S^{3}D-NeRF: Single-Shot Speech-Driven Neural Radiance Field for High Fidelity Talking Head Synthesis

Dongze Li[1,2] Kang Zhao[3], Wei Wang[2⋆], Yifeng Ma[3],
Bo Peng[2], Yingya Zhang[3] and Jing Dong[2]

[1] School of Artificial Intelligence, University of Chinese Academy of Sciences
[2] NLPR & MAIS, Institute of Automation, Chinese Academy of Sciences
[3] Alibaba Group
dongze.li@cripac.ia.ac.cn; {wwang,bo.peng,jdong}@nlpr.ia.ac.cn
{zhaokang.zk, yingya.zyy, mayifeng.myf}@alibaba-inc.com

Abstract. Talking head synthesis is a practical technique with wide applications. Current Neural Radiance Field (NeRF) based approaches have shown their superiority on driving one-shot talking heads with videos or signals regressed from audio. However, most of them failed to take the audio as driven information directly, unable to enjoy the flexibility and availability of speech. Since mapping audio signals to face deformation is non-trivial, we design a Single-Shot Speech-Driven Neural Radiance Field (S^{3}D-NeRF) method in this paper to tackle the following three difficulties: learning a representative appearance feature for each identity, modeling motion of different face regions with audio, and keeping the temporal consistency of the lip area. To this end, we introduce a Hierarchical Facial Appearance Encoder to learn multi-scale representations for catching the appearance of different speakers, and elaborate a Cross-modal Facial Deformation Field to perform speech animation according to the relationship between the audio signal and different face regions. Moreover, to enhance the temporal consistency of the important lip area, we introduce a lip-sync discriminator to penalize the out-of-sync audio-visual sequences. Extensive experiments have shown that our S^{3}D-NeRF surpasses previous arts on both video fidelity and audio-lip synchronization.

Keywords: Talking Head · Neural Radiance Fields

1 Introduction

Speech driven talking head synthesis is a promising technique and can be applied to a wide range of situations such as digital human, film making, virtual reality and video games. Current Neural Radiance Field (NeRF) [24] based methods [11,17,38] have shown their superiority on generating vivid talking portraits

⋆ Corresponding author

with high quality for their 3D consistency and view controllablity. However, they can only be applied to a specific identity, and require a long video sequence from the same speaker to train on, which hampers them from broader application scenarios. Afterwards some progress have been made in the generalization of NeRF-based talking head methods. Most of them [18, 23, 41] are driven by intermediate facial representations, such as 3DMM coefficients extracted from video or expression coefficients regressed from audio. These intermediate representations may introduce information loss more or less, leading to sub-optimal synthesis results.

Considering it is non-trivial to map audio signals to face deformation, we attribute the challenge of performing high-fidelity one-shot speech animation with Neural Radiance Fields to three aspects: **1) Learning Representative Appearance Features for Each Identity.** Different people have different facial appearance i.e. shape and texture, it is difficult for a vanilla NeRF to model these details of several speakers simultaneously. During the inference phase, it becomes even harder to generate a talking head sequence with high quality given that only a single image is available. **2) Modeling Motion of Different Face Regions with Audio.** Video driven techniques can utilize global driving signals which can describe the motion of the whole face, such as depth priors [12], 3DMM meshes [41] or PNCC [10,18] maps. However, in a speech driven task, the input audio merely has strong correlation with the lower face [27, 34]. Directly predicting face motions according to the input audio will cause insufficient modeling of different face regions, harming the fidelity of the synthesized portrait. **3) Keeping Temporal Consistency of the Lip Area.** The lip area has the greatest importance in speech animation. Due to the lack of constraints on this part, videos generated by NeRFs have sometimes shown incorrect mouth shapes.

In this paper, we propose S^3D-NeRF, namely Single-Shot Speech-Driven Neural Radiance Field, to synthesize high fidelity talking head videos. Given a single shot source image and a driven audio sequence, our method can synthesize vivid free view talking head videos with accurate mouth shapes.

To fully capture the appearance of an arbitrary speaker, a *Hierarchical Facial Appearance Encoder* is presented to extract the multi-scale facial features of the single-shot source image with a down sample convolutional network armed with a feature pyramid. These features contain rich structural and textural information of the source identity. Then, a multi-scale tri-plane [4] representation is constructed for neural rendering. To realize accurate speech animation, a *Cross-modal Facial Deformation Field* is proposed to faithfully catch the correlations between the speech signal and the visual features from different face regions through cross attention, and predicts the motion of the whole face precisely. To further enhance the consistency of the lip area, a *lip-sync discriminator* is imposed on the lower face of the generated video frames and the driving audio signal. By penalizing the out-of-sync audio-visual sequences, higher temporal consistency of the lip area can be achieved. During training, we adapt a coarse-to-fine image generation strategy to reduce the difficulty of modeling multiple speakers at the same time. Specifically, a coarse talking head frame which

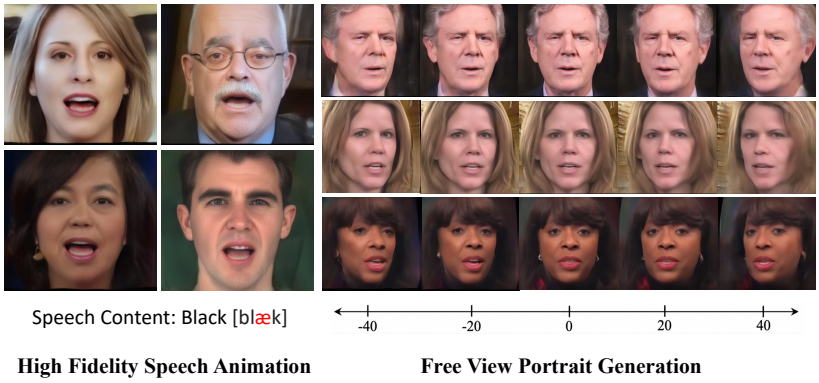

Fig. 1: A showcase of our S^{3}D-NeRF, which generates high quality face portraits with fine-grained face texture and mouth details .

contains the important inner face area is synthesized through standard volume neural rendering, while the texture details are refined with a super-resolution module, resulting in a high fidelity talking head frame.

Our main contributions can be summarized as below:

- We propose S^{3}D-NeRF, a single-shot NeRF-based talking head synthesis framework. Our method extend NeRF-based speech animation techniques to handle arbitrary unseen identities.
- Several key components are presented to assist the talking head synthesis procedure. Including a Hierarchical Facial Appearance Encoder for representative feature extraction, a Cross-modal Facial Deformation Field for accurate speech animation, and a lip-sync discriminator for better temporal consistency of lip area.
- Comprehensive experiments have shown the superiority of our S^{3}D-NeRF over previous arts on both video fidelity and audio-lip synchronization.

2 Related Work

2.1 Speech Driven Talking Head Synthesis

The goal of speech driven talking head synthesis is to animate a speaker according to input audios. According to the generalization ability of the model to new identities, current speech driven methods can be divided into two categories: identity agnostic methods and identity specific methods.

Identity agnostic methods are capable of generating speech videos of an arbitrary speaker once their training procedure is finished, with one or several images or video sequences as inputs. Among them, single-shot methods [22, 27, 43, 45] have attracted considerable interests due to their high generalization ability and data efficiency. Image-based methods [27, 37, 45] utilize deep generative

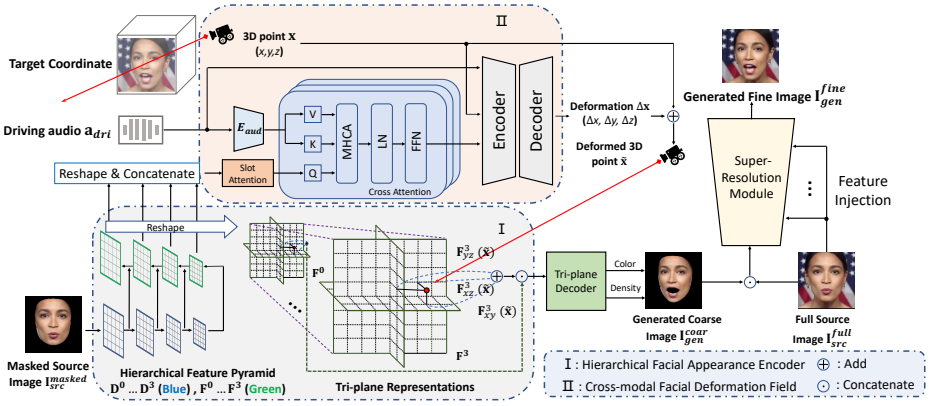

Fig. 2: The full pipeline of our S^{3}D-NeRF. The Hierarchical Facial Appearance Encoder extracts representative features from the masked face region of the single-shot source image, for high fidelity neural rendering of an arbitrary identity. The Cross-modal Facial Deformation Field accurately models the motion of different face regions, with the help of the correlation score calculated through cross attention between audio-visual features. Texture details are complemented with the super-resolution module.

networks [9] or auto-encoders [16] to generate continuous video clips. Model-based methods predict intermediate representations such as motion field [36], 2D landmarks [46] or 3D Morphable Model (3DMM) [3] coefficients [22, 29, 43] with the input audio signal, and animate the head with these representations. StyleTalk [22] predicts 3DMM through a transformer architecture and can generate portraits with diverse speaking styles under the help of a pretrained renderer [29]. SadTalker [43] uses several representation networks to learn lip and expression basis. Although remarkable progresses have been made, current single-shot methods still suffer in image quality and audio-lip synchronization because of the error accumulated in the audio to representation mapping.

Identity specific methods train an individual model for each identity with higher fidelity. Earlier methods [33, 40] utilize strong priors such as 3D meshes or dense key points for portrait generation. Current NeRF-based methods [11, 20, 31, 38] have shed light on high fidelity speaker synthesis. They optimize on the ground truth video clip of a single person, and can synthesize pose-controllable faces with fine-grained details. AD-NeRF [11] uses two separated NeRFs to model the head and the torso part respectively. SSP-NeRF [20] performs rays re-sampling based on the loss magnitude of different semantic regions. Currently, RAD-NeRF [32] and ER-NeRF [17] utilize hash table structures [25] to improve the rendering speed. GeneFace [39] proposes a facial radiance field with generalize ability but it also requires per identity finetuning and a complex multi-stage training process. Despite the above advantages, the above methods can only fit a single identity, and need tedious re-training when a new identity is encountered. At the same time, since no constraints are imposed on image sequences, the temporal consistency of the lip area is also not satisfactory.

2.2 NeRF-based Face Portrait Reconstruction

It is a natural idea to model human faces with NeRFs in a 3D aware manner. However, vanilla NeRFs struggle to handle dynamic objects, and are unsuitable for animating human faces which have diverse poses and expressions. There mainly exist two series of solutions for this problem. The former [14] is to condition the original Radiance Field with additional global control signals, e.g. expression parameters. The latter solve this problem by learning an additional deformation field [26, 28] to warp the 3D points. NerFace [8] is the first approach to reconstruct animatable face avatars with NeRFs conditioned on 3DMM expression parameters. HeadNeRF [13], NOFA [41] and OTAvatar [23] utilize GAN inversion to project onto the latent space of a 3D aware GAN [4] and drive the generated avatar through different latent codes. HiDe-NeRF [18] utilizes Projected Normalized Coordinate Code (PNCC) [10] as a global expression control indicator, and drives the face image with high fidelity. Comparing with the above mentioned methods, driving an arbitrary head with speech signal directly through NeRFs is a more challenging problem and is rarely explored.

3 Method

The overall pipeline of our S^{3}D-NeRF is shown in Fig.2. The 3D positions of the talking head can be derived through the given head pose of the target image $\mathbf{P}_{tar}$, which is available during both training and inference time. Firstly, our Hierarchical Facial Appearance Encoder takes a single-shot image as input, and models its appearance with a deep convolutional network, where multi-resolution feature maps are extracted to construct an efficient representation i.e. multi-scale appearance tri-planes, which are known to be capable of modeling several identities faithfully at the same time [4, 18]. Feature points from these planes are retrieved according to the 3D position for rendering. Then, the motion of the talking face are predicted by our Cross-modal Facial Deformation Field. It firstly calculates the correlation between an aggregated visual feature embedding and the audio signal through cross attention, then takes the result as a prior knowledge to assist the regression process of the deformation, which is used for calibrating the 3D position. A coarse face is rendered through the classic volume rendering process [24] with color and density predicted from the feature points on the tri-plane. Finally, a super-resolution module is used to refine the coarse face and yields the high fidelity portrait.

3.1 Hierarchical Facial Appearance Encoder

An effective feature representation is important to accurately model the appearance of a speaker. Initially proposed in [4], tri-plane has shown its strong representative power and is much faster comparing with the vanilla NeRFs. The Hierarchical Facial Appearance Encoder constructs this representation based on multi-scale feature maps from the one-shot source image, and yields the feature points for further volumetric rendering process.

Constructing Representative Appearance Features. We adopt a feature pyramid structure [19] to extract source features for its ability to learn multi-scale information from different resolutions, and reshape all the feature maps channel-wisely to construct several orthogonal planes as in [4]. To be more concrete, given the source image $\mathbf{I}_{src}$, feature maps with different resolutions $\{\mathbf{D}^0, \mathbf{D}^1, \mathbf{D}^2, \mathbf{D}^3\}$ are extracted through several convolutional downsample blocks $Down^i(.)$, while $\mathbf{D}^0 = Down^0(\mathbf{I}_{src})$ and $\mathbf{D}^i = Down^i(\mathbf{D}^{i-1})$. Then, hierarchical facial appearance features can be obtained by connecting the feature map at current i-th level with those at the higher $(i+1)$-th level, both of which are upsampled to the same resolution. 1×1 convolution layers are used to reduce the channel number to avoid overlarge computational cost:

$$\mathbf{F}^i = \begin{cases} Conv^i([Up^i(\mathbf{F}^{i+1}), \mathbf{D}^i]), & i = 0, 1, 2 \\ \mathbf{D}^i, & i = 3 \end{cases} \tag{1}$$

where $Up^i(.)$ denotes the i-th upsampling convolution layer, and $Conv^i(.)$ denotes the i-th 1×1 convolution layer mentioned above. [...] denotes channel-wise concatenation. Afterwards, the feature map at i-th level F^i has a shape of $B \times 3 \times C^i \times H^i \times W^i$, and is reshaped and splitted into three orthogonal sub feature maps, each of which has a shape of $B \times C^i \times H^i \times W^i$, representing coordinates in the three-dimensional space, i.e. plane $\mathbf{F}_{xy}$, $\mathbf{F}_{yz}$ and $\mathbf{F}_{xz}$ respectively. B is the batch size, and C^i, H^i and W^i are the channel number, the height and width at i-th level. Since this tri-plane representation is obtained from multiple identities, it is capable of fitting an unseen person during inference time.

Extracting Features for Neural Rendering. To render the target image $\mathbf{I}_{tar}$, which may have different head pose with the source image, we need to find the correspondence between the position in the target image space and the source feature tri-plane space. This can be realized through camera transformation between the source pose $\mathbf{P}_{src}$ and the target pose $\mathbf{P}_{tar}$, together with bilinear interpolation. Concretely, given a 3D point $\mathbf{x}$ on a ray $\mathbf{r}(t)$ constructed with target pose $\mathbf{P}_{tar} = \{\mathbf{R}_{tar}, \mathbf{T}_{tar}\}$, it is firstly calibrated by the deformation predicted from the Cross-modal Facial Deformation Field (Sec.3.2) to get the deformed point $\tilde{\mathbf{x}}$. Then, feature vectors $\mathbf{F}_{xy}(\tilde{\mathbf{x}})$, $\mathbf{F}_{yz}(\tilde{\mathbf{x}})$ and $\mathbf{F}_{xz}(\tilde{\mathbf{x}})$ are retrieved from three orthogonal planes $\mathbf{F}_{xy}$, $\mathbf{F}_{yz}$ and $\mathbf{F}_{xz}$, according to the pose of the source image $\mathbf{P}_{src} = \{\mathbf{R}_{src}, \mathbf{T}_{src}\}$ via bilinear interpolation $Interp(.)$:

$$\mathbf{F}^i_{xyz}(\tilde{\mathbf{x}}) = Interp(\mathbf{F}^i, Inv(\mathbf{P}_{src}) \cdot \tilde{\mathbf{x}}), \tag{2}$$

$Inv(\mathbf{P}_{src})$ is the inverse source pose which transforms 3D positions onto the source tri-plane space. These extracted feature vectors are further used for generating the talking head image through volume rendering.

3.2 Cross-modal Facial Deformation Field

To animate the target speaker, a crucial step is to get the precise deformation which describe face motions driven by the speech signal. Solely predicting the deformation from audio struggles to learn the global motion of the whole face, since

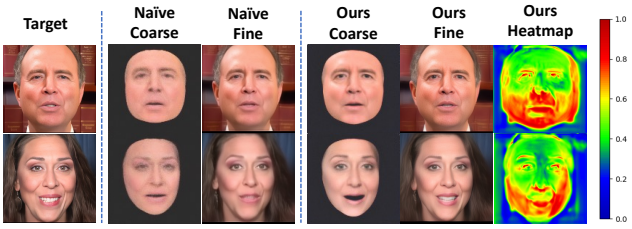

Fig. 3: Results with naive deformation module (left) and our Cross-modal Facial Deformation Field (right). Lower Face regions have the largest activations in the heatmap, which denote their strongest correlations with the driven speech signal.

audio is a local signal which only correlates strongly with the mouth area. Thus it tends to yield blurry results (shown in Fig.3, left). For correctly modeling the global motion, structural and textural information of the whole face is required. Our Cross-modal Facial Deformation Field firstly aggregates the multi-scale visual features to a unified embedding, which contains the global information of the entire face area, then calculates the corrlation score between the audio signal and this embedding through cross attention. The correlation score is used as a prior knowledge for determining the importance of motions from each part of the face, making deformation prediction more precise. As is shown in Fig.3, our Cross-modal Facial Deformation Field is free from the blurry issues because of the correctly learned motion.

Aggregating Multi-scale Facial Features. To learn the features which represents the whole face, all the hierarchical appearance feature maps $\mathbf{F}^0$ to $\mathbf{F}^3$ are first reshaped to the same resolution ($H_s \times W_s$) and concatenated along the channel. Then, these features are flattened and fed into a slot-attention module [21] to exchange information from different scales: $\mathbf{F}_{agg} = SlotAttention(\mathbf{F}^{0\ldots3})$. The aggregated slot feature embedding has a shape of $B \times (H_s \times W_s) \times D_{slot}$, where D_{slot} is the slot feature dimension. This learning process makes the embedding contain rich structural and textural information of the speaker.

Calculating Audio-Visual Correlation Scores. The input speech signal is firstly processed with a 1D full convolutional network to get the driving audio feature $\mathbf{a}_{dri}$. Relevance between audio and face motion is calculated through the Multi-Head Cross-Attention (MHCA) scores with the aggregated visual embedding $\mathbf{F}_{agg}$ as query, the driving audio feature $\mathbf{a}_{dri}$ as key and value:

$$\begin{aligned}\mathbf{F}_{cm} &= MHCA\left(\mathbf{F}_{agg}, \mathbf{a}_{dri}\right) \\ &= \text{Softmax}\left[\frac{\mathbf{F}_{agg}\boldsymbol{W}^Q\left(\mathbf{a}_{dri}\boldsymbol{W}^K\right)^T}{\sqrt{d}}\right]\mathbf{a}_{dri}\boldsymbol{W}^V,\end{aligned} \tag{3}$$

where $\boldsymbol{W}^Q$, $\boldsymbol{W}^K$, $\boldsymbol{W}^V$ are the projection matrices with hidden dimension d, respectively. Cross-attention scores between the speech signal and the whole face feature representation indicate the relevance of each face area with the speech

Fig. 4: Qualitative comparison with single-shot methods. Our S^{3}D-NeRF yields the most correct lip shapes and the clearest teeth. Note that different methods adopt different face alignment tools, and the ground truth row demonstrates the raw image without alignment, so the face poses from different methods are slightly different.

input, and can serve as a strong prior for modeling the dynamic talking face in a region-aware manner. We observed that the mouth area of the speaker, which moves with the audio signal has the largest deformation scales. Also shown in Fig.3, heatmaps are calculated by averaging $\mathbf{F}_{cm}$ channelwisely, and have shown larger activations on lower face regions which contains not only mouth area but also the muscles of the lower face.

Predicting Facial Deformation. The facial deformation prediction module $Deform(...)$ has a U-Net [30] like architecture, with the original 3D position $\mathbf{x}$, the cross-modal correlation score $\mathbf{F}_{cm}$, and the driving audio feature $\mathbf{a}_{dri}$ as inputs, this deformation prediction process can be formalized as

$$\begin{aligned} \Delta\mathbf{x} &= Deform(\mathbf{x}, \mathbf{a}_{dri}, \mathbf{F}_{cm}), \\ \tilde{\mathbf{x}} &= \mathbf{x} + \Delta\mathbf{x}. \end{aligned} \tag{4}$$

With the calibrated 3D points, source features from the tri-plane can be extracted faithfully through Eq.2.

3.3 High Fidelity Talking Head Generation

Volume Rendering. To generate an image with volume rendering, given the feature vector $\mathbf{F}^i_{xyz}(\tilde{\mathbf{x}})$ retrieved from the tri-plane, a tri-plane decoder takes it as input and predicts the color density $\mathbf{c}$ and occupancy σ at that position. The final color of a pixel $\mathbf{C}(\mathbf{r})$ along the ray is calculated through:

$$\mathbf{C}(\mathbf{r}) = \int_{t_n}^{t_f} T(t)\sigma(\mathbf{r}(t), \mathbf{a})\mathbf{c}(\mathbf{r}(t), \mathbf{a})dt, \tag{5}$$

Table 1: Quantitative comparison with single-shot methods, The **best** and the second best results are emphasized.

Dataset	Metric	GT	Wav2Lip	MakeItTalk	PC-AVS	StyleTalk	SadTalker	S^{3}D-NeRF
HDTF	SSIM ↑	1	0.749	0.559	0.582	0.776	0.678	**0.819**
	LPIPS ↓	0	0.332	0.422	0.411	0.289	0.386	**0.258**
	F-LMD ↓	0	3.582	6.846	3.925	3.623	4.621	**2.799**
	M-LMD ↓	0	3.652	6.152	3.912	3.273	3.654	**2.929**
	CPBD ↑	0.284	0.195	0.139	0.206	0.234	0.192	**0.263**
	Sync ↑	8.144	**8.365**	4.813	5.096	6.188	6.473	6.514
MEAD	SSIM ↑	1	0.829	0.626	0.673	**0.850**	0.755	0.845
	LPIPS ↓	0	0.212	0.367	0.359	0.199	0.223	**0.161**
	F-LMD ↓	0	2.126	6.081	8.372	2.622	3.835	**2.014**
	M-LMD ↓	0	3.377	6.790	5.192	3.061	3.631	**2.921**
	CPBD ↑	0.176	0.101	0.065	0.068	0.145	0.081	**0.152**
	Sync ↑	7.702	**7.713**	5.143	5.466	6.057	5.707	6.768

where $T(t) = \exp\left(-\int_{t_n}^{t} \sigma(s)ds\right)$ denotes for the accumulated transmittance along the ray from t_n to t, t_n and t_f are the lower and the upper bound of depth.

Coarse-to-fine Generation Strategy. Talking head frames contain various backgrounds and complex outer face textures which are hard for a NeRF to render at the same time. To decline the difficulty for training, we adopt a coarse-to-fine strategy. Specifically, the inner face part $\mathbf{I}^{masked}$ of an arbitrary frame $\mathbf{I}$ is decoupled in advance. During training, the inner face is rendered by the NeRF model, and the synthesized image is denoted as $\mathbf{I}_{gen}^{coar}$, while the fine texture details of the outer face and the background is complemented by an additional super-resolution module. We utilize Deep3DFaceRecon [7] to generate a 3D-aware face segmentation mask for inner-outer face area separation. This coarse-to-fine generation strategy guarantees that the NeRF can focus on the important face area.

Super Resolution Module. The super resolution module takes the coarse output face $\mathbf{I}_{gen}^{coar}$ and the full source face image $\mathbf{I}_{src}^{full}$ as input, and generates the talking head frame $\mathbf{I}_{fine}^{gen}$ with background and more details. The whole super resolution process can be written as $\mathbf{I}_{gen}^{fine} = Supres(\mathbf{I}_{gen}^{coar}, \mathbf{I}_{src}^{full})$. More details are shown in the appendix.

3.4 Lip-sync Discriminator

Our lip sync discriminator has a visual branch and an audio branch. The former takes T frames of the lip area as input and outputs their embedding $\mathbf{e}_l$, while the latter gets an audio clip and outputs its corresponding feature embedding $\mathbf{e}_a$. Then, the synchronization level between the audio and the video sequence

can be judged by calculating the cosine similarity between $\mathbf{e}_l$ and $\mathbf{e}_a$:

$$\cos(\mathbf{e}_l, \mathbf{e}_a) = \frac{\mathbf{e}_l \cdot \mathbf{e}_a}{\max\left(\|\mathbf{e}_l\|_2 \cdot \|\mathbf{e}_a\|_2, \epsilon\right)}, \tag{6}$$

where ϵ is a margin which is set to 0. Different from that used in Wav2Lip [27], our lip sync discriminator is trained on a larger mixed dataset containing HDTF [44] and LRS2 [1], with a single contrastive triplet loss for more distinctive feature representations:

$$\begin{aligned} \mathcal{L}_{sync}^{dis} = &\max\left(0, \eta + \cos(\mathbf{e}_l, \mathbf{e}_a^+) - \cos(\mathbf{e}_l, \mathbf{e}_a^-)\right) \\ &+ \max\left(0, \eta + \cos(\mathbf{e}_l^+, \mathbf{e}_a) - \cos(\mathbf{e}_l^-, \mathbf{e}_a)\right). \end{aligned} \tag{7}$$

$\mathbf{e}_a^+$ and $\mathbf{e}_a^-$ regard the positive and the negative audio embedding according to the lip sequence respectively, so as $\mathbf{e}_l^+$ and $\mathbf{e}_l^-$, η is a margin set to 0.5. While training our model, the lip-sync discriminator is frozen and T frames of the same person are generated once a time. The lip regions are cropped through the pre-extracted bounding boxes of the ground-truth. The loss used for S^{3}D-NeRF training can be formulated as

$$\mathcal{L}_{sync}^{gen} = \cos(\mathbf{e}_l^{gen}, \mathbf{e}_{gt}). \tag{8}$$

while $\mathbf{e}_l^{gen}$ is the embedding of the generated lip sequence and $\mathbf{e}_{gt}$ is the embedding of the ground-truth audio. By penalizing the out-of-sync audio-visual pair, more precise mouth shapes can be achieved.

3.5 Loss Functions

In addition to the lip synchronization loss $\mathcal{L}_{sync}^{gen}$, we use a pixelwise l_2 reconstruction loss $\mathcal{L}_{pix}$, a feature level perceptual loss $\mathcal{L}_{per}$ to match the generated face with the target face, and a GAN loss $\mathcal{L}_{adv}$ with the architecture of StyleGAN discriminator [15]. All the losses are imposed on both the coarse and the fine faces. An l_2 loss $\mathcal{L}_{deform}$ is used to regularize the prediction deformation. Our S^{3}D-NeRF is finally trained with the weighted sum of the above losses as:

$$\mathcal{L} = \mathcal{L}_{pix} + \lambda_{per}\mathcal{L}_{per} + \lambda_{adv}\mathcal{L}_{adv} + \lambda_{sync}^{gen}\mathcal{L}_{sync}^{gen} + \lambda_{deform}\mathcal{L}_{deform}, \tag{9}$$

with $\lambda_{per} = 0.01$, $\lambda_{adv} = 1$, $\lambda_{sync}^{gen} = 0.5$, $\lambda_{deform} = 0.001$.

4 Experiments

4.1 Experimental Settings

Datasets. Our S^{3}D-NeRF is trained on HDTF [44] training split, where each video ranges from tens of seconds to more than ten minutes in length, and evaluated on HDTF testing split and MEAD testing split [35] for comparison with single-shot methods. There is no overlap between the training and the testing

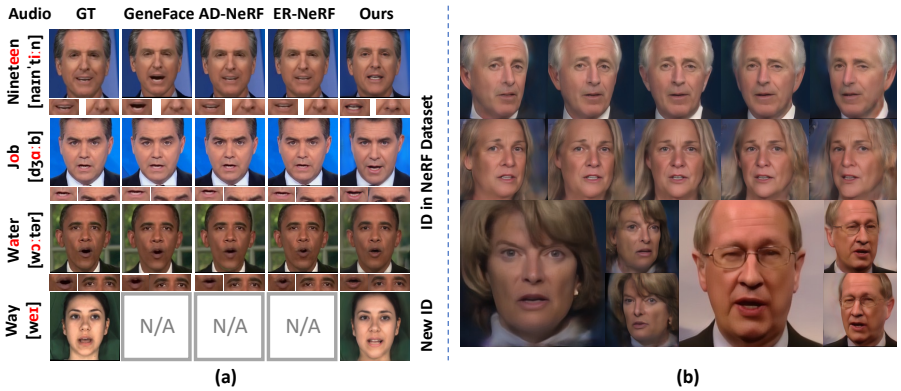

Fig. 5: Left (a): Qualitative comparison with NeRF-based methods, when encountering a new identity, our S^{3}D-NeRF successfully synthesizes a portrait faithfully without any retraining. Right (b): High fidelity generation results with multi-view consistency. Ground truth are the images with front view. Most of the face details are preserved.

split. For comparison with NeRF-based methods, we use the videos provided by AD-NeRF [11] and GeneFace [39]. The video sampling rate is set to 25 FPS. All the faces are cropped and aligned to a fix resolution of 256 × 256.

Implementation Details. The whole proposed framework is implemented by PyTorch and can be trained end-to-end. It takes about two days to train our S^{3}D-NeRF on 4 NVIDIA-A100 GPUs. The sample rate of the speech signal is set to 16000. We use Wav2Vec2.0 [2] to extract the driving audio features. More details and the user study can be found in the appendix.

4.2 Quantitative Results

Several widely used metrics are adopted to carry out our quantitative evaluation. We use SSIM and LPIPS [42] to measure the similarity of the generated speaker with the ground truth in pixel level and feature level respectively. To compare the identity fidelity, we use Facial Landmark Distance (F-LMD). To compare the clear extent of the images, we use Cumulative Probability of Blur Detection (CPBD). To evaluate the audio-lip synchronization, we use the SyncNet confidence score (Sync) [6] and Mouth Landmark Distance (M-LMD) [5].

Comparison with Single-Shot Methods. Our competitors includes Wav2Lip [27], MakeitTalk [46], PC-AVS [45], SadTalker [43] and StyleTalk [22]. For the above methods, we take their provided checkpoints and strictly follow the test protocol of current single-shot methods to conduct our evaluation. MakeItTalk has shown the worst image quality and lip shape. Wav2Lip also uses a pretrained lip-sync expert as supervision, and it has larger batch size and its model is simpler. Thus, it achieves a SyncNet score which is even better than ground truth. However, Wav2Lip tends to generate blurry faces, resulting in an inferior face fidelity. StyleTalk and SadTalker have shown competitive performance on image quality. Our S^{3}D-NeRF achieves the second-best lip-sync result, meanwhile sur-

Table 2: Quantitative comparison with NeRF-based methods. The **best** and the second best results are highlighted. Our S^{3}D-NeRF holds competitive image quality and audio-visual consistency, meanwhile the best generalize ability.

Metric	GT	GeneFace	AD-NeRF	ER-NeRF	S^{3}D-NeRF
SSIM ↑	1	0.828	0.846	**0.884**	0.852
LPIPS ↓	0	0.098	0.087	**0.068**	0.104
M-LMD ↓	0	2.436	1.982	1.659	**1.493**
F-LMD ↓	0	1.898	1.715	**1.158**	2.019
CPBD ↑	0.278	0.207	0.220	0.239	**0.244**
Sync ↑	8.970	5.366	5.626	6.781	**7.118**
FPS ↑	N/A	0.13	0.13	**29**	9.5
Fit Time / h ↓	N/A	42	36	4.5	<**0.01**

passes the competitors significantly on quality metrics at both image level and feature level. Moreover, our method has the lowest face landmark distance and mouth landmark distance, which indicates that our S^{3}D-NeRF keep facial and lip structure better than other methods. On MEAD dataset where all the identities are unseen before, our S^{3}D-NeRF still achieves competitive performance, which demonstrates the strong generalization ability of our approach.

Comparison with NeRF-based Methods. We also compare S^{3}D-NeRF with current identity specific NeRF-based methods. Three start-of-the-art approaches AD-NeRF [11] GeneFace [39] and ER-NeRF [17] are chosen as our opponents. It is worth to note that **our method has never seen the identity nor the driving audio in the evaluation set**, while the person-specific NeRFs are trained on a training set which has the same identity as the evaluation set (although no image overlap). In addition to the metrics mentioned above, we also take account of the data efficiency metrics, FPS: Frame Per Second during inference, which can reflects the speed of a rendering model. Fit Time: Times needed to generate the plausible talking head videos of a new identity with a pretrained model available. ER-NeRF has shown the strong image quality with the fastest rendering speed. GeneFace needs to train a postnet for each new identity, which takes extra time. Our model is also competitive in image quality. Moreover, it has the best lip synchronization score, together with the best mouth-LMD, which indicates the superiority on the specific audio-driven task. Moreover, our model has the highest CPBD, that means we generate the clearest images. Besides, S^{3}D-NeRF has the strongest generalization ability since it is free from retraining when coming up with an new speaker, and has shown an excellent inference speed.

4.3 Qualitative Results

Comparison with Single-Shot Methods. A direct comparison with other person agnostic methods are shown in Fig.4. Our S^{3}D-NeRF has shown more precise lip shapes and clearer teeth. At the same time, the detailed facial features such as skin texture, eyebrows and teeth splits are preserved by our method.

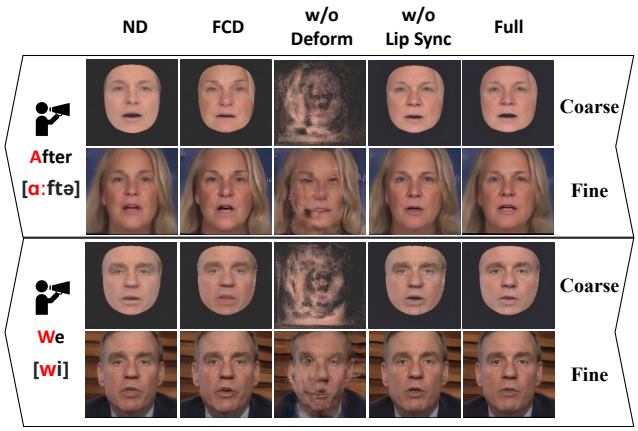

Fig. 6: Qualitative ablation study over key components.

Table 3: Quantitative ablation study. The full S^{3}D-NeRF achieves the best landmark distance and lip synthesis results, with a little sacrifice on image quality.

Method	SSIM ↑	LPIPS ↓	F-LMD ↓	M-LMD ↓	CPBD ↑	Sync ↑
ND	0.744	0.278	3.475	3.507	0.165	5.201
FCD	**0.834**	**0.242**	2.855	3.196	0.246	6.106
w/o Deform	0.556	0.457	N/A	N/A	0.140	0.240
w/o Lip Sync	0.801	0.263	3.276	3.342	0.255	5.649
S^{3}D-NeRF	0.829	0.258	**2.799**	**2.929**	**0.263**	**6.514**

Comparison with NeRF-based Methods. A qualitative comparison with NeRF-based methods are shown in Fig.5. AD-NeRF has shown the head-torso separation problem, which seriously harm the fidelity of the portrait (Obama face at the third row). Due to the identity specific training process, NeRF based methods can achieve excellent speaker reconstruction. Their generated face regions are very similar to the ground truth. **However, on a novel identity, all the person specific NeRFs failed to generate a plausible result.** On the contrary, our S^{3}D-NeRF has shown excellent generation results. The images from our S^{3}D-NeRF do not reach a similarity as high as NeRF-based methods, but they has more distinct facial details, each face region of the speaker are kept completely. Images generated by our method are sharper and their lip areas are much clearer.

Multi-view Generation Results. Benefited from the 3D consistency, our method can generate faces with various view angles with the same lip shape, which is shown in the right part of the Fig.5. The facial details such as wrinkles and ear-rings are well kept, indicating the generative power of our method.

4.4 Ablation Studies & Analysis

The Hierarchical Facial Appearance Encoder is a necessary structure which can not be discarded or replaced. To study the function of our Cross-modal Facial Deformation Field, we designed several kinds of variants. Firstly, we directly reconstruct the frame of the target speaker without deformation, denoted as w/o deform. In this case, the audio feature is concatenated with the tri-plane feature and fed into the NeRF model Secondly, we replace our Cross-modal Facial Deformation Field with a naive deformation module mentioned in Sec.3.2, denoted as Naive Deform (ND). Thirdly, to show the effect of the correlation score as a prior, we train a model with a deformation module which takes directly the tri-plane feature as an additional input, denoted as Feat-Concatenate Deform (FCD). A model without the lip-sync discriminator is denoted as w/o lip-sync. Qualitative and quantitative results are shown in Tab.3 and Fig.6.

The model without deformation modules totally fails to keep the appearance of the speaker, resulting in a poor video quality, not even performing speech animations. The Naive Deform model yields a coarse image similar to the average face, and its final generated talking head frames are less sharper. As for the Feat-Concatenate Deform model, it has kept the source feature better, resulting in a slight higher image quality metrics, however, because of it does not take the relationship between the speech signal and different face regions, some details of the mouth and the face area are missing. The model trained without the lip-sync discriminator synthesizes less precise lip shape. Our full model has the best performance on most of the metrics.

5 Limitations & Ethical Considerations

Limitations. Changing backgrounds like person specific-NeRFs [11, 17] is not supported currently. When the head pose is overlarge, the contour of the portrait sometimes tends to be blur. The above issues will be studied in future.
Ethical Considerations. Once trained, our S^{3}D-NeRF can synthesize videos of any given person saying something that they never actually said. This may cause some moral or legal problems. We are committed to offer our model to fight against the potential abuses.

6 Conclusion

In this paper, we propose a Single-Shot Speech-Driven Neural Radiance Field (S^{3}D-NeRF) for synthesizing audio driving talking heads with high fidelity. Several key components are proposed to assist the generation process, including a Hierarchical Facial Appearance Encoder, a Cross-modal Facial Deformation Field and a lip-sync discriminator. Our S^{3}D-NeRF is the first attempt to explore single-shot speech-driven talking head generation with NeRFs. Validated by comprehensive experiments, S^{3}D-NeRF has shown clear improvement against current state-of-the-art, both on video quality and generalization ability.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (NSFC) under Grants 62372452, 62272460, Youth Innovation Promotion Association CAS, and Alibaba Research Intern Program.

References

1. Afouras, T., Chung, J.S., Senior, A., Vinyals, O., Zisserman, A.: Deep audio-visual speech recognition. IEEE transactions on pattern analysis and machine intelligence **44**(12), 8717–8727 (2018)
2. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems **33**, 12449–12460 (2020)
3. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: Seminal Graphics Papers: Pushing the Boundaries, Volume 2, pp. 157–164 (2023)
4. Chan, E.R., Lin, C.Z., Chan, M.A., Nagano, K., Pan, B., De Mello, S., Gallo, O., Guibas, L.J., Tremblay, J., Khamis, S., et al.: Efficient geometry-aware 3d generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16123–16133 (2022)
5. Chen, L., Maddox, R.K., Duan, Z., Xu, C.: Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7832–7841 (2019)
6. Chung, J.S., Zisserman, A.: Out of time: automated lip sync in the wild. In: Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13. pp. 251–263. Springer (2017)
7. Deng, Y., Yang, J., Xu, S., Chen, D., Jia, Y., Tong, X.: Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In: IEEE Computer Vision and Pattern Recognition Workshops (2019)
8. Gafni, G., Thies, J., Zollhöfer, M., Nießner, M.: Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8649–8658 (June 2021)
9. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
10. Guo, J., Zhu, X., Yang, Y., Yang, F., Lei, Z., Li, S.Z.: Towards fast, accurate and stable 3d dense face alignment. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020)
11. Guo, Y., Chen, K., Liang, S., Liu, Y.J., Bao, H., Zhang, J.: Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5784–5794 (2021)
12. Hong, F.T., Zhang, L., Shen, L., Xu, D.: Depth-aware generative adversarial network for talking head video generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3397–3406 (2022)
13. Hong, Y., Peng, B., Xiao, H., Liu, L., Zhang, J.: Headnerf: A real-time nerf-based parametric head model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20374–20384 (2022)

14. Jang, W., Agapito, L.: Codenerf: Disentangled neural radiance fields for object categories. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12949–12958 (2021)
15. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8110–8119 (2020)
16. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
17. Li, J., Zhang, J., Bai, X., Zhou, J., Gu, L.: Efficient region-aware neural radiance fields for high-fidelity talking portrait synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 7568–7578 (October 2023)
18. Li, W., Zhang, L., Wang, D., Zhao, B., Wang, Z., Chen, M., Zhang, B., Wang, Z., Bo, L., Li, X.: One-shot high-fidelity talking-head synthesis with deformable neural radiance field (2023)
19. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)
20. Liu, X., Xu, Y., Wu, Q., Zhou, H., Wu, W., Zhou, B.: Semantic-aware implicit neural audio-driven video portrait generation. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII. pp. 106–125. Springer (2022)
21. Locatello, F., Weissenborn, D., Unterthiner, T., Mahendran, A., Heigold, G., Uszkoreit, J., Dosovitskiy, A., Kipf, T.: Object-centric learning with slot attention. Advances in Neural Information Processing Systems **33**, 11525–11538 (2020)
22. Ma, Y., Wang, S., Hu, Z., Fan, C., Lv, T., Ding, Y., Deng, Z., Yu, X.: Styletalk: One-shot talking head generation with controllable speaking styles. In: Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI) (2023)
23. Ma, Z., Zhu, X., Qi, G.J., Lei, Z., Zhang, L.: Otavatar: One-shot talking face avatar with controllable tri-plane rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16901–16910 (2023)
24. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: ECCV (2020)
25. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Transactions on Graphics (ToG) **41**(4), 1–15 (2022)
26. Park, K., Sinha, U., Barron, J.T., Bouaziz, S., Goldman, D.B., Seitz, S.M., Martin-Brualla, R.: Nerfies: Deformable neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5865–5874 (2021)
27. Prajwal, K., Mukhopadhyay, R., Namboodiri, V.P., Jawahar, C.: A lip sync expert is all you need for speech to lip generation in the wild. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 484–492 (2020)
28. Pumarola, A., Corona, E., Pons-Moll, G., Moreno-Noguer, F.: D-nerf: Neural radiance fields for dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10318–10327 (2021)
29. Ren, Y., Li, G., Chen, Y., Li, T.H., Liu, S.: Pirenderer: Controllable portrait image generation via semantic neural rendering (2021)
30. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted

Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)
31. Shen, S., Li, W., Zhu, Z., Duan, Y., Zhou, J., Lu, J.: Learning dynamic facial radiance fields for few-shot talking head synthesis. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XII. pp. 666–682. Springer (2022)
32. Tang, J., Wang, K., Zhou, H., Chen, X., He, D., Hu, T., Liu, J., Zeng, G., Wang, J.: Real-time neural radiance talking portrait synthesis via audio-spatial decomposition. arXiv preprint arXiv:2211.12368 (2022)
33. Thies, J., Elgharib, M., Tewari, A., Theobalt, C., Nießner, M.: Neural voice puppetry: Audio-driven facial reenactment. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16. pp. 716–731. Springer (2020)
34. Wang, J., Zhao, K., Zhang, S., Zhang, Y., Shen, Y., Zhao, D., Zhou, J.: Lipformer: High-fidelity and generalizable talking face generation with a pre-learned facial codebook. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13844–13853 (2023)
35. Wang, K., Wu, Q., Song, L., Yang, Z., Wu, W., Qian, C., He, R., Qiao, Y., Loy, C.C.: Mead: A large-scale audio-visual dataset for emotional talking-face generation. In: ECCV (2020)
36. Wang, S., Li, L., Ding, Y., Yu, X.: One-shot talking face generation from single-speaker audio-visual correlation learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 2531–2539 (2022)
37. Wiles, O., Koepke, A., Zisserman, A.: X2face: A network for controlling face generation using images, audio, and pose codes. In: Proceedings of the European conference on computer vision (ECCV). pp. 670–686 (2018)
38. Ye, Z., Jiang, Z., Ren, Y., Liu, J., He, J., Zhao, Z.: Geneface: Generalized and high-fidelity audio-driven 3d talking face synthesis. arXiv preprint arXiv:2301.13430 (2023)
39. Ye, Z., Jiang, Z., Ren, Y., Liu, J., He, J., Zhao, Z.: Geneface: Generalized and high-fidelity audio-driven 3d talking face synthesis. arXiv preprint arXiv:2301.13430 (2023)
40. Yi, R., Ye, Z., Zhang, J., Bao, H., Liu, Y.J.: Audio-driven talking face video generation with learning-based personalized head pose. arXiv preprint arXiv:2002.10137 (2020)
41. Yu, W., Fan, Y., Zhang, Y., Wang, X., Yin, F., Bai, Y., Cao, Y.P., Shan, Y., Wu, Y., Sun, Z., et al.: Nofa: Nerf-based one-shot facial avatar reconstruction. In: ACM SIGGRAPH 2023 Conference Proceedings. pp. 1–12 (2023)
42. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. Proceedings of the IEEE conference on computer vision and pattern recognition pp. 586–595 (2018)
43. Zhang, W., Cun, X., Wang, X., Zhang, Y., Shen, X., Guo, Y., Shan, Y., Wang, F.: Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. arXiv preprint arXiv:2211.12194 (2022)
44. Zhang, Z., Li, L., Ding, Y., Fan, C.: 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. pp. 3661–3670 (2021)
45. Zhou, H., Sun, Y., Wu, W., Loy, C.C., Wang, X., Liu, Z.: Pose-controllable talking face generation by implicitly modularized audio-visual representation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021)

46. Zhou, Y., Han, X., Shechtman, E., Echevarria, J., Kalogerakis, E., Li, D.: Makelttalk: speaker-aware talking-head animation. ACM Transactions On Graphics (TOG) **39**(6), 1–15 (2020)