# NOFA++: Tuning-free NeRF-based One-shot Facial Avatar Reconstruction

Wangbo Yu, Chaoran Feng, Yanbo Fan†, Yong Zhang, Xuan Wang, Fei Yin, Yunpeng Bai, Baoyuan Wu, Yan-Pei Cao, Li Yuan†, and Yonghong Tian†, *Fellow, IEEE*

**Abstract**—3D facial avatar reconstruction, a crucial topic in computer graphics and computer vision, has achieved significant advancements due to the development of neural radiance fields (NeRFs). However, most NeRF-based 3D facial avatars primarily focus on subject-specific reconstruction, necessitating numerous multi-view training images with various expressions, and the learned model cannot generalize to new subjects, limiting their wider applicability. To address these challenges, we propose a generalizable one-shot 3D facial avatar reconstruction framework capable of reconstructing high-fidelity 3D facial avatar from a single source image. To overcome the challenges in obtaining generalization ability and lacking multi-view supervision, we employ the generative prior of pretrained 3D GAN and develop an efficient encoder-generator pipeline to reconstruct the canonical neural volume of the source image, then further leverage a coarse-to-fine generation strategy to synthesize canonical volume with image-specific details, enabling reconstructing high-fidelity facial avatars without the need for test-time finetuning. To facilitate fine-grained control over facial dynamics, we incorporate a deformation field to warp the canonical volume into target expressions, enabling accurate motion modeling that can be controlled by both video and audio signals. Through comprehensive experiments, we demonstrate that our method achieves superior reconstruction and reenactment results compared to state-of-the-art methods. Please watch our **Demo video** for a more comprehensive experience.

**Index Terms**—3D facial avatar reconstruction, 3D morphable models, Neural radiance field, 3D-aware GAN.

## 1 Introduction

Facial avatar reconstruction holds a pivotal position in the realms of computer graphics and computer vision, given its exceptional applications in virtual reality (VR), augmented reality (AR), the film industry, and teleconferencing. The ability to produce high-fidelity face reconstructions and achieve fine-grained face reenactment is crucial in unlocking the vast potential of these innovative domains.

To reconstruct facial avatars, several 2D-based methods [1], [2], [3], [4], [5], [6], [7] have been proposed. These methods utilize flow-based warping in image or feature spaces for motion transfer, along with encoder-decoder networks for synthesizing appearance. Through training on large-scale face video datasets [3], [8], [9] containing a large number of identities, these methods can generalize to new identities and produce vivid reenactment results even provided with just a single source image. However, 2D-based methods lack constraints on 3D facial geometry and struggle to generate multi-view consistent images, making them prone to artifacts when handling large driving poses or expressions. Conventional parametric face model [10], [11] -based methods [12], [13], [14], [15], [16], [17] model 3D faces with template mesh and 3DMM parameters [10]. While they support flexible control over poses and expressions, these mesh-based methods suffer from inferior texture quality, memory inefficiency, and are less effective in modeling non-face regions, such as teeth, hair, and accessories. Recently, the photo-realistic and multi-view consistent rendering ability of Neural Radiance Fields (NeRFs) [18] have led to the development of NeRF-based facial avatar reconstruction [19], [20], [21]. These works achieve facial reenactment by learning deformation fields or rendering functions conditioned on control signals, with the pose and expression parameters of 3DMMs [10] being the most commonly used conditions. Although achieves outstanding visual quality and view consistency, NeRF-based methods suffer from two limitations: First, they necessitate a substantial number of images featuring various poses and expressions of the target face for training, which may not always be accessible in real-world scenarios; Second, they are subject-dependent and 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 for extensive multi-view training data limit their wider applicability.

In this work, our objective is to develop a NeRF-based, generalizable, one-shot facial avatar reconstruction method that requires only a single source image and can seamlessly adapt to new identities without test-time finetuning. This goal presents several formidable challenges: **1)** Accurately reconstructing a 3D face from a single source image is difficult due to the inherent complexity of facial distributions and the absence of multi-view information; **2)** Endowing a personalized NeRF with the ability to generalize to different identities is a demanding task; **3)** Precise control over facial motion in NeRF remains elusive, particularly in one-shot scenarios.

To overcome the aforementioned challenges, we devised several innovative solutions. First, we capitalize on the generative prior of a NeRF-based 3D GAN [22] to facilitate

• † *denotes corresponding authors. Wangbo Yu, Chaoran Feng, Li Yuan and Yonghong Tian are with Peking University and PengCheng Laboratory. Yanbo Fan, Xuan Wang are with Ant Research. Yong Zhang is with Tencent AI Lab. Yan-Pei Cao is with VAST. Fei Yin is with University of Cambridge. Yunpeng Bai is with UT Austin. Baoyuan Wu is with CUHK-Shenzhen.*

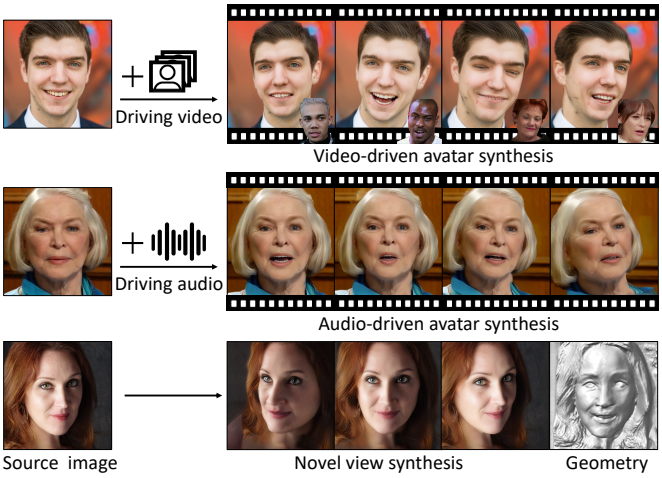



Fig. 1. We propose a NeRF-based, generalizable one-shot facial avatar reconstruction method, which enables high-fidelity facial avatar reconstruction and reenactment given a single source image, and supports both video and audio as control signals.

3D reconstruction from a single image. The 3D GAN's latent space encodes a robust 3D-consistent generative prior, significantly contributing to the synthesis of diverse neural volumes for human faces. To align this latent space with real images, we develop an efficient vision transformer [23]-based image encoder that takes real images as input and produces a set of latent codes, which are then fed into the 3D GAN's generator for neural volume synthesis. This encoder-generator pipeline empowers our method to achieve 3D reconstruction from just a single source image. Nevertheless, due to the compression loss caused by the image encoder, the reconstructed volume exhibits poor identity preservation and low texture quality. To supplement the loss of detailed information, we further leverage a coarse-to-fine generation strategy and learns a detail network. This detail network processes the residual between the source image and its coarse reconstruction, generating a residual feature rich in intricate information, which is then fused with the coarse feature of the 3D GAN generator, steering a refined neural volume generation process. Second, in pursuit of accurate motion modeling, unlike conventional GAN inversion methods [24], [25], [26], [27] that faithfully reconstruct the source image, we project the input image, encompassing any potential expressions, into a shared canonical space featuring an aligned expression. Following this, we learn a deformation field to model facial dynamics, which establishes a conditional mapping that transforms each sampled point in the target expression space into the canonical space based on 3DMM [10] driving signals. Distinct from personalized NeRF avatars [19], [21], [28], [29], we endow our deformation field with generalization ability by conditioning it on both identity and expression parameters of 3DMMs, and train it in conjunction with the canonical space on a large-scale video dataset [9]. Consequently, our method can effectively adapt to any input identity and accurately model their facial motions. In addition to addressing the aforementioned challenges, we also venture into various applications of our method. To enhance multi-modal facial motion control, we finetune the deformation field using video-audio data pairs [30], which allows our method to extend to audio-driven avatar synthesis, creating a more immersive experience. Furthermore, beyond modeling real-world faces, we explore the potential of our method in synthesizing virtual avatars, thereby broadening its synthesis capabilities.

In conclusion, our work makes the following contributions: **1)** We present a NeRF-based, generalizable one-shot facial avatar reconstruction method capable of synthesizing high-fidelity 3D facial avatars from a single source image. Our method, once trained, effectively adapts to unseen test images without test-time finetuning, offering greater efficiency and practicality compared to personalized approaches. **2)** Our method supports real-world and virtual avatar creation, and facilitates multi-modal facial motion control, enabling both video-driven and audio-driven 3D facial avatar synthesis. **3)** We conducted comprehensive experiments, comparing our method with both 2D and 3D facial avatar synthesis methods and demonstrate its superior performance in reconstruction and reenactment tasks.

This work extends NOFA [31], which was previously presented in a conference. We have addressed several key limitations of NOFA, and implemented novel applications as well as conducted comprehensive experiments to validate this improved method. First, we significantly improve NOFA's efficiency by employing a vision transformer-based image encoder, a residual-based coarse-to-fine generation strategy, and a synthetic data-assisted progressive training strategy, enabling high-fidelity 3D avatar synthesis without laborious test-time finetuning. Unlike NOFA, which requires time-consuming and GPU-intensive test-time finetuning (approximately 20 minutes on a 24G GPU) for accurate reconstruction, our improved method achieves similar reconstruction quality, and even better performance in large face poses, in a tuning-free manner. This substantial improvement paves the way for real-time applications. Second, NOFA only supports video-driven avatar synthesis. In this improved method, we further endow it with the capability of audio-driven facial avatar synthesis, enabling vivid facial motion generation from audio input. Third, beyond synthesizing real-world facial avatars, we explore application on virtual avatar synthesis through latent space sampling, which can not be achieved by NOFA. Lastly, we carry out extensive experiments to validate our improved method, incorporating additional comparisons with the latest state-of-the-art approaches and more detailed ablation studies.

## 2 Related Work

### 2.1 Neural scene representation

Neural radiance fields (NeRF) [18] represents 3D scenes using MLP-based implicit function and achieve compelling rendering quality in 3D reconstruction tasks. Leveraging 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, conventional NeRF cannot handle dynamic subjects. Several approaches have been devoted to work around this limitation [19], [20], [21], [32], [33], [34], [35], [36], [37], [38]. The solutions can be roughly categorized into two groups: 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 [19], [20], [21], [28], [29], [32], [39], [40], the pose and expression coefficients of 3D Morphable Face Models (3DMMs) [10] are mostly employed as control signals to model facial deformations. These studies focus on subject-dependent reconstruction and are not generalizable to different identities. Additionally, a large set of facial images of a specific identity is required for training. In contrast to above works, we investigate the subject-agnostic problem, where only a single portrait image is available for reconstruction, and propose a generalizable model capable of handling various testing faces.

### 2.2 3D-aware Generative Networks

Inspired by the breakthroughs achieved by 2D Generative Adversarial Networks (GANs) [41], [42], [43], recent researches [22], [44], [45], [46] have extended 2D image generation into 3D settings by combining GANs with Implicit Neural Representations (INRs). These unconditional 3D GANs can generate photorealistic renderings and enable control over views. However, they do not support fine-grained and explicit expression controls. Recently, [47] 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 finetuning 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). Several concurrent works [48], [49], [50], [51] leverage the parameters of 3D Morphable Face Models (3DMMs) to explicitly control the expression of rendered faces. However, these models are specifically designed for unconditionally generating virtual avatars and cannot be readily applied in real-world applications. In contrast to these methods, we utilize 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 serve as a bridge to bring GANs into real-world applications, such as image editing and reenactment. Existing GAN inversion approaches can be roughly divided into three categories: The optimization-based methods [52], [53] 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 [24] [25], 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 [54], the generator parameters are also optimized in [26] to achieve better results. Additionally, [27] proposes a residual-based compensation network to assist the GAN inversion process, achieving expressive reconstruction and editing results. Meanwhile, the combination of GAN inversion techniques and 3D GANs has also brought hope for single-image to 3D reconstruction. Several works explored optimization-based inversion with view regularization for 3D reconstruction [55]. There have also emerged learning-based methods [31], [56], [57], [58] that achieve 3D face reconstruction using a learned encoder.

### 2.4 One-shot Talking Head Synthesis

One-shot talking head synthesis aims to generate talking face videos from a source image and driving signals (typically video or audio signals). The generated videos should maintain the facial characteristics of the source image and exhibit reasonable facial motion in response to the driving signal. 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 [2], [5] predict warping flow from key-points to warp features of source images into target motion. The work of [59] uses 3DMM parameters to modulate flow generator and a refine network to supplement fine details. [1] further leverages the prior of StyleGAN [41] to enhance appearance. These methods are trained on the large-scale face video datasets [3], [8], [9], [60] 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 [3], [6] 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. There are also 3D-based methods that reconstruct 3D facial avatar from a single image [61], [62]. These methods are resolution-limited, prone to generating overly blurry textures, and cannot model backgrounds.

## 3 METHOD

We propose NOFA++, a NeRF-based one-shot facial avatar reconstruction framework, which will be described in detail in this section. First, in Section 3.1, we introduce an efficient encoder-generator network tailored for coarse 3D volume synthesis from a single image. Next, we illustrate a residual-based coarse-to-fine generation strategy in Section 3.2, which enables fine volume synthesis with image-specific details. Subsequently, we introduce the architecture of the deformation field designed for accurate facial motion modeling in Section 3.3. In Section 3.4, we present a synthetic data-assisted progressive training strategy that utilizes both real and synthetic data for effective model training.

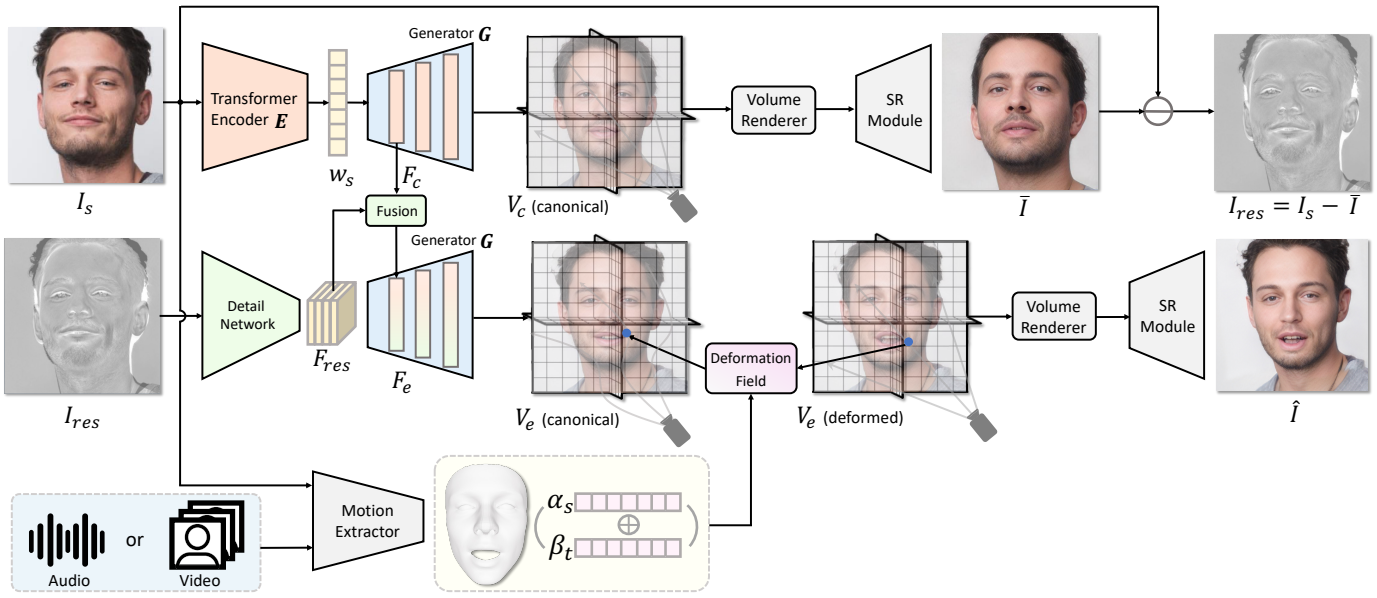



Fig. 2. **Inference pipeline of NOFA++.** We adopt a coarse-to-fine strategy to generate high-fidelity facial avatar. In the coarse stage, given a source image $I_s$, a transformer encoder $E$ is adopted to encode the image into the latent space of the tri-plane generator $G$, which will produce a coarse canonical volume $V_c$ with an aligned expression, while preserve the identity of $I_s$. Then, a coarse image $\bar{I}$ can be rendered from $V_c$. In the refine stage, a detail network is employed to accept the images residual $I_{res} = I_s - \bar{I}$ as input, and generate a residual $F_{res}$ to supplement image-specific details for the coarse tri-plane feature $F_c$, forming a refined feature $F_e$ for fine volume synthesis. To achieve explicit motion control, we employ a 3DMM extractor $M$ to extract 3DMM parameters from driving signals and source image $I_s$, then use the combination of source identity and target expression parameters as control signal of the deformation field, which will deform the reconstructed fine canonical volume $V_e$ into target expression.

### 3.1 Learning coarse volume synthesis with generative prior

To reconstruct high-fidelity facial avatars, the first and most crucial step is to construct a 3D representation of the given subject. As shown in Fig. 2, we leverage the generative prior of a pretrained 3D GAN [22] to synthesize 3D representation of the input image. It consists of an image SR module, a volume renderer, and a tri-plane generator $\boldsymbol{G}$. The tri-plane generator $\boldsymbol{G}$ employs a StyleGAN2 [42] backbone to synthesize features of size $96 \times 256 \times 256$, which will be reshaped into a tri-plane of size $3 \times 32 \times 256 \times 256$. Subsequently, the features at each coordinate in the tri-plane can be efficiently queried using coordinates and decoded into neural volumes for volumetric rendering, resulting in view-consistent images.

The 3D GAN was originally designed to sample latent codes from its latent space and synthesize virtual faces. To align the unconditional latent space with real images, we draw inspirations from GAN inversion approaches [24], [25], [27] and learn a vision transformer [23]-based encoder to achieve image-to-volume synthesis. Specifically, given a source image $I_s$, we train an encoder $\boldsymbol{E}$ to project the source image into the latent space of the tri-plane generator $\boldsymbol{G}$, *i.e.*, encoding $I_s$ into a set of latent codes $w_s$ that will modulate the intermediate features of $\boldsymbol{G}$ for volume generation. Different from NOFA [31] that employed a CNN-based encoder for image encoding, in this improved method, we alternate to train a more effective transformer network [23] to improve the quality of image-to-volume synthesis. The volume synthesis process is formulated as:

$$V_c = \boldsymbol{G}(\boldsymbol{E}(I_s) + \overline{w}), \tag{1}$$

where $V_c$ denotes the coarse neural volume produced by $\boldsymbol{G}$, and $\overline{w}$ is the pre-computed average latent code of $\boldsymbol{G}$, which remains fixed during training. It is worth noting that, unlike conventional GAN inversion approaches that faithfully reconstruct the input image, the coarse volume $V_c$ produced by our method is defined in the canonical space, *i.e.*, with an aligned expression instead of preserving the original expression of $I_s$. In our method, learning a canonical space is critical for modeling facial dynamics, as it enables the application of backward deformation for accurate motion control. The canonical space is naturally learned by jointly training the entire framework in an end-to-end manner, without explicit internal supervision. Visualization of the learned canonical expression is shown in Fig. 8.

### 3.2 Learning fine volume synthesis with detail network

Although the vision transformer-based encoder facilitates a more accurate volume synthesis than CNN-based encoder adopted in NOFA [31], it still faces challenges in recovering fine details of the input image, since encoding the input image into the extremely low-rate latent codes will inevitably result in information loss [63]. To alleviate this issue, NOFA [31] adopts a compensation network to take the intermediate feature of the tri-plane generator $\boldsymbol{G}$ as input and produce a compensation volume, which is directly added with the simultaneously generated coarse volume to supplement detailed information. Nevertheless, experiments demonstrate that this strategy lacks sufficiency

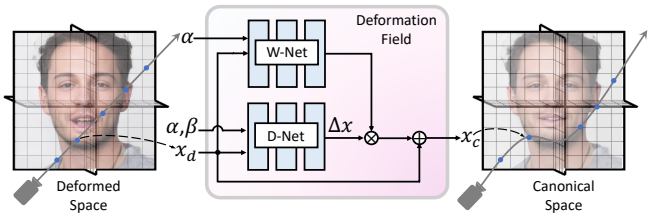



Fig. 3. **Illustration of the deformation field.** The deformation field in our method models the backward deformation, transforming 3D points in the deformed space back to the canonical space. It comprises two key components: a deformation network (D-Net) and a weighting network (W-Net). The D-Net predicts coordinate offsets between $x_c$ and $x_d$, by taking the concatenation of positional embedded $x_d$ and $(\alpha, \beta)$ as input, predicts a 3D offset $\Delta x$. Furthermore, we train a W-Net that accepts the concatenation of positional embedded $x_d$ and $\alpha$ as input and predicts the offset weights to be multiplied with $\Delta x$, enabling more accurate motion modeling. Finally, $x_c$ is derived by adding the weighted $\Delta x$ to $x_d$.

in faithfully supplementing information loss to achieve high fidelity reconstruction. Thus, NOFA [31] still requires time-consuming and GPU-intensive test time finetuning to achieve more accurate reconstruction, which impacted its broader applications.

Different from NOFA [31], we explore a more effective coarse-to-fine generation strategy, which enables the generation of high-fidelity 3D avatars without the need for laborious test-time finetuning. As shown in Fig. 2, after generating the coarse volume, we learn a detail network to take the residual [27] of the source image $I_s$ and its coarse canonical reconstruction $\overline{I}$ as input and produce a residual feature that contains intricate details of the source image, expressed as:

$$F_{res} = \boldsymbol{C}(I_{res}), \tag{2}$$

where $I_{res} = I_s - \overline{I}$ denotes the residual image. The output residual feature $F_{res}$ is then applied to modulate the coarse tri-plane feature $F_c$ through a feature fusion network to supplement the image-specific details, expressed as:

$$F_e = \gamma(F_{res}) \odot F_c + \eta(F_{res}), \tag{3}$$

where $\gamma$ and $\eta$ are two light-weight CNNs, and $F_e$ denotes the refined tri-plane feature. Through feeding the refined feature $F_e$ to the generator $\boldsymbol{G}$ and repeating the volume generation process, the final generated fine canonical volume $V_e$ will contain more details of the source image, enabling high-fidelity 3D reconstruction.

### 3.3 Dynamic modeling with 3DMM guidance

In order to achieve explicit motion control on the reconstructed neural volume, we exploit a deformation field to model facial dynamics and utilize the semantic parameters of a 3DMM face model [11] as control signals. In 3DMM, the face shape is defined as:

$$\mathbf{S} = \bar{\mathbf{S}} + \alpha \mathbf{B_{id}} + \beta \mathbf{B_{exp}}, \tag{4}$$

where $\bar{\mathbf{S}}$ represents the average face shape, $\mathbf{B_{id}}$ and $\mathbf{B_{exp}}$ are the identity and expression basis computed by PCA [64]. We adopt the coefficients $\alpha$ and $\beta$ as control signals, which are semantically meaningful and enable fine-detailed expression control.

For video-driven avatar synthesis, let $I_s$ represent the source image, and $I_t$ represent the target image with desired expression, we employ an off-the-shelf 3D face reconstruction model [12] to estimate source identity parameter $\alpha_s$ from $I_s$, and derive target expression parameter $\beta_t$ from $I_t$. We use a concatenation of $(\alpha_s, \beta_t)$ as the control signal of the deformation field, since we found that additionally use source identity $\alpha_s$ as condition helps to preserve the identity of the deformed volume, which is crucial for endowing the deformation field with generalization ability.

In particular, the deformation field models the backward deformation that deforms 3D points in the deformed space to the canonical space. As shown in Fig. 3, to render facial image with a specific expression $(\alpha, \beta)$, we shoot straight rays in the deformed space and sample points along the rays. For each sampled point $x_d$ in the deformed space, we use the deformation field to predict its corresponding canonical coordinate $x_c$ and query the tri-plane feature at $x_c$, which is then used to regress the density and color at $x_d$ for neural rendering. The deformation field consists of a deformation network (D-Net) and a weighting network (W-Net). We employ D-Net to predict coordinate offsets between $x_c$ and $x_d$, it takes the concatenation of positional embedded $x_d$ and $(\alpha, \beta)$ as input and predicts a 3D offset $\Delta x$. Drawing inspiration from the FLAME mesh model [65] which assigns skinning weights on mesh vertices for smooth blending, we additionally train a W-Net, which takes the concatenation of positional embedded $x_d$ and $\alpha$ as input and predicts the offset weights to be multiplied with $\Delta x$, enabling more accurate motion modeling. Finally, $x_c$ is derived by adding the weighted $\Delta x$ to $x_d$.

Following the deformation field, a volume renderer is utilized to perform volume rendering along the deformed bent rays, generating an observation in the deformed space. Subsequently, we use an image SR module inherited from the 3D GAN [22] to produce the final output $\hat{I}$.

### 3.4 Synthetic data-assisted progressive training

**Reconstruction training with synthetic data.** During this training stage, we first fix the pretrained generator $G$ and train the transformer-based image encoder $E$ to accomplish image to coarse volume synthesis. The training is carried out using synthetic 3D data generated by the adopted 3D GAN [22]. At each training iteration, we randomly sample a volume from its latent space, and sample two camera poses $P_s$ and $P_t$ to render a set of training images with the same identity, involving high-resolution image pair $I_s$ and $I_t$, raw image pair $I_s^r$ and $I_t^r$ generated before the SR module, and pair of rendered depth map $I_s^d$, $I_t^d$. Subsequently, we use $I_s$ as the input of the image encoder, employing $I_s$, $I_t$ and $I_s^r$, $I_t^r$ as multi-view supervision. We additionally utilize depth map pair $I_s^d$, $I_t^d$ for geometry regularization.

The loss function for training the image encoder is defined as:

$$\mathcal{L}_{\text{enc}} = \lambda_{\text{LPIPS}}\mathcal{L}_{\text{LPIPS}} + \lambda_{\text{L2}}\mathcal{L}_{\text{L2}} + \lambda_{\text{depth}}\mathcal{L}_{\text{depth}}, \tag{5}$$

where $\lambda_{\text{LPIPS}}$, $\lambda_{\text{L2}}$ and $\lambda_{\text{depth}}$ are set to 1.0, 1.0 and 2.0 respectively. Here, the perceptual loss [66] $\mathcal{L}_{\text{LPIPS}}$ is computed as:

$$\begin{aligned}\mathcal{L}_{\text{LPIPS}} = \text{LPIPS}(I_s, \overline{I}_s) + \text{LPIPS}(I_t, \overline{I}_t) + \\ \text{LPIPS}(I_s^r, \overline{I}_s^r) + \text{LPIPS}(I_t^r, \overline{I}_t^r),\end{aligned} \tag{6}$$

where $\overline{I}_s$, $\overline{I}_t$, $\overline{I}_s^r$ and $\overline{I}_t^r$ are rendered from the reconstructed coarse volume $V_c$. The L2 loss $\mathcal{L}_{\text{L2}}$ is computed as:

$$\mathcal{L}_{\text{L2}} = ||I_s - \overline{I}_s||_2 + ||I_t - \overline{I}_t||_2 + ||I_s^r - \overline{I}_s^r||_2 + ||I_t^r - \overline{I}_t^r||_2. \tag{7}$$

We also apply a depth loss $\mathcal{L}_{\text{depth}}$ on the reconstructed depth pair $\overline{I}_s^d$, $\overline{I}_t^d$ and ground truth depths to regularize the reconstructed geometry, expressed as:

$$\mathcal{L}_{\text{depth}} = ||I_s^d - \overline{I}_s^d||_2 + ||I_t^d - \overline{I}_t^d||_2. \tag{8}$$

After training the image encoder $E$, we fix both it and the generator $G$, and proceed to train the detail network along with the feature fusion network for fine volume synthesis. The utilized loss functions and hyperparameters are the same as those used for training the image encoder, expressed as:

$$\mathcal{L}_{\text{detail}} = \lambda_{\text{LPIPS}}\mathcal{L}_{\text{LPIPS}} + \lambda_{\text{L2}}\mathcal{L}_{\text{L2}} + \lambda_{\text{depth}}\mathcal{L}_{\text{depth}}, \tag{9}$$

where we compute the above loss terms on the reconstructed fine volume $V_e$, using final upsampled images $\hat{I}_s$ and $\hat{I}_t$, raw images $\hat{I}_s^r$ and $\hat{I}_t^r$, as well as the rendered depth $\hat{I}_s^d$, $\hat{I}_t^d$.

**Deformation training with large-scale video data.** In this training stage, we jointly train the entire model, including the deformation field, on a large-scale video dataset [9]. At each iteration, we sample a source image $I_s$ and a target image $I_t$ from the same video clip, using $I_s$ as model input and $I_t$ as the supervision. As mentioned before, we use the 3DMM parameter $(\alpha_s, \beta_t)$ estimated from $I_s$ and $I_t$ as control signal of the deformation field. We use multiple objectives to ensure accurate motion modeling. Firstly, we apply a reconstruction loss consisting of perceptual loss [66] and L2 loss between the synthetic image $\hat{I}$ and the target image $I_t$, expressed as:

$$\mathcal{L}_{\text{rec}} = ||I_t - \hat{I}||_2 + \text{LPIPS}(I_t, \hat{I}). \tag{10}$$

Additionally, we employ a local reconstruction loss to further enhance the critical mouth and eye regions. To accomplish this, we estimate facial landmarks using MTCNN [67] to create bounding boxes and crop the mouth and eye regions from $I_t$ and $\hat{I}$. We then apply the reconstruction loss to the cropped region, formulated as:

$$\mathcal{L}_{\text{bbox}} = ||bbox(I_t) - bbox(\hat{I})||_2 + \text{LPIPS}(bbox(I_t), bbox(\hat{I})). \tag{11}$$

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

$$\mathcal{L}_{\text{id}} = 1 - \left\langle F(I_t), F(\hat{I}) \right\rangle, \tag{12}$$

where $F(\cdot)$ is the pretrained ArcFace [68], $\langle \cdot, \cdot \rangle$ denotes cosine distance.

The total objective of this training stage is defined as:

$$\mathcal{L}_{\text{dfm}} = \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{bbox}}\mathcal{L}_{\text{bbox}} + \lambda_{\text{id}}\mathcal{L}_{\text{id}}, \tag{13}$$

where $\lambda_{\text{rec}}, \lambda_{\text{bbox}}, \lambda_{\text{id}}$ are set as 1.0, 1.0 and 0.1 respectively.

**Refinement training with hybrid data.** Since the video dataset [69] has relatively lower texture quality compared to the synthetic 3D data used in the reconstruction training stage, training on it would lead to a decline in visual quality. To address this issue, we incorporate an additional refinement training stage that utilizes both the synthetic 3D data and video data to enhance visual quality. During this stage, we alternate between training on the synthetic dataset and the video dataset. Specifically, we conduct one iteration on the synthetic dataset after every two iterations on the video dataset. At the synthetic iteration, we train the entire model using the same loss functions and hyperparameters as defined in Eq. 9. At the video iteration, we fix other modules and only train the deformation field, employing the loss functions in Eq. 11 and Eq. 12, with the hyperparameters set as $\lambda_{\text{bbox}} = 2.0$ and $\lambda_{\text{id}} = 1.0$.

**Audio-video joint training.** As our model utilizes 3DMM parameters as control signals for modeling facial dynamics, ideally, we can employ off-the-shelf audio-to-3DMM estimators to accomplish audio-driven motion modeling. Nonetheless, we found that this naive approach yields inferior results, characterized by inaccurate and flickering motion, due to the inevitable domain gap when transferring video to audio signals. To address this issue, we finetune the deformation field on paired audio-video data [30]. Specifically, for data preparing, given an audio-video pair, we extract 3DMM expression parameter $\beta_{aud}$ from audio signals using [70], and derive 3DMM identity parameter $\alpha_{vid}$ and camera poses from its video counterpart using [12]. Then, the finetuning is conducted in the same manner as training on video data. We sample a source frame and a target frame from the same video clip, utilizing the source frame as the model input and the target frame as the supervision. The input to the deformation field is a combination of $(\alpha_{vid}, \beta_{aud})$. During finetuning, we fix other modules and only train the deformation field, using the local reconstruction loss (Eq. 11) on the mouth region and the identity loss (Eq. 12) to preserve identity. The hyperparameters are set as $\lambda_{\text{bbox}} = 1.0$ and $\lambda_{\text{id}} = 0.2$.

**Fast adaptation on challenging cases.** Similar to existing tuning-free methods [1], [59], [61], [62], our method can well handle common faces with an tuning-free manner, but encounters artifacts when dealing with challenging cases, such as out-of-domain faces and faces with heavy make-ups. To address this issue, we explore a fast adaptation strategy that requires optimizing only a very limited number of parameters within a short time, enabling our model to handle challenging cases. Specifically, given a challenging case $I_s$, we use it as both the source image and the driving image to get its reconstruction $\hat{I}$ using our model, then directly optimize the refined tri-plane feature $F_e$ in Eq. 3 to make $\hat{I}$ match $I_s$. We apply the reconstruction loss (Eq. 10) between $I_s$ and the generated $\hat{I}$, and additionally use a depth regularization loss $\mathcal{L}_{\text{reg}} = ||\hat{I}_{opt}^d - \hat{I}_{reg}^d||_2$ for geometry regularization, where $\hat{I}_{reg}^d$ represents the rendered depth obtained from the original tri-plane feature $F_e$, $\hat{I}_{opt}^d$ denotes the depth generated from the feature during optimization. The total loss function is formulated as:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}, \tag{14}$$

where we set $\lambda_{\text{rec}} = 1$ and $\lambda_{\text{reg}} = 1$. After this adaption process, our model is capable of faithfully reconstructing challenging faces using the optimized tri-plane feature, examples are shown in Fig. 10.

TABLE 1
**Quantitative comparison on self-reenactment and cross-reenactment.** The results indicate that our method achieves the best reconstruction quality in most metrics, as well as the best pose accuracy, and attains a comparable motion accuracy compared with 2D-based methods.

| Methods | Self-reenactment | | | | | | | Cross-reenactment | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | FID↓ | LPIPS↓ | PSNR ↑ | SSIM ↑ | CSIM↑ | AED↓ | APD↓ | FID↓ | CSIM↑ | AED↓ | APD↓ |
| Face-vid2vid [3] | 36.75 | 0.204 | 20.22 | 0.651 | 0.768 | **0.129** | 0.024 | 49.98 | 0.603 | **0.210** | 0.035 |
| PIRenderer [59] | 43.90 | 0.238 | 20.69 | 0.633 | 0.712 | 0.132 | 0.038 | 56.39 | 0.482 | 0.224 | 0.042 |
| StyleHEAT [1] | 46.32 | 0.227 | 21.80 | 0.638 | 0.701 | 0.146 | 0.031 | 51.20 | 0.477 | 0.237 | 0.045 |
| ROME [61] | 50.55 | 0.243 | 17.44 | 0.614 | 0.726 | 0.155 | 0.025 | 66.01 | 0.548 | 0.259 | 0.029 |
| HideNeRF [62] | 54.47 | 0.251 | 18.17 | 0.609 | 0.720 | 0.159 | 0.023 | 62.14 | 0.530 | 0.308 | 0.031 |
| Real3D [71] | 38.93 | 0.215 | 21.07 | 0.643 | 0.765 | 0.151 | 0.029 | 51.77 | 0.582 | 0.215 | 0.033 |
| VOODOO3D [72] | 42.19 | 0.219 | 20.85 | 0.647 | 0.733 | 0.152 | 0.023 | 53.84 | 0.490 | 0.301 | 0.030 |
| NOFA [31] | 32.08 | **0.159** | 21.55 | 0.668 | 0.785 | 0.142 | 0.021 | **42.38** | 0.665 | 0.256 | 0.024 |
| Ours | **29.91** | 0.176 | **21.83** | **0.672** | **0.789** | 0.135 | **0.021** | 43.75 | **0.691** | 0.234 | **0.023** |

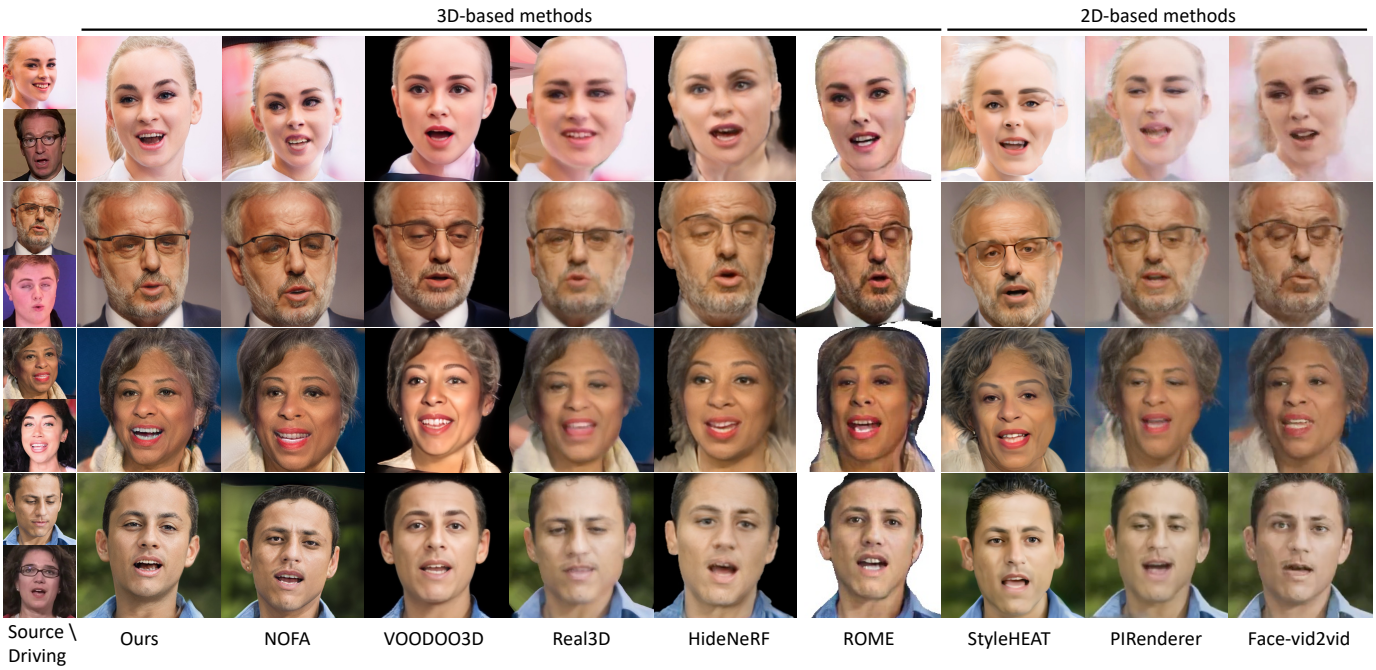



Fig. 4. **Comparison of video-driven avatar synthesis.** We compared our method with both 2D-based and 3D-based methods. The results demonstrate that our method can produce high-fidelity facial avatars with accurate motion.

## 4 Experiments

### 4.1 Implementation details

**Dataset and preprocessing.** We train our model using synthetic 3D data produced by EG3D [22], a filtered CelebV-HQ video dataset [69] containing 20000 video clips, and a subset of Voxceleb [30] dataset which includes 1500 aligned audio-video pairs. For video preprocessing, we crop and align the videos in the same way with [22], then extract per-frame 3DMM parameters including identity, expression, and camera poses for training.

**Training details.** Our framework is trained across three stages using 4 Nvidia Tesla V100 GPUs. The batch size is set to 4, and we utilize the ADAM optimizer to optimize model weights. In the reconstruction training stage, we first train the image encoder for 1000K iterations, using a learning rate of $10^{-4}$. We then fix the image encoder and proceed to train the detail network and the feature fusion network for 200K iterations, using the same learning rate. In the deformation training stage, we add a deformation field and jointly train the entire model for 1000K iterations, using a smaller learning rate of $2 * 10^{-5}$. In the hybrid data training stage, we train the model for 200K iterations, with the learning rate set as $10^{-4}$. For audio-video joint training, we finetune the deformation field for 200K iterations, using a learning rate of $2 * 10^{-5}$. The entire training process takes approximately 5 days.

**Modeling head rotation and background.** In our implementation, similar to [48], [49], [50], [51], we use camera poses to model head rotation. This helps improve the pose accuracy in video-driven facial avatar generation, but will lead to background movements when applying different poses, thus affecting the overall realism of the rendered video. To address this issue, we adopt a post-processing to handle background movements. Specifically, we mask out the foreground of the input image then use an inpainting model [74] to fill the background. During face rendering,

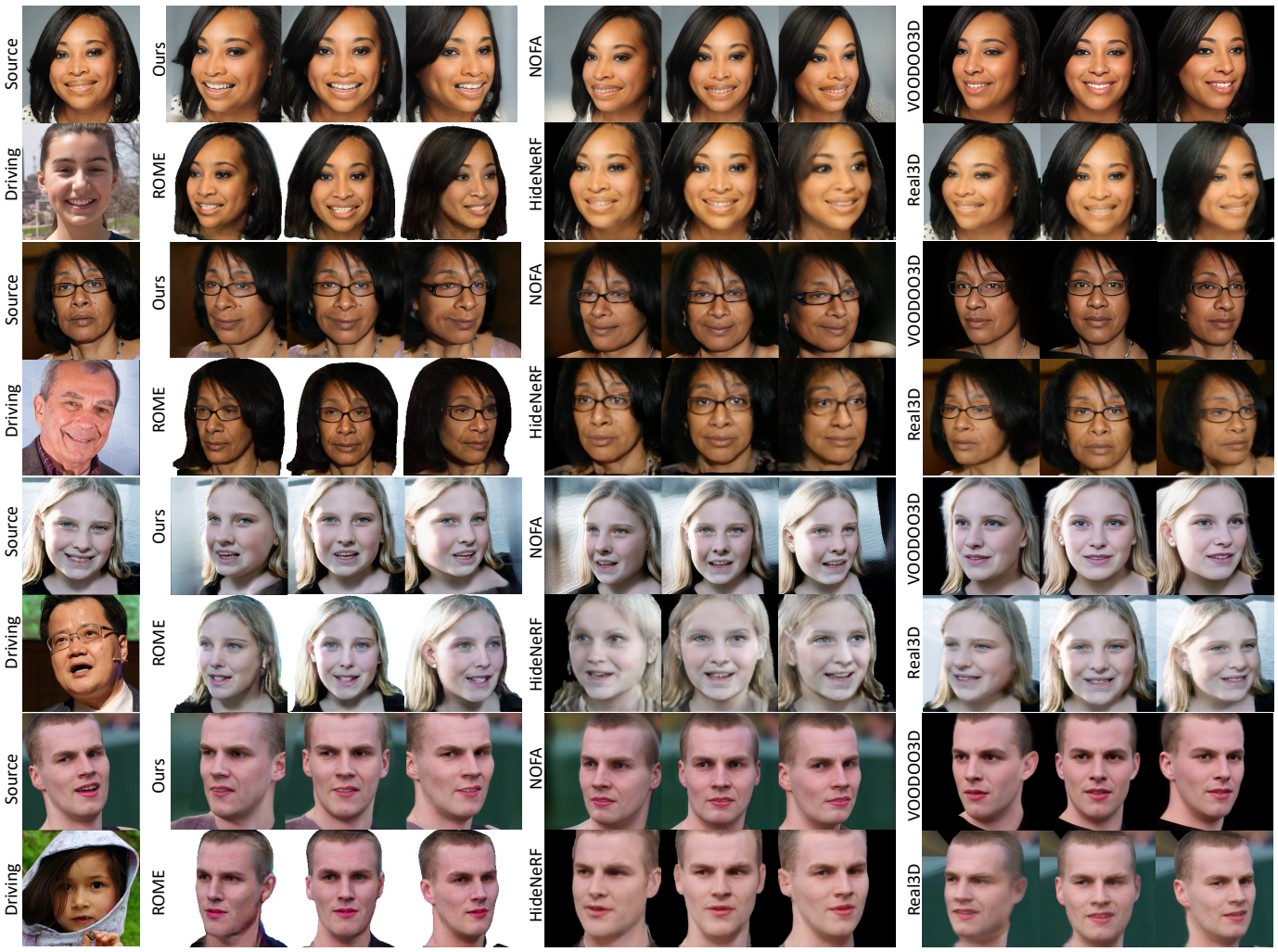



Fig. 5. **Comparison with 3D-based methods on novel view synthesis.** We render three novel views for each method, with their expressions controlled by the driving image. The results show that our method achieves comparable reconstruction quality with the optimization-based NOFA [31], while exhibiting improved motion accuracy, and better reconstruction quality compared to the other baselines.

TABLE 2
**Quantitative comparison of audio-driven avatar synthesis.** Our method achieves better reconstruction quality and pose accuracy than the baselines, and it also achieves a competitive facial motion accuracy compared to the 2D-based method.

| Methods | FID↓ | CSIM↑ | AED↓ | APD↓ |
|---|---|---|---|---|
| SadTalker [73] | 59.86 | 0.645 | **0.224** | 0.047 |
| Real3D [71] | 55.16 | 0.667 | 0.285 | 0.029 |
| Ours | **52.47** | **0.683** | 0.236 | **0.024** |

we utilize a real-time portrait matting model [75] to obtain the foreground of the rendered face and combine it with the inpainted background, finally resulting in a static video background.

**Evaluation metrics.** We utilize several metrics to assess the quality of reconstruction and reenactment. The peak signal-to-noise ratio (PSNR), Structural Similarity (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS) [66] are employed for evaluating image synthesis quality. Furthermore, we employ the Frechet Inception Distance (FID) [76] to measure the distance between synthetic and real image distributions. We calculate the cosine similarity (CSIM) between the source and generated images to evaluate identity preservation. Referring [59], to assess reenactment quality, we extract 3DMM expression and pose parameters from synthetic and real images and compute their Average Expression Distance (AED) and Average Pose Distance (APD).

### 4.2 Comparison of video-driven facial avatar synthesis.

**Baselines and benchmarks.** We evaluate our method by comparing it with several state-of-the-art facial avatar synthesis methods, including the following 2D methods: PIRenderer [59], Face-vid2vid [3], and StyleHEAT [1], as well as the following 3D methods: VOODOO3D [72], Real3D [71], ROME [61], HideNeRF [62] and NOFA [31]. In the evaluation of video-driven face reenactment, we conduct two types of reenactment tasks: self-reenactment and cross-reenactment. In self-reenactment, the identity of the source image is the same as that of the driving frames. In cross-

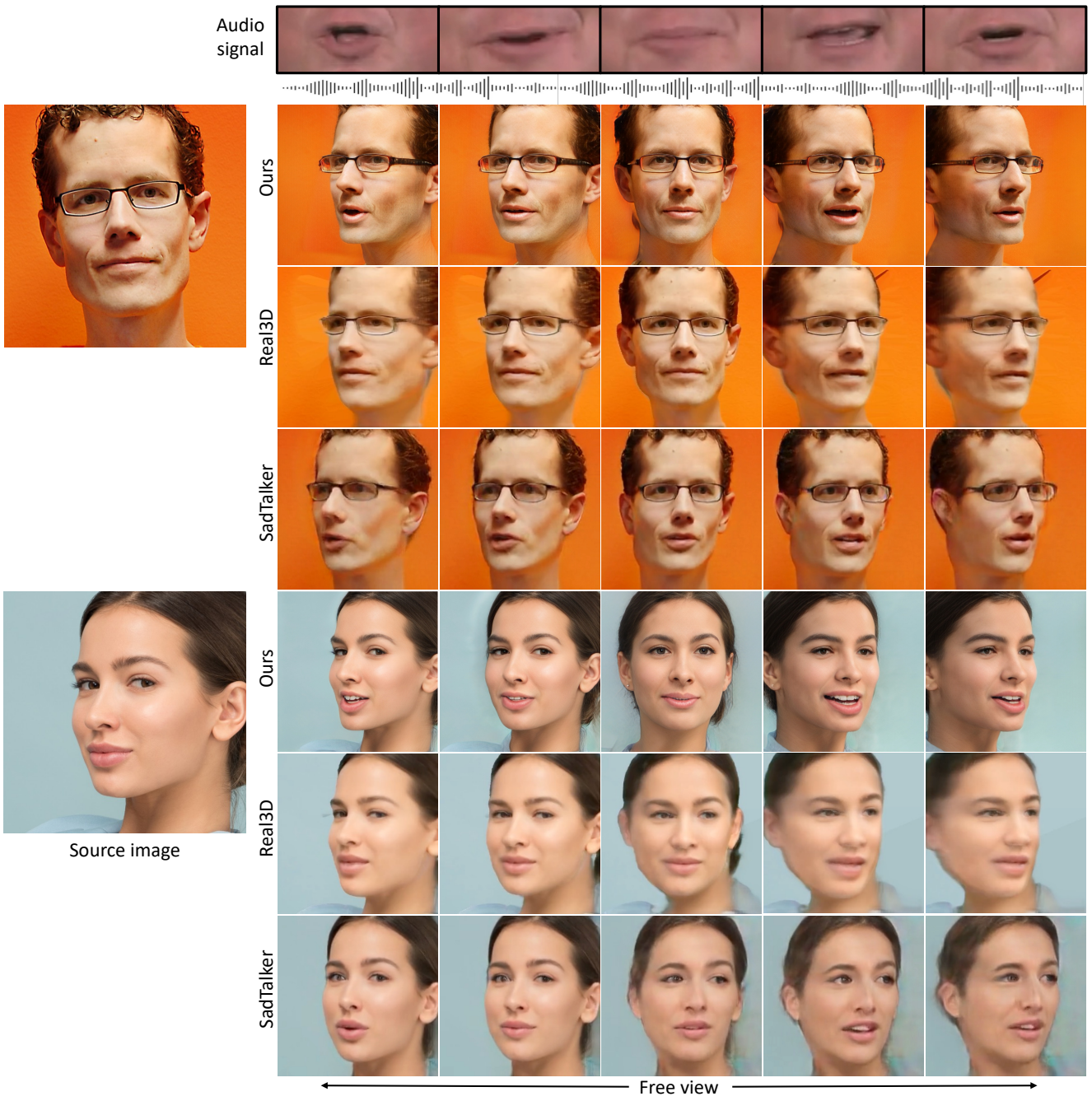



Fig. 6. **Comparison of audio-driven avatar synthesis.** The top row shows the driving audio and the corresponding ground truth lip motion, below shows the generated free views with lip motion controlled by the driving audio. The results demonstrate that our method can produce view-consistent audio-driven facial avatars with accurate lip motion.

reenactment, the source image and driving frames come from two different identities. The latter setting is much more challenging due to the facial feature gap between the source and driving faces. For self-reenactment evaluation, we use 40 video clips comprising a total of 12,000 frames from the HDTF dataset [77] and the VFHQ dataset [78]. For cross-reenactment evaluation, we randomly select 2,000 images from the FFHQ dataset [41] and the LPFF dataset [79] that contains large poses as source images, and use the 40 video clips as driving videos.

**Qualitative results.** Fig. 4 shows the qualitative reenactment results of our method and other state-of-the-art methods, demonstrating that our method achieves better reconstruction quality. Specifically, when reconstructing input faces with large poses and dealing with driving faces with deviated poses, the 2D-based methods fail to infer reasonable frontal faces and suffer from severe artifacts, whereas our method successfully reconstructs accurate and reasonable facial geometry. Compared to 3D-based methods, HideNeRF [62] ROME [61] and Real3D [71] suffer from over-smoothed appearance, fail to maintain fidelity under novel views, and experience inaccuracies in motion as well as

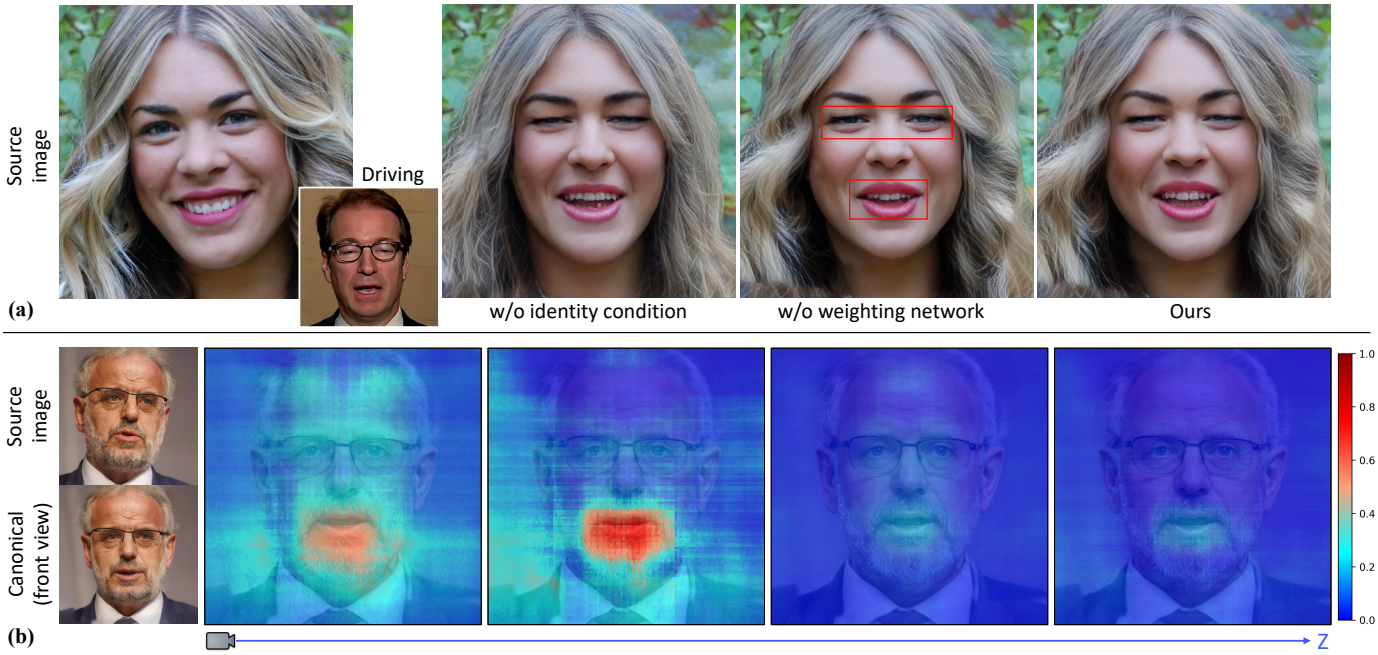



Fig. 7. **Evaluation of the deformation field.** Fig. (a) shows qualitative ablation of the design space of the deformation field. Fig. (b) shows visualization of the weight scalars predicted by the weighting network. Given a source image, we shoot camera rays in the front view onto the reconstructed canonical volume, and sample four 2D coordinate planes along the camera's z-axis. The heat maps show the normalized weights of the sampled positions


Fig. 8. **Visualization of the canonical space and evaluation of the coarse-to-fine generation strategy.** The second row shows rendering results of the coarse canonical volume, the third row shows the rendering results of the fine canonical volume produced by the detail network. With the detail network, the reconstruction quality is significantly improved, exhibiting better texture and identity preservation.

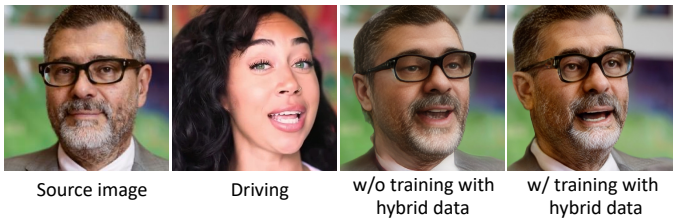



Fig. 9. **Qualitative ablation of training with hybrid data.** It can be found that the reconstruction is significantly improved after refinement training with hybrid data.


Fig. 10. We employ a fast adaptation strategy to address out-of-domain faces and heavy make-ups. The example above shows that direct reconstruct the complex case using our model yields inferior results, while adopting fast adaptation effectively improves the reconstruction quality.

teeth artifacts; VOODOO3D [72] struggles to preserve fidelity of the source image. While NOFA [31] can produce satisfying detailed textures through test-time finetuning, the finetuning process may introduce artifacts and affect motion accuracy. In contrast, our method guarantees fine-grained motion control and high fidelity across views. We also compare our method with 3D-based approaches in terms of novel view synthesis. The results are illustrated in Fig. 5, where we render three novel views for each method, with their expressions controlled by the driving image. The results indicate that our method attains a reconstruction quality comparable to the optimization-based NOFA [31], while exhibiting improved motion accuracy and superior quality compared to the other 3D-based methods.

**Quantitative evaluations.** The quantitative results are listed

TABLE 3
**Quantitative ablation of training with hybrid data.** The reconstruction quality is significantly improved with refinement training with hybrid data.

| Training with hybrid data | FID↓ | LPIPS↓ | PSNR ↑ | SSIM ↑ | CSIM↑ |
|---|---|---|---|---|---|
| w/o | 39.67 | 0.234 | 19.08 | 0.625 | 0.695 |
| w/ | **29.91** | **0.176** | **21.83** | **0.672** | **0.785** |

TABLE 4
**Quantitative evaluation of coarse-to-fine generation.** The reconstruction quality is significantly improved with the help of the detail network in the refine stage.

| Stage | FID↓ | LPIPS↓ | PSNR ↑ | SSIM ↑ | CSIM↑ |
|---|---|---|---|---|---|
| Coarse | 37.44 | 0.213 | 20.25 | 0.649 | 0.703 |
| Fine | **29.91** | **0.176** | **21.83** | **0.672** | **0.785** |

in Table. 1. In terms of reconstruction quality, our method achieves comparable performance to the optimization-based NOFA [31], and surpasses the other strong baselines. In facial motion modeling, the average expression distance (AED) shows that our 3D-based method produces competitive animation results compared to 2D-based methods. Furthermore, for head pose modeling, our method significantly outperforms the baselines in terms of the average pose distance (APD).

**User study.** In order to assess the effectiveness of our proposed method in terms of human perception quality, we conducted a comprehensive user study. Each participant was presented with 20 questions randomly selected from the test data, and asked to identify their most preferred result for each question, taking into consideration both the reconstruction quality and motion quality. The study involved a total of 60 participants, with an analysis illustrated in Fig. 11. On average, our method was favored by 75% of the participants, substantially surpassing the performance of the baselines. This outcome provides compelling evidence in support of the robustness and superior quality inherent in our proposed method.

### 4.3 Comparison of audio-driven facial avatar synthesis.

**Baselines and benchmarks.** For audio-driven facial avatar synthesis comparison, we compare with SadTalker [73] and Real3D [71]. We utilize 100 images from the FFHQ dataset [41] and the LPFF dataset [79] as source images, and 20 audio-video pairs from the HDTF dataset [77] to provide the driving signal. Since the underlying generated head poses can be diverse, it is inadequate to evaluate the pose accuracy of audio-driven avatar generation using ground truth head poses, we therefore manually set the generated head poses to range from $-60°$ to $60°$. Then, we use CSIM to measure identity preservation, FID to assess image quality, and employ AED and APD to evaluate motion accuracy.

**Qualitative results.** Fig. 6 shows the qualitative results. It can be observed that, SadTalker [73] produces artifacts in novel views that deviate significantly from the source image and fails to infer reasonable novel views given the source images with large poses, while Real3D [71] suffers from inaccurate lip motion. In contrast, our method can maintain high-fidelity and accurate lip motion across various poses and can effectively handle source images with large poses.

**Quantitative evaluations.** The quantitative results are reported in Table. 2. The results validate that our method achieves better reconstruction quality and pose accuracy than the comparison baselines, and produces accurate facial motion that is comparable to SadTalker [73].

TABLE 5
**Quantitative evaluation of the weighting network and identity condition in the deformation field.** The improved quantitative metrics validates the effectiveness of the adopted designs.

| Weighting network | w/ | w/o | Identity condition | w/ | w/o |
|---|---|---|---|---|---|
| AED ↓ | **0.135** | 0.144 | CSIM ↑ | **0.785** | 0.719 |

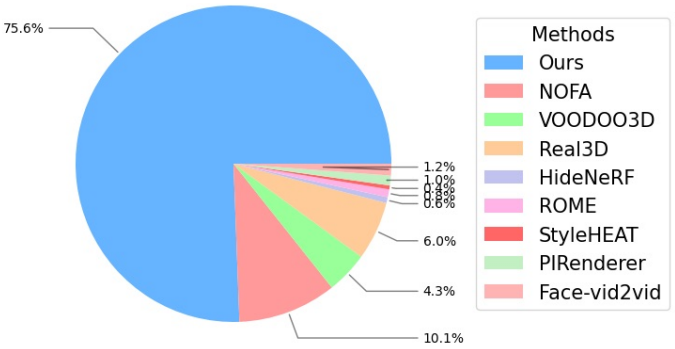



Fig. 11. **Statistical results of the user study.** Our method was favored by 75% of the participants, substantially surpassing the performance of the baselines.

### 4.4 Ablation Study

**Effectiveness of the coarse-to-fine generation strategy.** As depicted in Fig. 2, given a source image $I_s$, we adopt a coarse-to-fine generation strategy, including firstly employing an encoder-generator network to get a coarse canonical volume $V_c$, then using a detail network to help generate a refined volume $V_e$. The second row of Fig. 8 shows rendering results of $V_c$. It can be observed that although the rendered images possess similar structure and colors to the source image, they lack detailed texture and identity information. The third row of Fig. 8 shows rendering results of $V_e$, its reconstruction quality is significantly improved compared to the coarse reconstruction results, exhibiting more detailed texture and identity that closely resemble the source images. For quantitative evaluation, we conduct self-reenactment experiment employed in Sec. 4.2, and report the reconstruction quality of the two stages in Table. 4. The results further validate the effectiveness of the coarse-to-fine generation strategy. Compared to the naive compensation strategy and network structure proposed in NOFA [31], which were insufficient to faithfully supplement the information loss and required time-consuming and GPU-intensive test-time finetuning to achieve accurate reconstruction, the novel network structure and refine strategy in this work is more effective. It allows for a more precise extraction of detailed information from the source image, enabling the generation


Fig. 12. **Application of virtual avatar synthesis.** The top row shows the driving video, the following shows generated virtual avatars.

of high-fidelity 3D avatars without the need for laborious test-time finetuning. Moreover, it achieves a reconstruction quality comparable to the optimization-based NOFA [31], as validated in Sec. 4.

**Evaluation of the deformation field.** We learn a deformation field to model facial dynamics. As shown in Fig. 3, we employ 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. Qualitative ablation of these designs are shown in Fig. 7. (a). It can be found that without the identity parameter as condition, the generated face suffers from identity leakage from the driving frame, and can not preserve the identity information of the source image. Furthermore, without the weighting network, the generated expressions are inaccurate and inconsistent with the driving frame. We also conduct self-reenactment experiment employed in Sec. 4.2, using 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 validates the effectiveness of the proposed designs. In Fig. 7. (b), we visualize the weight scalars predicted by the weighting network. Specifically, given a source image, we shoot camera rays in the front view onto the reconstructed canonical volume, and sample four 2D coordinate planes along the camera's z-axis. The heat maps show the normalized weights of the sampled positions. Their distribution

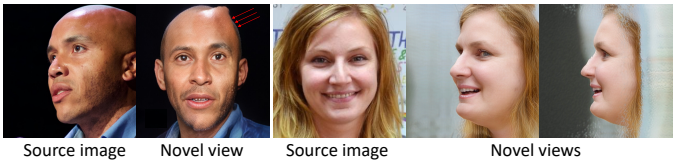



Fig. 13. **Limitations of our method.** When handling source images with large pose, our method may encounter artifacts. Additionally, since our training data contains only frontal faces, the generated avatar only includes the frontal face region.

indicates that the weighting network has learned to assign higher weights to the facial region, particularly the dynamic areas such as the eyes and mouth. Conversely, it assigns smaller weights to the background areas that remain static. It further validates the weighting network helps to improve the motion accuracy.

**Effectiveness of training with hybrid data.** After training the image encoder, the compensation network and the deformation field separately, we perform a refinement training stage that utilize both synthetic and real data to enhance the texture quality of the generated avatars, as the real video dataset [69] exhibits relatively inferior texture quality. As shown in Fig. 9, without refinement training with hybrid data, the generated avatar displays low texture quality and poor identity preservation. After training with hybrid data, the generation quality is substantially enhanced. We conduct self-reenactment experiment employed in Sec. 4.2 for quantitative evaluation, results are listed in Table. 3, which also validate the effectiveness of the refinement training stage.

**Effectiveness of fast adaption.** While our method can well handle common faces with an tuning-free manner, it may encounter artifacts when dealing with complex cases such as out-of-domain faces or faces with heavy make-ups. We address this issue by employing a fast adaptation strategy. As shown in Fig. 10, given a source image with heavy make-up, directly reconstruct it using our model produces inferior results. In comparison, adopting fast adaption on the source image effectively helps to improve the reconstruction quality.

### 4.5 Application on virtual avatar synthesis

In addition to modeling real-world faces, we also investigate applications on virtual avatar synthesis. As our method is based on a pretrained 3D GAN [22] that is capable of generating virtual 3D faces from its latent space, our method inherently supports randomly generating virtual avatars. We show examples on video-driven avatars synthesis in Fig. 12, which demonstrates that our method is capable of generating realistic virtual avatars with accurate motion.

## 5 Limitations and Ethical Issues

**Limitations.** Although our method can reconstruct high-fidelity 3D avatar from a single source image, it still has some limitations. As shown in Fig. 13, when dealing with source images with large pose, our method may encounter artifacts. Additionally, since our training data contains only frontal faces, the generated avatar only includes the frontal face region. Furthermore, the facial motion is modeled by a deformation field that relies on a pretrained 3DMM extractor to estimate the 3DMM parameters as control signals, such pretrained model may produce inaccurate results especially when handling extreme expressions beyond its capacity. Applying $360^\circ$ data for full head reconstruction and more robust motion representation for motion modeling would be our future plans.

**Ethical issues.** Since our framework can reconstruct high-fidelity facial avatars using just a single image, it may pose a risk for malicious uses, such as deep-fakes. We are acutely aware of the potential for our approach to be misused. Therefore, we plan to investigate the implementation of robust video watermarks for the synthesized videos when we release the code. On the other hand, we hope that our work can also advance the researches on privacy protection and deep-fake detection.

## References

[1] F. Yin, Y. Zhang, X. Cun, M. Cao, Y. Fan, X. Wang, Q. Bai, B. Wu, J. Wang, and Y. Yang, "Styleheat: One-shot high-resolution editable talking face generation via pretrained stylegan," *ECCV*, 2022.

[2] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "First order motion model for image animation," *Advances in Neural Information Processing Systems (NIPS)*, vol. 32, 2019.

[3] T.-C. Wang, A. Mallya, and M.-Y. 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 (CVPR)*, 2021, pp. 10 039–10 049.

[4] Y. Wang, D. Yang, F. Bremond, and A. Dantcheva, "Latent image animator: Learning to animate images via latent space navigation," *arXiv preprint arXiv:2203.09043*, 2022.

[5] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "Animating arbitrary objects via deep motion transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 2377–2386.

[6] N. Drobyshev, J. Chelishev, T. Khakhulin, A. Ivakhnenko, V. Lempitsky, and E. Zakharov, "Megaportraits: One-shot megapixel neural head avatars," *arXiv preprint arXiv:2207.07621*, 2022.

[7] K. Cheng, X. Cun, Y. Zhang, M. Xia, F. Yin, M. Zhu, X. Wang, J. Wang, and N. Wang, "Videoretalking: Audio-based lip synchronization for talking head video editing in the wild," in *SIGGRAPH Asia 2022*, 2022.

[8] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[9] H. Zhu, W. Wu, W. Zhu, L. Jiang, S. Tang, L. Zhang, Z. Liu, and C. C. Loy, "Celebv-hq: A large-scale video facial attributes dataset," in *European Conference on Computer Vision (ECCV)*, 2022, pp. 650–667.

[10] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 1999, pp. 187–194.

[11] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3d face model for pose and illumination invariant face recognition," in *2009 sixth IEEE international conference on advanced video and signal based surveillance*, 2009, pp. 296–301.

[12] Y. Deng, J. Yang, S. Xu, D. Chen, Y. Jia, and X. 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*, 2019, pp. 0–0.

[13] Y. Feng, H. Feng, M. J. Black, and T. Bolkart, "Learning an animatable detailed 3d face model from in-the-wild images," *ACM Transactions on Graphics (TOG)*, vol. 40, no. 4, pp. 1–13, 2021.

[14] P. Garrido, M. Zollhöfer, D. Casas, L. Valgaerts, K. Varanasi, P. Pérez, and C. Theobalt, "Reconstruction of personalized 3d face rigs from monocular video," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 3, pp. 1–15, 2016.

[15] S. Sanyal, T. Bolkart, H. Feng, and M. J. Black, "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)*, 2019, pp. 7763–7772.

[16] A. Tewari, F. Bernard, P. Garrido, G. Bharaj, M. Elgharib, H.-P. Seidel, P. Pérez, M. Zollhofer, and C. Theobalt, "Fml: Face model learning from videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 10 812–10 822.

[17] J. Xing, M. Xia, Y. Zhang, X. Cun, J. Wang, and T.-T. Wong, "Codetalker: Speech-driven 3d facial animation with discrete motion prior," *arXiv preprint arXiv:2301.02379*, 2023.

[18] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *European Conference on Computer Vision (ECCV)*, 2020.

[19] S. Athar, Z. Xu, K. Sunkavalli, E. Shechtman, and Z. Shu, "Rignerf: Fully controllable neural 3d portraits," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 20 364–20 373.

[20] P.-W. Grassal, M. Prinzler, T. Leistner, C. Rother, M. Nießner, and J. Thies, "Neural head avatars from monocular rgb videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 18 653–18 664.

[21] G. Gafni, J. Thies, M. Zollhofer, and M. Nießner, "Dynamic neural radiance fields for monocular 4d facial avatar reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 8649–8658.

[22] E. R. Chan, C. Z. Lin, M. A. Chan, K. Nagano, B. Pan, S. D. Mello, O. Gallo, L. Guibas, J. Tremblay, S. Khamis, T. Karras, and G. Wetzstein, "Efficient geometry-aware 3D generative adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision (CVPR)*, 2021.

[23] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021.

[24] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, and D. Cohen-Or, "Encoding in style: a stylegan encoder for image-to-image translation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[25] Y. Alaluf, O. Patashnik, and D. Cohen-Or, "Restyle: A residual-based stylegan encoder via iterative refinement," in *Proceedings of the IEEE International Conference on Computer Vision (CVPR)*, 2021.

[26] D. Roich, R. Mokady, A. H. Bermano, and D. Cohen-Or, "Pivotal tuning for latent-based editing of real images," *arXiv preprint arXiv:2106.05744*, 2021.

[27] T. Wang, Y. Zhang, Y. Fan, J. Wang, and Q. Chen, "High-fidelity gan inversion for image attribute editing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[28] Y. Zheng, V. F. Abrevaya, M. C. Bühler, X. Chen, M. J. Black, and O. Hilliges, "Im avatar: Implicit morphable head avatars from videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 13 545–13 555.

[29] X. Zhao, L. Wang, J. Sun, H. Zhang, J. Suo, and Y. Liu, "Havatar: High-fidelity head avatar via facial model conditioned neural radiance field," *ACM Trans. Graph.*, 2023.

[30] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," in *INTERSPEECH*, 2017.

[31] A. authors, "Nofa: Nerf-based one-shot facial avatar reconstruction," in *ACM SIGGRAPH 2023 Conference Proceedings*, 2023.

[32] Y. Guo, K. Chen, S. Liang, Y.-J. Liu, H. Bao, and J. Zhang, "Ad-nerf: Audio driven neural radiance fields for talking head synthesis," in *Proceedings of the IEEE International Conference on Computer Vision*, 2021.

[33] Z. Wang, T. Bagautdinov, S. Lombardi, T. Simon, J. Saragih, J. Hodgins, and M. Zollhofer, "Learning compositional radiance fields of dynamic human heads," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 5704–5713.

[34] A. Raj, M. Zollhofer, T. Simon, J. Saragih, S. Saito, J. Hays, and S. Lombardi, "Pixel-aligned volumetric avatars," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 11 733–11 742.

[35] S.-Y. Su, F. Yu, M. Zollhöfer, and H. Rhodin, "A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose," *Advances in Neural Information Processing Systems (NIPS)*, pp. 12 278–12 291, 2021.

[36] W. Xian, J.-B. Huang, J. Kopf, and C. Kim, "Space-time neural irradiance fields for free-viewpoint video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 9421–9431.

[37] E. Tretschk, A. Tewari, V. Golyanik, M. Zollhöfer, C. Lassner, and C. Theobalt, "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)*, 2021, pp. 12 959–12 970.

[38] K. Park, U. Sinha, J. T. Barron, S. Bouaziz, D. B. Goldman, S. M. Seitz, and R. Martin-Brualla, "Nerfies: Deformable neural radiance fields," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 5865–5874.

[39] J. Zhang, X. Li, Z. Wan, C. Wang, and J. Liao, "Fdnerf: Few-shot dynamic neural radiance fields for face reconstruction and expression editing," in *SIGGRAPH Asia 2022*, 2022.

[40] X. Gao, C. Zhong, J. Xiang, Y. Hong, Y. Guo, and J. Zhang, "Reconstructing personalized semantic facial nerf models from monocular video," *ACM Transactions on Graphics (TOG)*, 2022.

[41] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4401–4410.

[42] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[43] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila, "Alias-free generative adversarial networks," *arxiv:2106.12423*, 2021.

[44] E. R. Chan, M. Monteiro, P. Kellnhofer, J. Wu, and G. Wetzstein, "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)*, 2021.

[45] K. Schwarz, Y. Liao, M. Niemeyer, and A. Geiger, "Graf: Generative radiance fields for 3d-aware image synthesis," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[46] X. Chen, Y. Deng, and B. Wang, "Mimic3d: Thriving 3d-aware gans via 3d-to-2d imitation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[47] D. Wang, P. Chandran, G. Zoss, D. Bradley, and P. Gotardo, "Morf: Morphable radiance fields for multiview neural head modeling," in *SIGGRAPH 2022*, 2022.

[48] A. W. Bergman, P. Kellnhofer, Y. Wang, E. R. Chan, D. B. Lindell, and G. Wetzstein, "Generative neural articulated radiance fields," *arXiv preprint arXiv:2206.14314*, 2022.

[49] Y. Wu, Y. Deng, J. Yang, F. Wei, Q. Chen, and X. Tong, "Anifacegan: Animatable 3d-aware face image generation for video avatars," *arXiv preprint arXiv:2210.06465*, 2022.

[50] J. Sun, X. Wang, L. Wang, X. Li, Y. Zhang, H. Zhang, and Y. Liu, "Next3d: Generative neural texture rasterization for 3d-aware head avatars," *arXiv preprint arXiv:2211.11208*, 2022.

[51] J. Tang, B. Zhang, B. Yang, T. Zhang, D. Chen, L. Ma, and F. Wen, "Explicitly controllable 3d-aware portrait generation," *arXiv preprint arXiv:2209.05434*, 2022.

[52] R. Abdal, Y. Qin, and P. Wonka, "Image2stylegan: How to embed images into the stylegan latent space?" in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[53] Abdal, Rameen and Qin, Yipeng and Wonka, Peter, "Image2stylegan++: How to edit the embedded images?" in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8296–8305.

[54] J. Zhu, Y. Shen, D. Zhao, and B. Zhou, "In-domain gan inversion for real image editing," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2020.

[55] J. Xie, H. Ouyang, J. Piao, C. Lei, and Q. Chen, "High-fidelity 3d gan inversion by pseudo-multi-view optimization," in *CVPR*, 2023.

[56] Y. Deng, B. Wang, and H.-Y. Shum, "Learning detailed radiance manifolds for high-fidelity and 3d-consistent portrait synthesis from monocular image," in *CVPR*, 2023.

[57] Z. Yuan, Y. Zhu, Y. Li, H. Liu, and C. Yuan, "Make encoder great again in 3d gan inversion through geometry and occlusion-aware encoding," in *ICCV*, 2023.

[58] Y. Lan, X. Meng, S. Yang, C. C. Loy, and B. Dai, "Self-supervised geometry-aware encoder for style-based 3d gan inversion," in *CVPR*, 2023.

[59] Y. Ren, G. Li, Y. Chen, T. H. Li, and S. Liu, "Pirenderer: Controllable portrait image generation via semantic neural rendering,"

in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 13 759–13 768.
[60] S. Wang, Y. Ma, Y. Ding, Z. Hu, C. Fan, T. Lv, Z. Deng, and X. Yu, "Styletalk++: A unified framework for controlling the speaking styles of talking heads," *IEEE TPAMI*, 2024.
[61] T. Khakhulin, V. Sklyarova, V. Lempitsky, and E. Zakharov, "Realistic one-shot mesh-based head avatars," in *ECCV 2022*, 2022.
[62] W. Li, L. Zhang, D. Wang, B. Zhao, Z. Wang, M. Chen, B. Zhang, Z. Wang, L. Bo, and X. Li, "One-shot high-fidelity talking-head synthesis with deformable neural radiance field," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.
[63] Y. Blau and T. Michaeli, "Rethinking lossy compression: The rate-distortion-perception tradeoff," in *International Conference on Machine Learning*, 2019, pp. 675–685.
[64] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, p. 20150202, 2016.
[65] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, "Learning a model of facial shape and expression from 4d scans." *ACM Transactions on Graphics (TOG)*, vol. 36, no. 6, pp. 194–1, 2017.
[66] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2018, pp. 586–595.
[67] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE signal processing letters*, 2016.
[68] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2019, pp. 4690–4699.
[69] H. Zhu, W. Wu, W. Zhu, L. Jiang, S. Tang, L. Zhang, Z. Liu, and C. C. Loy, "CelebV-HQ: A large-scale video facial attributes dataset," in *ECCV*, 2022.
[70] W. Zhang, X. Cun, X. Wang, Y. Zhang, X. Shen, Y. Guo, Y. Shan, and F. Wang, "Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.
[71] Z. Ye, T. Zhong, Y. Ren, J. Yang, W. Li, J. Huang, Z. Jiang, J. He, R. Huang, J. Liu *et al.*, "Real3d-portrait: One-shot realistic 3d talking portrait synthesis," in *ICLR*, 2024.
[72] P. Tran, E. Zakharov, L.-N. Ho, A. T. Tran, L. Hu, and H. Li, "Voodoo 3d: Volumetric portrait disentanglement for one-shot 3d head reenactment," in *CVPR*, 2024.
[73] W. Zhang, X. Cun, X. Wang, Y. Zhang, X. Shen, Y. Guo, Y. Shan, and F. Wang, "Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation," in *CVPR*, 2023.
[74] W. Li, Z. Lin, K. Zhou, L. Qi, Y. Wang, and J. Jia, "Mat: Mask-aware transformer for large hole image inpainting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.
[75] Z. Ke, J. Sun, K. Li, Q. Yan, and R. W. Lau, "Modnet: Real-time trimap-free portrait matting via objective decomposition," in *AAAI*, 2022.
[76] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
[77] Z. Zhang, L. Li, Y. Ding, and C. Fan, "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)*, 2021, pp. 3661–3670.
[78] L. Xie, X. Wang, H. Zhang, C. Dong, and Y. Shan, "Vfhq: A high-quality dataset and benchmark for video face super-resolution," in *The IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022.
[79] Y. Wu, J. Zhang, H. Fu, and X. Jin, "Lpff: A portrait dataset for face generators across large poses," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

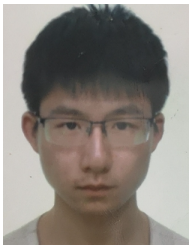
**Wangbo Yu** is currently a Ph.D. student at the School of Computer Science, Peking University. He received a B.E. degree in telecommunications engineering from Xidian University in 2021. His research interests include low-level computer vision, 3D vision and Generative models.

**Chaoran Feng** is a master student of Computer Science in Peking University. He pursued his undergraduate studies at the School of Future Technology, Dalian University of Technology, majoring in Artificial Intelligence. His research involves 3D reconstruction for large scale scene and sparse views, 3D generation, SLAM, and spiking neural network.

**Yanbo Fan** is a Research Scientist at Ant Research. He received his Ph.D. degree from Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China, in 2018, and his B.S. degree in Computer Science and Technology from Hunan University in 2013. His research interests are computer vision and machine learning.

**Yong Zhang** is a senior researcher at Tencent AI Lab. He received his Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences in 2018. From 2015 to 2017, he was a visiting scholar at the Rensselaer Polytechnic Institute. His research interests include computer vision and machine learning.

**Xuan Wang** is a scientific researcher in Interaction Intelligence Lab, Ant Research. He recieved a Ph.D. degree from School of Computer Science and Technology, Xi'an Jiaotong University in 2019. He was a visiting student at NICTA in 2015. He received a Master degree from School of Software Engineering, Xi'an Jiaotong University in 2010, and received Bachelor degree from Department of Computer Science and Technology, Xi'an University of Science and Technology in 2007. His research interests include neural rendering, non-rigid 3D reconstruction, performance capture, image synthesis and relevant applications.



**Fei Yin** is a first year Ph.D. student in the department of Computer Science, University of Cambridge. Previously, he obtained his Master degree from Tsinghua University. He received his bachelor's degree in Department of Software Engineering from Northeastern University in 2020. His research interest includes 2D and 3D Generation.

**Yunpeng Bai** is a Ph.D. student in Computer Science at UT Austin. His interests lie in using computer algorithms to create visually striking content, with a focus on 3D foundational models and 3D visual content generation. He previously graduated with a master's degree from Tsinghua University.

**Baoyuan Wu** is an Associate Professor of School of Data Science, the Chinese University of Hong Kong, Shenzhen. He received the PhD degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, on June 2014. From November 2016 to August 2020, he was a Senior and Principal Researcher at Tencent AI lab. His research interests are AI security and privacy, machine learning, computer vision and optimization.

**Yan-Pei Cao** received the bachelor's and Ph.D. degrees in computer science from Tsinghua University in 2013 and 2018, respectively. He is currently the Head of Research and Founding Team at VAST. His research interests include computer graphics and 3D computer vision.

**Li Yuan** received the B.E. degree from University of Science and Technology of China, in 2017, and the PhD degree from National University of Singapore, in 2021. He is currently a tenure-track assistant professor with School of Electrical and Computer Engineering with Peking University. He has published more than 40 papers on top conferences/journals. His research interests include deep learning, image processing, and computer vision.

**Yonghong Tian** (Fellow, IEEE) is currently the Dean of the School of Electronics and Computer Engineering, a Boya Distinguished Professor with the School of Computer Science, Peking University, China, and the Deputy Director of the Artificial Intelligence Research, Peng Cheng Laboratory, Shenzhen, China. He is the author or coauthor of over 350 technical papers in refereed journals and conferences. His research interests include neuromorphic vision, distributed machine learning, and AI for science. He is a TPC Member of more than ten conferences, such as CVPR, ICCV, ACM KDD, AAAI, ACM MM, and ECCV. He is a Senior Member of CIE and CCF and a member of ACM. He was a recipient of the Chinese National Science Foundation for Distinguished Young Scholars in 2018, two National Science and Technology Awards, and three ministerial-level awards in China. He received the 2015 Best Paper Award for *EURASIP Journal on Image and Video Processing*, the Best Paper Award from IEEE BigMM 2018, and the 2022 IEEE SA Standards Medallion and SA Emerging Technology Award. He served as the TPC Co-Chair for BigMM 2015, the Technical Program Co-Chair for IEEE ICME 2015, IEEE ISM 2015, and IEEE MIPR 2018/2019, and the General Co-Chair for IEEE MIPR 2020 and ICME 2021. He was/is an Associate Editor of IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY from January 2018 to December 2021, IEEE TRANSACTIONS ON MULTIMEDIA from August 2014 to August 2018, *IEEE Multimedia Magazine* from January 2018 to August 2022, and IEEE ACCESS from January 2017 to December 2021. He co-initiated the IEEE International Conference on Multimedia Big Data (BigMM).