3D constraints, causing poor multi-view consistency and can hardly be used in novel view synthesis.

3 METHOD

We propose NOFA, a NeRF-based one-shot facial avatar reconstruction framework, which will be described in detail in this section. We first introduce the networks employed for image to volume synthesis in Sec. 3.1. Then, we present the details of the 3DMM-guided deformation field for dynamic modeling in Sec. 3.2. Finally, we provide the loss functions used in the training stage as well as explaining the training strategy in Sec. 3.3.

3.1 Volume Synthesis with Generative Prior

In order to reconstruct high-fidelity facial avatars, the first and most important step is to build 3D representation of the given subject. As shown in Fig. 2, we leverage the generator of a pretrained state-of-the-art 3D GAN [8] to synthesize 3D representation and render images, which consists of a tri-plane generator G , a volumetric rendering module, and a super resolution module. G uses a StyleGAN2 [27] backbone to synthesize features of size $96 \times 256 \times 256$, which will be reshaped into a tri-plane of size $3 \times 32 \times 256 \times 256$. Then, features of each position in the tri-plane could be efficiently queried by coordinates and decoded to neural volumes for volumetric rendering, producing high-fidelity and view-consistent images.

We draw ideas from general GAN inversion approaches and achieve image to volume synthesis by exploiting learning-based GAN inversion. Specifically, given a source facial image I_s , we adopt a standard e4e [46] encoder to project the image to the latent space of G , *i.e.*, embedding I_s into a set of latent codes w_s that will modulate features in the tri-plane generator G for volume generation:

$$V_c = G(E(I_s) + \bar{w}), \quad (1)$$

where \bar{w} is the average latent code of G , and V_c denotes the output neural volume. Different from conventional inversion approaches that faithfully reconstruct the input image, the generated volume V_c is defined in the canonical space, *i.e.*, with an aligned expression instead of preserving the original expression of I_s . This is crucial for the following deformation process, where we gain explicit expression control by exploiting backward deformation. In our implementation, the canonical space is naturally learned by jointly training the whole framework on videos in an end-to-end manner, without explicit inner supervision.

As conjectured in [7] that the low-rate latent codes produced by the encoder are insufficient for high-fidelity reconstruction, we thus additionally learn a compensation network C to compensate the information loss caused by E .

Fig. 9 (c) shows the architecture of the compensation network. It contains several spatially-adaptive de-normalization (SPADE) residual blocks [32] and convolution, which takes the $64 \times 64 \times 512$ feature of the tri-plane generator G as input and progressively modulate and upscale the feature with source image I_s . The details of SPADE-ResBlock are given in Fig. 9 (a) and Fig. 9 (b). We discard the batch normalization (BN) layers in SPADE to better preserve the tri-plane features. The final output compensation volume is of the same size with the output of G . They are summed together to produce the neural volume with better identity and appearance preservation.

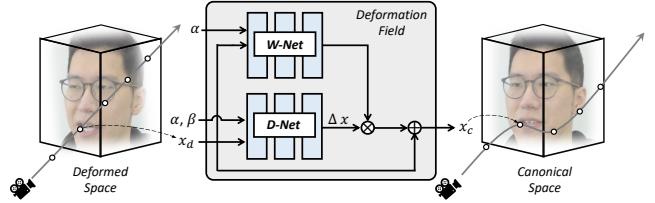


Fig. 3. Illustration of the deformation field. It consists of a deformation network (D-Net) and a weighting network (W-Net). D-Net regress the coordinates offsets from the deformed space to the canonical space. W-Net predicts the per-location weight scalars to multiply with the offsets. With the weighted offsets, we can query volume features defined in the canonical space for volumetric rendering.

3.2 Dynamic Modeling with 3DMM Guidance

In order to achieve explicit motion control on the reconstructed neural volume, we exploit a deformation field D to model facial dynamics and employ the semantic parameters of a 3DMM face model [33] as control signals. In 3DMM, the face shape is defined as:

$$S = \bar{S} + \alpha B_{id} + \beta B_{exp}, \quad (2)$$

where \bar{S} represents the average face shape, B_{id} and B_{exp} are the identity and expression basis computed by PCA [23]. We adopt the coefficients α and β as control signals, which are semantically meaningful and enable fine-detailed expression control.

During training, we employ an off-the-shelf 3D face reconstruction model [13] to estimate source identity parameter α_s , target expression parameter β_t , and target camera parameter c_t from target image I_t and source image I_s , and use the combination of α_s and β_t as the input control signal of the deformation field. Then, a hybrid neural renderer consisting of volume rendering and super-resolution (SR) is adopted to render the final output \hat{I} given target camera parameters c_t that model head rotation and translation.

In particular, the deformation field models the backward deformation that deforms 3D points in the target space to the canonical space. As shown in Fig. 3, the deformation field consists of a deformation network (D-Net) and a weighting network (W-Net). For each location x_d in the deformed space, we use D-Net to predict its canonical location x_c , and then query the canonical tri-Plane feature at x_c , which is used to regress the density and color of x_d for neural rendering. In addition to the target expression, we also condition D-Net on source identity to preserve the input identity, which is crucial for endowing it generalization ability. Inspired by the FLAME mesh model [29] that assigns skinning weights on mesh vertices for smooth blending, we additionally learn a W-Net to predict offset weights for each location.

3.3 Loss Functions and Training Strategy

General Training Stage. In the first stage, we train the base model without the compensation branch on a large-scale video dataset in an end-to-end fashion, by sampling source image I_s and target image I_t in the same video clip. During training, we use multiple objectives to ensure faithful reconstruction and vivid reenactment.

First, we apply a reconstruction loss between the synthesis image \hat{I} and the target I_t :

$$\mathcal{L}_{\text{rec}} = \|I_t - \hat{I}\|_2 + \text{LPIPS}(I_t, \hat{I}), \quad (3)$$

where $\text{LPIPS}(\cdot, \cdot)$ is the perceptual loss [59]. We also use the pixel-wise $L2$ distance to constrain high-level image contents.

We additionally utilize a mouth regularization loss in order to further enhance the mouth region thus derive more accurate facial motion. Specifically, we crop the mouth region from I_t and \hat{I} using ROI align [21], and apply the reconstruction loss on the cropped region, formulated as:

$$\mathcal{L}_{\text{mouth}} = \|crop(I_t) - crop(\hat{I})\|_2 + \text{LPIPS}(crop(I_t), crop(\hat{I})), \quad (4)$$

where $crop(I)$ represents the cropped region of image I .

For better identity preservation, we incorporate a face recognition loss between the synthesis image and the target image:

$$\mathcal{L}_{\text{id}} = 1 - \langle F(I_t), F(\hat{I}) \rangle, \quad (5)$$

where $F(\cdot)$ is the pre-trained ArcFace [12]. $\langle \cdot, \cdot \rangle$ is cosine distance.

Finally, we adopt a latent space regularization loss to force the encoder to produce latent codes closer to the average latent code of the pre-trained 3D GAN [8]:

$$\mathcal{L}_{\text{latent}} = \|w_s - \bar{w}\|_2. \quad (6)$$

It is helpful to regularize the 3D shape of the synthesized volume.

The total objective function of the general training is defined as:

$$\mathcal{L}_{\text{general}} = \lambda_{\text{rec}} \mathcal{L}_{\text{rec}} + \lambda_{\text{mouth}} \mathcal{L}_{\text{mouth}} + \lambda_{\text{id}} \mathcal{L}_{\text{id}} + \lambda_{\text{latent}} \mathcal{L}_{\text{latent}}, \quad (7)$$

where λ_{rec} , λ_{mouth} , λ_{id} , and λ_{latent} are trade-off hyperparameters. They are set as 1, 0.5, 0.1, and 0.01 respectively.

Teeth Refinement. After the first training stage, our model can produce nearly satisfying reconstruction and animation results except for the synthetic teeth, which are of great influence on the visual quality yet ignored by most other approaches. It is challenging to synthesize clear teeth due to the limited resolution and quality of the training videos. To tackle this problem, we exploit a face restoration model GFPGAN [50] to provide clear teeth as supervision and continue training the base model for teeth refinement while fixing the deformation field. Empirically, we observed that applying GFPGAN on the synthetic image \hat{I} as supervision produce better results than applying it on the target image I_t . We therefore apply the mouth regularization loss between \hat{I} and $\hat{I}^* = \text{GFPGAN}(\hat{I})$ for teeth refinement:

$$\mathcal{L}_{\text{teeth}} = \|crop(\hat{I}^*) - crop(\hat{I})\|_2 + \text{LPIPS}(crop(\hat{I}^*), crop(\hat{I})), \quad (8)$$

We also incorporate Eq. 3, Eq. 5, and Eq. 6 in the teeth refinement stage, and the objective function of this stage is:

$$\mathcal{L}_{\text{refine}} = \lambda_{\text{rec}} \mathcal{L}_{\text{rec}} + \lambda_{\text{teeth}} \mathcal{L}_{\text{teeth}} + \lambda_{\text{id}} \mathcal{L}_{\text{id}} + \lambda_{\text{latent}} \mathcal{L}_{\text{latent}}, \quad (9)$$

where we set $\lambda_{\text{rec}} = 0.5$, $\lambda_{\text{teeth}} = 1$, $\lambda_{\text{id}} = 0.1$, and $\lambda_{\text{latent}} = 0.01$.

Training Compensation Network. We add a compensation network to supplement fine details for the base model. Specifically, we fix the base model and train the compensation network on face image datasets in a self-supervised way. Given the input source image I_s , we adopt loss functions in Eq. 3 and Eq. 5 between I_s and the reconstructed image \hat{I} . In order to ensure depth and view consistency for the canonical volumes after compensation, we further enforce the output compensation volumes of the compensation network to share similar distribution with the original canonical volumes, using the following regularization term:

$$\mathcal{L}_{\text{depth}} = \|\mu V_c - V_m\|_2, \quad (10)$$

where V_m and V_c are the compensation volumes and canonical volumes respectively. $\mu = 0.1$ denotes a scale factor. The objective function of this stage is:

$$\mathcal{L}_{\text{comp}} = \lambda_{\text{rec}} \mathcal{L}_{\text{rec}} + \lambda_{\text{id}} \mathcal{L}_{\text{id}} + \lambda_{\text{depth}} \mathcal{L}_{\text{depth}}, \quad (11)$$

where we set $\lambda_{\text{rec}} = 1$, $\lambda_{\text{id}} = 0.1$ and $\lambda_{\text{depth}} = 0.01$.

One-shot Fine-tuning. When facing challenging cases, we can apply fast adaption on the input source image by fine-tuning the compensation network, which is much faster than fine-tuning the base model using PTI [37]. Specifically, we optimize the parameters of the compensation network using Eq. 3, Eq. 4, Eq. 5 and Eq. 10. The total loss function is:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{rec}} \mathcal{L}_{\text{rec}} + \lambda_{\text{mouth}} \mathcal{L}_{\text{mouth}} + \lambda_{\text{id}} \mathcal{L}_{\text{id}} + \lambda_{\text{depth}} \mathcal{L}_{\text{depth}}, \quad (12)$$

where we set $\lambda_{\text{rec}} = 1$, $\lambda_{\text{mouth}} = 0.5$, $\lambda_{\text{id}} = 0.1$ and $\lambda_{\text{depth}} = 0.01$.

4 EXPERIMENTS

4.1 Implementation Details

Datasets. We train our base model on the CelebV-HQ dataset [63] which contains 35,666 video clips involving 15,653 identities, and train the compensation network on the FFHQ dataset [26]. We crop and align faces from the videos and extract per-frame 3DMM parameters including identity, expression, and camera parameters for training. The training videos and images are resized into 512×512 . During inference, we apply camera movement over the reconstructed volumes to obtain head rotation and translation, and can produce continuous head movements without the limitation of the alignment.

Evaluation Metrics. We adopt several metrics to evaluate reconstruction and reenactment quality. The peak signal-to-noise ratio (PSNR), Structural Similarity (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS) [59] are exploited to evaluate synthetic quality. We also use Frechet Inception Distance (FID) [22] to measure difference between the synthetic and real distributions. We calculate the cosine similarity (CSIM) between the source and generated images to evaluate identity preservation. For reenactment quality, we extract 3DMM expression and pose parameters from synthetic and real images to compute their Average Expression Distance (AED) and the Average Pose Distance (APD), following [35].

Training Details. Our framework is trained on 8 Nvidia Tesla V100 GPUs in three stages. During training, the ADAM optimizer is adopted with a learning rate of 10^{-4} . In the general training

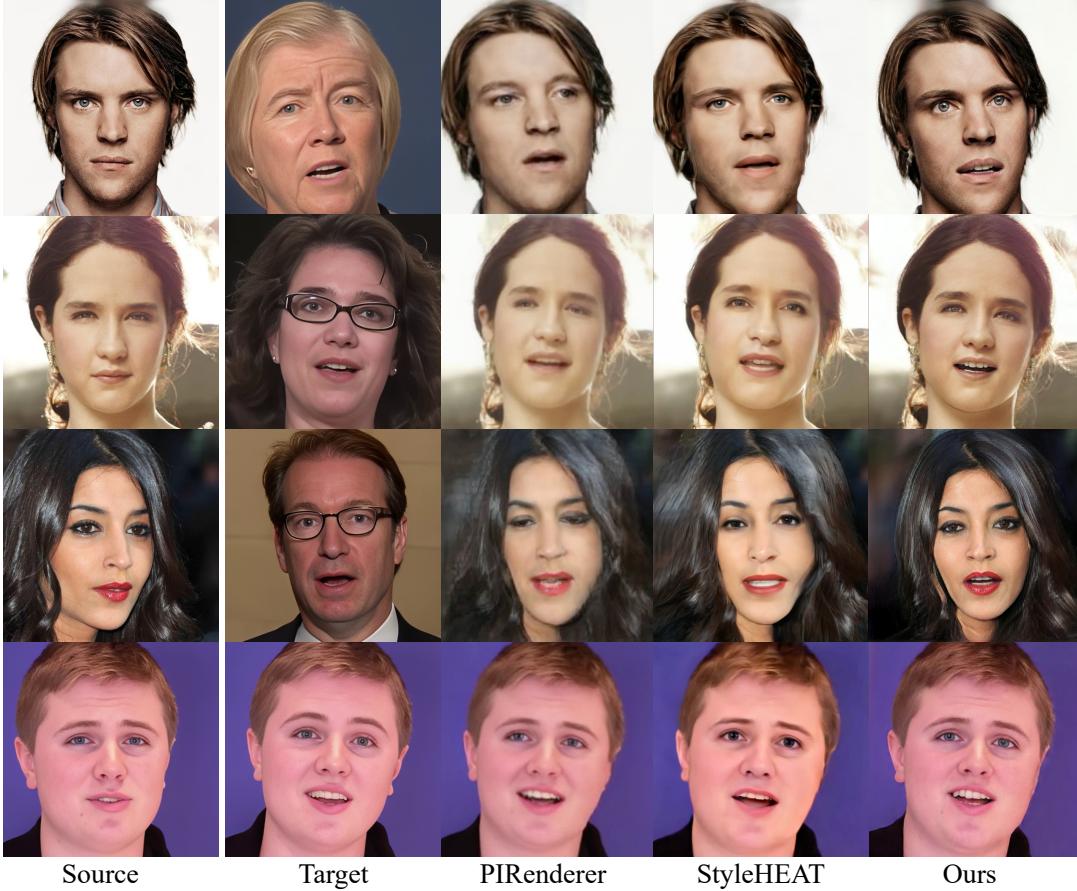


Fig. 4. Qualitative comparison with 2D approaches on cross-reenactment (top 3 rows) and self-reenactment (last row). Our approach achieves better reconstruction quality than previous state-of-the-arts.

stage, we train the base model without the compensation network on videos for 250K iterations with a batch size of 16. In the teeth refinement stage, we fix the deformation field and train the other parts of the base model with the same batch size for 25K iterations. Finally, we fix the base model and train the compensation network on images for 100K iterations with the batch size set to 32. The training takes about 4 days on 8 V100 GPUs.

Modeling Head Translation. Similar to [5, 43, 44, 53], we use camera poses to model head rotation and translation in our implementation. However, the face tracking model [13] used for estimating camera parameters requires face alignment as pre-processing. As a result, the estimated translation parameters are relative translations, and using them directly would cause center-aligned videos, as demonstrated in [5, 43, 44, 53]. To describe the absolute head translation, we use facial key-points estimated from the unaligned driving videos to compute pixel offsets between the face center of each driving frame and the source image, and convert them into camera coordinate offsets to rectify camera translation. In this way, the generated face can move freely within a fixed bounding box without the limitation of center align.

Same-ID	FID ↓	LPIPS ↓	PSNR ↑	SSIM ↑	CSIM ↑	AED ↓	APD ↓
PIRenderer [35]	27.14	0.2252	30.96	0.6028	0.7797	0.1073	0.01459
StyleHEAT [56]	18.02	0.1729	31.21	0.6019	0.7475	0.1151	0.01664
Ours	16.94	0.1481	31.78	0.6175	0.8031	0.1091	0.01142
Cross-ID	FID ↓	—	—	—	CSIM ↑	AED ↓	APD ↓
PIRenderer [35]	108.56	—	—	—	0.4812	0.2554	0.02962
StyleHEAT [56]	91.28	—	—	—	0.4890	0.2630	0.03484
Ours	84.47	—	—	—	0.5397	0.2581	0.01633

Table 1. Quantitative comparison with 2D approaches on face reenactment. The metrics indicate that our approach achieves the best reconstruction quality, comparable expression accuracy and the best pose accuracy compared with 2D approaches.

4.2 Comparisons

Comparison with 2D Reenactment Approaches. We first evaluate our approach in comparison with two state-of-the-art 3DMM-based 2D talking face generation approaches: PIRenderer [35] and StyleHEAT [56]. Both of them supports pose control via 3DMM parameters and video-driven face reenactment.

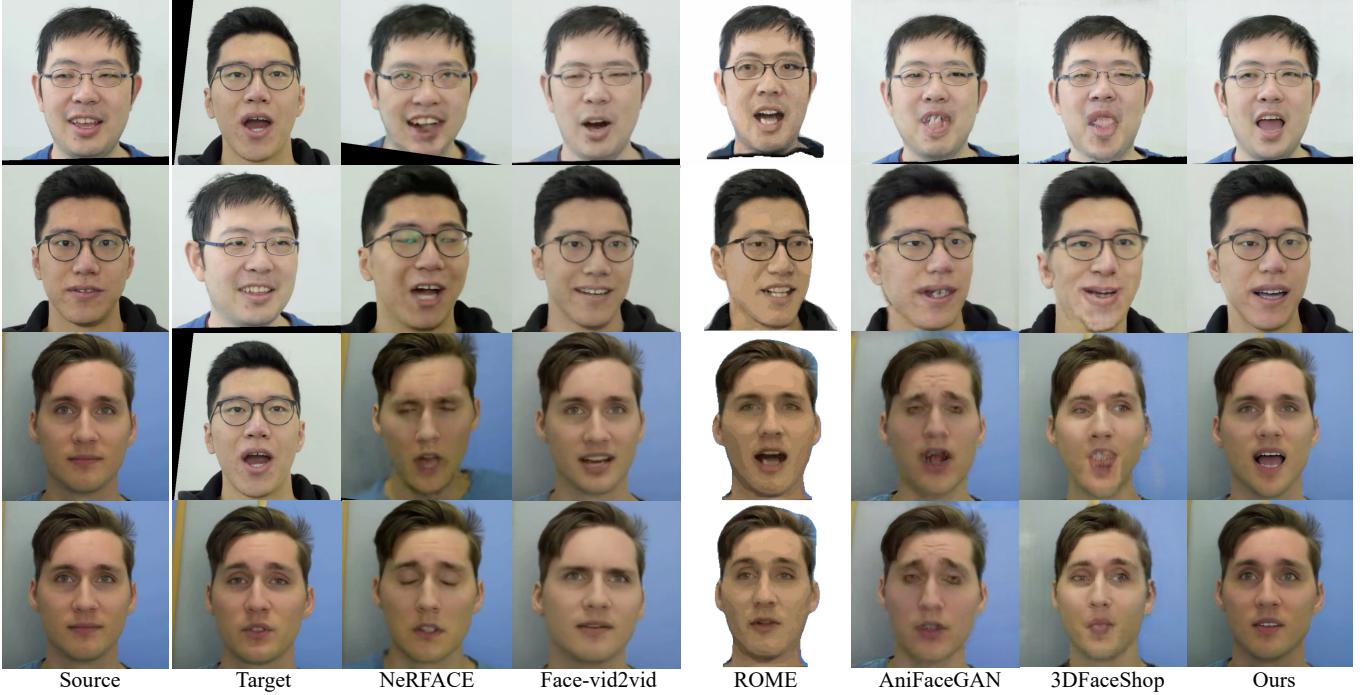


Fig. 5. Qualitative comparison with 3D-aware approaches on cross-reenactment (top 3 rows) and self-reenactment (last row). Our approach synthesizes facial images with higher fidelity and less artifacts comparing with previous state-of-the-arts.

Same-ID	FID ↓	LPIPS ↓	PSNR ↑	SSIM ↑	CSIM ↑	AED ↓	APD ↓
NeRFACE [16]	12.58	0.0925	30.87	0.6637	0.7846	0.1103	0.01071
Face-vid2vid [49]	20.69	0.2035	30.98	0.6191	0.7912	0.0995	0.01250
ROME [28]	25.83	0.2314	30.85	0.6007	0.7206	0.1224	0.01223
AniFaceGAN [53]	24.39	0.2105	29.86	0.6033	0.7754	0.1480	0.01288
3DFaceShop [44]	22.75	0.2154	30.07	0.6095	0.7516	0.1391	0.01167
Ours	16.41	0.1377	31.09	0.6175	0.7953	0.1195	0.01128
Cross-ID	FID ↓	—	—	—	CSIM ↑	AED ↓	APD ↓
NeRFACE [16]	157.38	—	—	—	0.3504	0.2554	0.02391
Face-vid2vid [49]	92.41	—	—	—	0.5024	0.2369	0.02541
ROME [28]	95.07	—	—	—	0.4693	0.2670	0.02011
AniFaceGAN [53]	93.47	—	—	—	0.4944	0.2721	0.02150
3DFaceShop [44]	92.53	—	—	—	0.5172	0.2855	0.02084
Ours	88.61	—	—	—	0.5524	0.2438	0.01872

Table 2. Quantitative comparison with 3D-aware approaches. In self-reenactment, our one-shot approach achieves comparable performance against NeRFACE that is trained on 1K frames, and surpasses the other approaches. In cross-reenactment, our approach outperforms the other approaches across most of the metrics, and achieves the second-highest expression accuracy.

For face reenactment evaluation, we conduct two types of reenactment tasks, *i.e.*, self-reenactment and cross-reenactment. For self-reenactment, the identity of source image is the same with the driving frames; for cross-reenactment, the source image and driving frames come from two different identities. The latter setting is much more challenging because of the facial feature gap between the source and driving faces. Following [56], we use 20 video clips with a total of 10K frames from HDTF dataset [60] for self-reenactment

evaluation. For cross-reenactment evaluation, we use the first 1,000 images from the CelebA-HQ dataset [24] as source images and the 20 HDTF videos as driving videos. Similar as [56], we perform one-shot fine-tuning on the source image to achieve better visual quality.

Fig. 4 shows the qualitative reenactment results of our approach and other state-of-the-arts, where our approach achieves better reconstruction quality in terms of identity preservation and detail textures. When dealing with side faces, feature warping-based 2D approaches fails to inference reasonable frontal faces and suffers from severe artifacts, while our approach successfully reconstruct accurate and reasonable frontal face. Table 1 shows the quantitative evaluation results, which demonstrate that our approach achieves better reconstruction results. For facial motion modeling, the average expression distance (AED) indicates that our 3D approach yields competitive animation results compared with 2D PIRenderer [35]. For head rotation modeling, our approach surpasses the two baselines with a large margin in terms of the average pose distance (APD), because the head poses are directly controlled by the external camera parameters in our implementation.

Comparison with 3D-aware Reenactment Approaches. We further compare with several 3D-aware approaches: NeRFACE [16], Face-vid2vid [49], ROME [28], AnifaceGAN [53] and 3DFaceShop [44]. Specifically, NeRFACE is a multi-shot NeRF avatar reconstruction approach. AnifaceGAN and 3DFaceShop are unconditional 3D GANs, we adopt PTI [37] to apply them on real images. We conduct self- and cross-reenactment experiments on the NeRFACE dataset [16],



Fig. 6. Visualization of canonical space. The synthetic images are directly rendered from canonical volumes without deformation, sharing the same aligned expression

where we use the first 1K frames of each video to train NeRFACE, while keep the one-shot setting of other approaches.

Fig. 5 shows the qualitative results. For self-reenactment, NeRFACE generates several expressions inconsistent with the driving image; for cross-reenactment, NeRFACE suffers from severe artifacts and produces inconsistent expressions. What's more, Face-vid2vid produces inaccurate poses because it doesn't model the 3D geometry. ROME suffers from oversmoothed appearance due to its mesh representation. AnifaceGAN and 3DFaceShop fail to keep fidelity under unseen views, and suffer from inaccurate motion and teeth artifacts. In contrast, our model ensures fine-grained motion control and high-fidelity across views. The quantitative results are listed in Table. 2. For self-reenactment evaluation, our one-shot approach achieves comparable performance against NeRFACE that is trained on 1K frames, and surpasses the other approaches. More importantly, In the more challenging and applied cross-reenactment task, our approach outperforms the other competitors across most of the metrics, and achieves the second-highest expression accuracy.

4.3 Ablation Study

Visualization of Canonical Space. Given the source image, our model produces neural volumes in the canonical space instead of preserving the original expression of source image. Given the target expressions, the back-ward deformation is applied by querying volumes in the canonical space, so that the canonical volumes are deformed into target expression. In our implementation, the canonical space is naturally learned by jointly training the whole framework on videos in an end-to-end manner, without explicit inner supervision. Fig. 6 shows examples of the canonical space, the output images are directly rendered from canonical volumes without deformation, sharing the same aligned expression.

Evaluation of Compensation Network. We design a compensation network to supplement identity and texture information for the

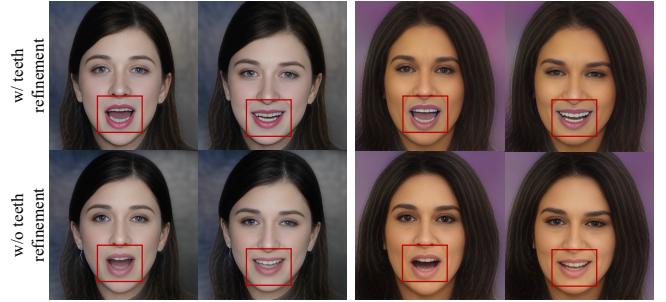


Fig. 7. Qualitative evaluation of teeth refinement. It helps to synthesize clearer teeth with less artifacts.

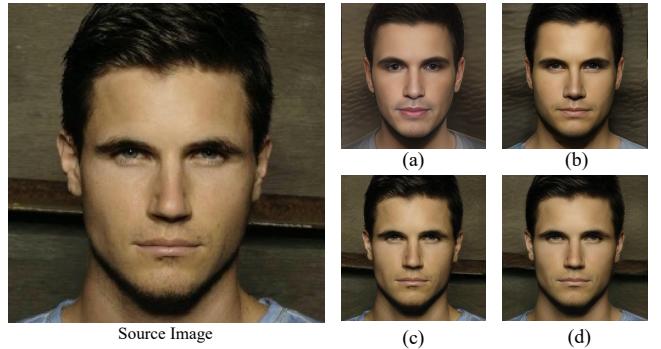


Fig. 8. Qualitative evaluation of the compensation network. (a). w/o compensation network, (b). w/ compensation network, (c). optimize generator using PTI [37], (d). optimize compensation network.

canonical volumes. As shown in Fig. 8 (a) and (b), given the source image, our base model yields correct 3D shape and expression yet fails to preserve the source identity. With the compensation network, the reconstruction quality is greatly improved. Our approach also supports one-shot fine-tuning to achieve more accurate reconstruction. As shown in Fig. 8 (c) and (d), fine-tuning the whole generator using PTI[37] produces photo-realistic reconstruction results, yet the training costs about **20** minutes on **4** V100 GPUs. With the compensation network, we can achieve comparable reconstruction results in a more time- and memory- efficient way, simply by optimizing the compensation network, which only takes about **10** minutes on a **singe** V100 GPU. We conducted self-reenactment experiments in Sec. 4.2, with the four different settings. Table. 3 shows the quantitative evaluation, which further demonstrate the effectiveness of the compensation network.

Effectiveness of Teeth Refinement. We incorporated a teeth refinement training stage to enable the base model to synthesize clearer teeth. Fig. 7 shows the faces generated before and after the teeth refinement stage. After teeth refinement, the synthetic teeth are clearer and show less artifacts, while without it the synthetic teeth tend to be blurry. We conduct self-reenactment experiment on the HDTF [60] dataset, and crop the mouth region of the synthetic images to compute metrics, the results are listed in Table.4. Through

Settings	FID ↓	LPIPS ↓	PSNR ↑	SSIM ↑	CSIM ↑
w/o Compensation	36.45	0.2429	29.75	0.5836	0.6024
w/ Compensation	25.63	0.2240	31.05	0.6007	0.7219
Optimize compensation network	18.75	0.1697	31.94	0.6042	0.7981
Optimize generator	16.94	0.1481	31.78	0.6175	0.8031

Table 3. Quantitative evaluation of the compensation network. The reconstruction quality is greatly improved with compensation. What's more, optimizing the compensation network results in a speed increase of eight times while achieving performance gains comparable to optimizing the entire generator.

Settings	FID ↓	LPIPS ↓	SSIM ↑
w/o teeth refinement	28.45	0.2678	0.3297
w/ teeth refinement	25.31	0.2215	0.3478

Table 4. Evaluation of teeth refinement. The quality of the synthetic teeth improved through the teeth refinement stage.

Weighting network	w/	w/o
AED ↓	0.1091	0.1143
Identity condition	w/	w/o
CSIM ↑	0.8031	0.7210

Table 5. Evaluation of weighting network and identity condition in the deformation field.

the teeth refinement stage, the quality of the reconstructed teeth improved.

Evaluation of the Deformation Field. We model facial dynamics using a deformation field. Typically, we learn a weighting network to increase motion accuracy, and use the 3DMM identity parameter as the condition for both the deformation network and the weighting network to preserve the source identity. We conduct self-reenactment experiment on the HDTF dataset to evaluate the designs. Specifically, we use the average expression distance (AED) to evaluate the effectiveness of the weighting network, and use CSIM to evaluate the effectiveness of identity condition. The results are listed in Table 5, which demonstrate the effectiveness of the proposed designs.

5 LIMITATIONS AND ETHICAL ISSUES

In our implementation, the head poses are modeled as camera poses, thus the background rotates as head rotates. We will explore the background modeling in future works.

Since our framework is capable of reconstructing high-fidelity facial avatar using only a single image, it may pose the risk of nefarious use such as deep-fakes. We are keenly aware of the potential for abuse of our approach, and we will explore the implementation of robust video watermarks for the synthesized videos, as well as develop tools to verify the authenticity.

REFERENCES

- [1] Rameen Abdal, Yipeng Qin, and Peter Wonka. 2019. Image2StyleGAN: How to Embed Images Into the StyleGAN Latent Space?. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.
- [2] Rameen Abdal, Yipeng Qin, and Peter Wonka. 2020. Image2stylegan++: How to edit the embedded images?. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [3] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. 2021. Restyle: A residual-based stylegan encoder via iterative refinement. In *Proceedings of the IEEE International Conference on Computer Vision (CVPR)*.
- [4] ShahRukh Athar, Zexiang Xu, Kalyan Sunkavalli, Eli Shechtman, and Zhixin Shu. 2022. RigNeRF: Fully Controllable Neural 3D Portraits. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 20364–20373.
- [5] Alexander W Bergman, Petr Kellnhofer, Yifan Wang, Eric R Chan, David B Lindell, and Gordon Wetzstein. 2022. Generative neural articulated radiance fields. *arXiv preprint arXiv:2206.14314* (2022).
- [6] Volker Blanz and Thomas Vetter. 1999. A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*. 187–194.
- [7] Yochai Blau and Tomer Michaeli. 2019. Rethinking lossy compression: The rate-distortion-perception tradeoff. In *International Conference on Machine Learning*. 675–685.
- [8] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Samih Khamis, Tero Karras, and Gordon Wetzstein. 2021. Efficient Geometry-aware 3D Generative Adversarial Networks. In *Proceedings of the IEEE International Conference on Computer Vision (CVPR)*.
- [9] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. 2021. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [10] Kun Cheng, Xiaodong Cun, Yong Zhang, Menghan Xia, Fei Yin, Mingrui Zhu, Xuan Wang, Jue Wang, and Nannan Wang. 2022. VideoReTalking: Audio-based Lip Synchronization for Talking Head Video Editing In the Wild. In *SIGGRAPH Asia 2022*.
- [11] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. 2018. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622* (2018).
- [12] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. 4690–4699.
- [13] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. 2019. 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*. 0–0.
- [14] Nikita Drobyshev, Jenya Chelishhev, Taras Khakhulin, Aleksei Ivakhnenko, Victor Lempitsky, and Egor Zakharov. 2022. Megaportraits: One-shot megapixel neural head avatars. *arXiv preprint arXiv:2207.07621* (2022).
- [15] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. 2021. Learning an animatable detailed 3D face model from in-the-wild images. *ACM Transactions on Graphics (TOG)* 40, 4 (2021), 1–13.
- [16] Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias Nießner. 2021. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8649–8658.
- [17] Xuan Gao, Chenglai Zhong, Jun Xiang, Yang Hong, Yudong Guo, and Juyong Zhang. 2022. Reconstructing personalized semantic facial nerf models from monocular video. *ACM Transactions on Graphics (TOG)*.
- [18] Pablo Garrido, Michael Zollhöfer, Dan Casas, Levi Valgaerts, Kiran Varanasi, Patrick Pérez, and Christian Theobalt. 2016. Reconstruction of personalized 3D face rigs from monocular video. *ACM Transactions on Graphics (TOG)* 35, 3 (2016), 1–15.
- [19] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. 2022. Neural head avatars from monocular RGB videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 18653–18664.
- [20] Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. 2021. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *Proceedings of the IEEE International Conference on Computer Vision*.
- [21] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision (ICCV)*. 2961–2969.
- [22] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*.

- [23] Ian T Jolliffe and Jorge Cadima. 2016. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374, 2065 (2016), 20150202.
- [24] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *International Conference on Learning Representations (ICLR)*.
- [25] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2021. Alias-free generative adversarial networks. *arXiv:2106.12423* (2021).
- [26] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4401–4410.
- [27] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- [28] Taras Khakhulin, Vanessa Sklyarova, Victor Lempitsky, and Egor Zakharov. 2022. Realistic one-shot mesh-based head avatars. In *ECCV 2022*.
- [29] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics (TOG)* 36, 6 (2017), 194–1.
- [30] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2020. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision (ECCV)*.
- [31] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. 2021. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 5865–5874.
- [32] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. 2019. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [33] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. 2009. A 3D face model for pose and illumination invariant face recognition. In *2009 sixth IEEE international conference on advanced video and signal based surveillance*. 296–301.
- [34] Amit Raj, Michael Zollhofer, Tomas Simon, Jason Saragih, Shunsuke Saito, James Hays, and Stephen Lombardi. 2021. Pixel-aligned volumetric avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 11733–11742.
- [35] Yurui Ren, Ge Li, Yuanqi Chen, Thomas H Li, and Shan Liu. 2021. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 13759–13768.
- [36] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. 2021. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [37] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. 2021. Pivotal Tuning for Latent-based Editing of Real Images. *arXiv preprint arXiv:2106.05744* (2021).
- [38] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael J Black. 2019. Learning to regress 3D face shape and expression from an image without 3D supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 7763–7772.
- [39] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. 2020. Graf: Generative radiance fields for 3d-aware image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- [40] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019. Animating arbitrary objects via deep motion transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2377–2386.
- [41] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019. First order motion model for image animation. *Advances in Neural Information Processing Systems (NIPS)* 32 (2019).
- [42] Shih-Yang Su, Frank Yu, Michael Zollhöfer, and Helge Rhodin. 2021. A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. *Advances in Neural Information Processing Systems (NIPS)* (2021), 12278–12291.
- [43] Jingxiang Sun, Xuan Wang, Lizhen Wang, Xiaoyu Li, Yong Zhang, Hongwen Zhang, and Yebin Liu. 2022. Next3D: Generative Neural Texture Rasterization for 3D-Aware Head Avatars. *arXiv preprint arXiv:2211.11208* (2022).
- [44] Junshu Tang, Bo Zhang, Binbin Yang, Ting Zhang, Dong Chen, Lihuang Ma, and Fang Wen. 2022. Explicitly controllable 3d-aware portrait generation. *arXiv preprint arXiv:2209.05434* (2022).
- [45] Ayush Tewari, Florian Bernard, Pablo Garrido, Gaurav Bharaj, Mohamed Elgharib, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. 2019. Fml: Face model learning from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10812–10822.
- [46] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. 2021. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)* 40, 4 (2021), 1–14.
- [47] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. 2021. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 12959–12970.
- [48] Daoya Wang, Prashanth Chandran, Gaspard Zoss, Derek Bradley, and Paulo Goto. 2022. Morf: Morphable radiance fields for multiview neural head modeling. In *SIGGRAPH 2022*.
- [49] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. 2021. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. 10039–10049.
- [50] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. 2021. Towards real-world blind face restoration with generative facial prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 9168–9178.
- [51] Yaohui Wang, Di Yang, Francois Bremond, and Antitza Dantcheva. 2022. Latent Image Animator: Learning to Animate Images via Latent Space Navigation. *arXiv preprint arXiv:2203.09043* (2022).
- [52] Ziyang Wang, Timur Bagautdinov, Stephen Lombardi, Tomas Simon, Jason Saragih, Jessica Hodgins, and Michael Zollhofer. 2021. Learning compositional radiance fields of dynamic human heads. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5704–5713.
- [53] Yue Wu, Yu Deng, Jiaolong Yang, Fangyun Wei, Qifeng Chen, and Xin Tong. 2022. Anifacegan: Animatable 3d-aware face image generation for video avatars. *arXiv preprint arXiv:2210.06465* (2022).
- [54] Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. 2021. Space-time neural irradiance fields for free-viewpoint video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 9421–9431.
- [55] Jingbo Xing, Menghan Xia, Yuechen Zhang, Xiaodong Cun, Jue Wang, and Tien-Tsin Wong. 2023. CodeTalker: Speech-Driven 3D Facial Animation with Discrete Motion Prior. *arXiv preprint arXiv:2301.02379* (2023).
- [56] Fei Yin, Yong Zhang, Xiaodong Cun, Mingdeng Cao, Yanbo Fan, Xuan Wang, Qingyan Bai, Baoyuan Wu, Jue Wang, and Yujiu Yang. 2022. Styleheat: One-shot high-resolution editable talking face generation via pretrained stylegan. *ECCV* (2022).
- [57] Fei Yin, Yong Zhang, Xuan Wang, Tengfei Wang, Xiaoyu Li, Yuan Gong, Yanbo Fan, Xiaodong Cun, Ying Shan, Cengiz Oztireli, et al. 2022. 3D GAN Inversion with Facial Symmetry Prior. *arXiv preprint arXiv:2211.16927* (2022).
- [58] Jingbo Zhang, Xiaoyu Li, Ziyu Wan, Can Wang, and Jing Liao. 2022. Fdnerf: Few-shot dynamic neural radiance fields for face reconstruction and expression editing. In *SIGGRAPH Asia 2022*.
- [59] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 586–595.
- [60] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. 2021. Flow-Guided One-Shot Talking Face Generation With a High-Resolution Audio-Visual Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 3661–3670.
- [61] Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C Bühlér, Xu Chen, Michael J Black, and Otmar Hilliges. 2022. Im avatar: Implicit morphable head avatars from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 13545–13555.
- [62] Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. 2022. CelebV-HQ: A large-scale video facial attributes dataset. In *European Conference on Computer Vision (ECCV)*. 650–667.
- [63] Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. 2022. CelebV-HQ: A Large-Scale Video Facial Attributes Dataset. In *ECCV*.
- [64] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. 2020. In-domain GAN Inversion for Real Image Editing. In *Proceedings of European Conference on Computer Vision (ECCV)*.

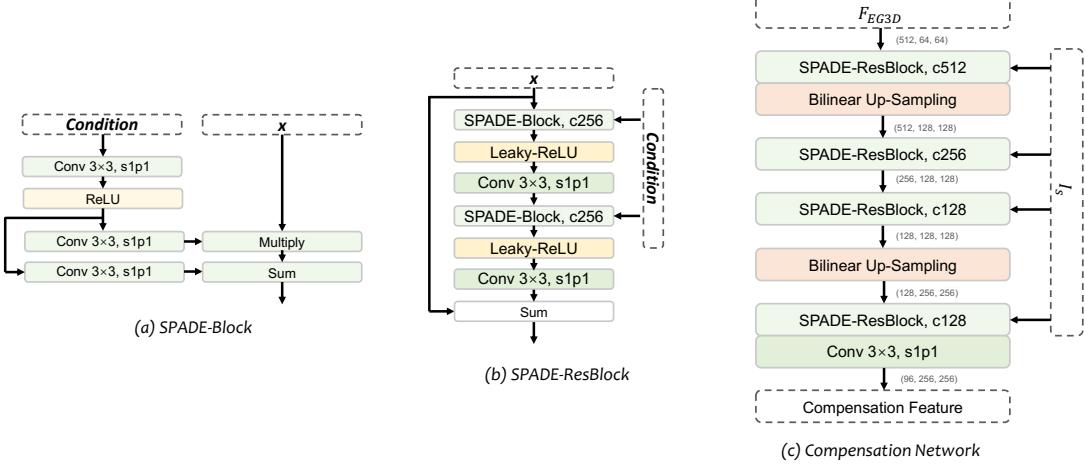


Fig. 9. Architecture of the compensation network, which comprises multiple blocks of spatially-adaptive de-normalization (SPADE) and convolution. It takes the intermediate feature of tri-plane generator G and source image I_s as input and outputs the compensation volume for supplementing identity and texture information. In SPADE, the intermediate features are progressively modulated with a set of scale and shift parameters predicted from I_s . We discard the batch normalization (BN) layers in SPADE to better preserve the information of the intermediate features.