Learning Dynamic Facial Radiance Fields for Few-Shot Talking Head Synthesis

Shuai Shen[1,2], Wanhua Li[1,2], Zheng Zhu[3], Yueqi Duan[4], Jie Zhou[1,2], and Jiwen Lu[1,2,*]

[1] Department of Automation, Tsinghua University, China
[2] Beijing National Research Center for Information Science and Technology, China
[3] PhiGent Robotics
[4] Department of Electronic Engineering, Tsinghua University, China
shens19@mails.tsinghua.edu.cn; li-wh17@tsinghua.org.cn;
zhengzhu@ieee.org; {duanyueqi, jzhou, lujiwen}@tsinghua.edu.cn

Abstract. Talking head synthesis is an emerging technology with wide applications in film dubbing, virtual avatars and online education. Recent NeRF-based methods generate more natural talking videos, as they better capture the 3D structural information of faces. However, a specific model needs to be trained for each identity with a large dataset. In this paper, we propose Dynamic Facial Radiance Fields (DFRF) for few-shot talking head synthesis, which can rapidly generalize to an unseen identity with few training data. Different from the existing NeRF-based methods which directly encode the 3D geometry and appearance of a specific person into the network, our DFRF conditions face radiance field on 2D appearance images to learn the face prior. Thus the facial radiance field can be flexibly adjusted to the new identity with few reference images. Additionally, for better modeling of the facial deformations, we propose a differentiable face warping module conditioned on audio signals to deform all reference images to the query space. Extensive experiments show that with only tens of seconds of training clip available, our proposed DFRF can synthesize natural and high-quality audio-driven talking head videos for novel identities with only 40k iterations. We highly recommend readers view our supplementary video for intuitive comparisons. Code is available in https://sstzal.github.io/DFRF/.

Keywords: few-shot talking head synthesis, neural radiance fields

1 Introduction

Audio-driven talking head synthesis is an ongoing research topic with a variety of applications including filmmaking, virtual avatars, video conferencing and online education [4,17,45,51,53,55]. Existing talking head generation methods can be roughly divided into 2D-based and 3D-based ones. Conventional 2D-based methods usually depend on GAN model [6, 11, 16] or image-to-image translation [12, 53–55]. However, due to the lack of 3D structure modeling, most of

*Corresponding author

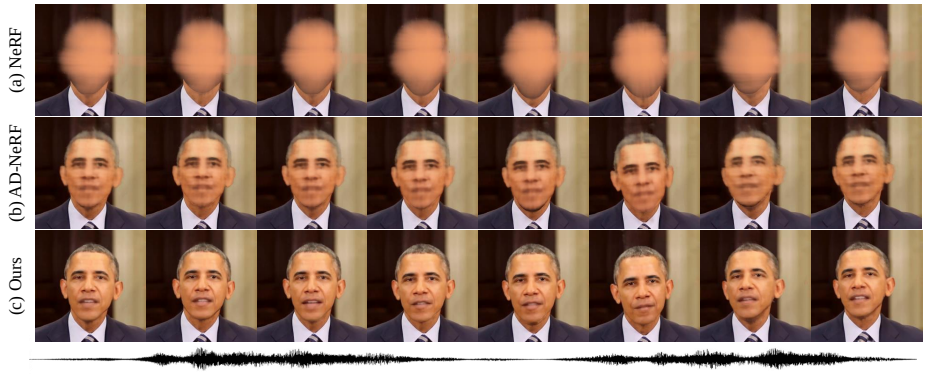

Fig. 1. We propose Dynamic Facial Radiance Fields (DFRF), a learning framework for few-shot talking head synthesis within a small number of training iterations. Given only a 15s video clip of Obama for 10k iterations training, our DFRF rapidly generalizes to this specific identity including the scene, and synthesizes photo-realistic talking head sequence as shown in row (c). In contrast, NeRF [27] and AD-NeRF [17] fail to produce plausible results in such a few-shot setting within limited training iterations.

these approaches struggle in generating vivid and natural talking styles. Another genre for talking head synthesis [4, 36, 39, 49] relies on the 3D morphable face model (3DMM) [2,40,57]. Benefit from the 3D-aware modeling, they can generate more vivid talking faces than 2D-based methods. Since the use of intermediate 3DMM parameters leads to some information loss, the audio-lip consistency of the generated videos may be affected [17].

More recently, the emerging Neural Radiance Fields (NeRF) based talking head methods [17, 27, 47] have achieved great performance improvement. They map audio features to a dynamic radiance field for talking portraits rendering without introducing extra intermediate representation. However, they directly encode the 3D geometry and appearance of a specific person into the radiance field, thereby failing to generalize to novel identities. A specific model needs to be trained for each novel identity with high computational cost. Moreover, a large training dataset is required, which cannot meet some practical scenarios where only a few data is available. As shown in Fig. 1, given only a 15s training clip, AD-NeRF [17] renders some blurry faces after 10k training iterations.

In this paper, we study this more challenging setting, few-shot talking head synthesis, for the aforementioned practical application scenarios. For an arbitrary new identity with merely a short training video clip available, the model should generalize to this specific person within a few iterations of fine-tuning. There are three key features of the few-shot talking head synthesis *i.e.* limited training video, fast convergence, and realistic generation results. To this end, we propose a Dynamic audio-driven Facial Radiance Field (DFRF) for few-shot talking head synthesis. A reference mechanism is designed to learn the generic mapping from a

few observed frames to the talking face with corresponding appearance (including the same identity, hairstyle and makeup). Specifically, with some 2D observations as references, the 3D query point can be projected back to the 2D image space of these references respectively and draw the corresponding pixel information to guide the following synthesis and rendering. A prior assumption for such projection operation is that two intersecting rays in 3D volume space should correspond to the same color [27, 29]. This conception holds for static scenes, yet talking heads are deformable objects and such naive warping may lead to some mismatch. We therefore introduce a differentiable face warping module for better modeling the facial dynamics when talking. This face warping module is realized as a 3D point-wise deformation field conditioned on audio signals to warp all reference images to the query space.

Extensive experiments show that our proposed DFRF can generate realistic and natural talking head videos with few training data and training iterations. Fig. 1 shows the visual comparison with NeRF [27] and AD-NeRF [17]. Given only a 15-second video clip of Obama for 10k training iterations, our proposed DFRF quickly generalizes to this specific identity and synthesizes photo-realistic talking head results. In contrast, NeRF and AD-NeRF fail to produce plausible results in such few-shot setting within limited training iterations. To summarize, we make the following contributions:

- We propose a dynamic facial radiance field conditioned on the 3D aware reference image features. The facial field can rapidly generalize to novel identities with only 15s clip for fine-tuning.
- For better modeling the face dynamics of talking head, we learn a 3D point-wise face warping module conditioned on audio signals for each reference image to warp it to the query space.
- The proposed DFRF can generate vivid and natural talking head videos using only a handful of training data with limited iterations, which far surpasses other NeRF-based methods under the same setting. We highly recommend readers view the supplementary videos for better comparisons.

2 Related Work

2D-Based Talking-Head Synthesis. Talking-head synthesis aims to animate portraits with given audios. 2D-based methods usually employ GANs [6,11,16,31] or image-to-image translation [12, 53–56] as the core technologies, and use some intermediate parameters such as 2D landmarks [5, 7, 11, 25, 55] to realize the synthesis task. There are also some works focusing on the few-shot talking head generation [12, 23, 26, 44, 51]. Zakharov *et al.* [51] propose a few-shot adversarial learning approach through pre-training high-capacity generator and discriminator via meta-learning. Wang *et al.* [44] realize one-shot talking head generation by predicting flow-based motion fields. Meshry *et al.* [26] disentangle the spatial and style information for few-shot talking head synthesis. However, since these 2D-based methods cannot grasp the 3D structure of head, the naturalness and vividness of the generated talking videos are inferior to the 3D-based methods.

3D-Based Talking-Head Synthesis. A series of 3D model-based methods [4, 9, 13, 19–21, 36, 37, 39] generate talking heads by utilizing 3D Morphable Models (3DMM) [2, 34, 40, 57]. Taking advantage of 3D structure modeling, these approaches can achieve more natural talking style than 2D methods. Representative methods [37, 39] have generated realistic and natural talking head videos. However, since their networks are optimized on a specific identity for idiosyncrasies learning, per-identity training on a large dataset is needed. Another common limitation is the information loss brought by the use of intermediate 3DMM parameters [2]. In contrast, our proposed method gets rid of such computationally expensive per-identity training settings while generating high-quality videos. More recently, the emerging NeRF [27] provides a new technique for 3D-aware talking head synthesis. Guo *et al.* [17] are the first to apply NeRF into the area of talking head synthesis and have achieved better visual quality. Yao *et al.* [47] further disentangle lip movements and personalized attributes. However both of them suffer in the few-shot learning setting.

Neural Radiance Fields. Neural Radiance Fields (NeRF) [27] store the information of 3D geometry and appearance in terms of voxel grids [35, 38] with a fully-connected network. The invention of this technology has inspired a series of following works. pi-GAN [3] proposes a generative model with NeRF as the backbone for static face generation while our method learns a dynamic radiance field. Since the original NeRF is designed for static scenes, some works try to extend this technique to the dynamic domain [14, 15, 29, 32, 41]. Gafni *et al.* [14] encode the expression parameters into the NeRF for dynamic faces rendering. [29, 32, 41] encode non-rigid scenes via ray bending into a canonical space. [45] represents face as compact 3D keypoints and performs keypoint driven animation. i3DMM [48] generates faces relying on geometry latent code. However, these methods need to optimize the model to every scene independently requiring a large dataset, while our method realizes fast generalization across identities based on easily accessible 2D reference images. There are also some other works that try to improve NeRF's generalization capabilities [42, 43, 50], yet their research are limited to static scenes.

3 Methodology

3.1 Problem Statement

Some limitations of existing talking head technologies hinder them from practical applications. 2D-based methods struggle to generate a natural talking style [39]. Classical 3D-based approaches have information loss due to the use of 3DMM intermediate representations [17]. NeRF-based ones synthesize superior talking head videos, however the computational cost is relatively high since a specific model needs to be trained for each identity. And a large dataset is required for training. We therefore focus on a more challenging setting for the talking head synthesis task. For an arbitrary person with merely a short training video clip available, a personalized audio-driven portrait animation model with high-quality synthesis results should be constructed within only a few iterations of

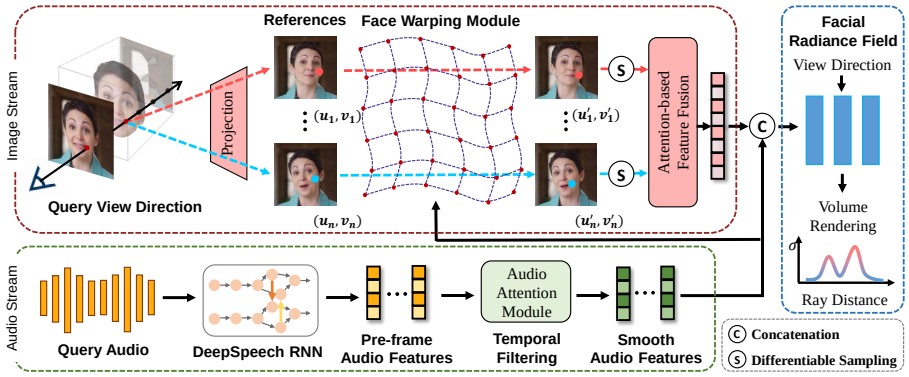

Fig. 2. Overview of the proposed Dynamic Facial Radiance Fields (DFRF).

fine-tuning. Three core features of this setting can be summarized as: limited training data, fast convergence and excellent generation effect.

To this end, we propose a Dynamic Facial Radiance Field (DFRF) for few-shot talking head synthesis. The image features are introduced as a condition to build a fast mapping from reference images to the corresponding facial radiance field. For better modeling the facial deformations, we further design a differentiable face warping module to warp reference images to the query space. Specifically, for fast convergence, a base model is firstly trained across different identities to capture the structure information of the head and establish a generic mapping from audio to lip motions. On this basis, efficient fine-tuning is performed to quickly generalize to a new target identity. In the following, we will detail these designs.

3.2 Dynamic Facial Radiance Field

The emerging NeRFs [27] provide a powerful and elegant framework for 3D scene representation. It encodes a scene into a 3D volume space with a MLP $\mathcal{F}_\theta$. The 3D volume can then be rendered into images by integrating colors and densities along camera rays [10, 28, 33]. Specifically, using $\mathcal{P}$ as the collection of all 3D points in the voxel space, with a 3D query point $p = (x, y, z) \in \mathcal{P}$ and a 2D view direction $d = (\theta, \phi)$ as input, this MLP infers the corresponding RGB color c and density σ, which can be formulated as $(c, \sigma) = \mathcal{F}_\theta(p, d)$.

In this work, we employ NeRF as the backbone for 3D-aware talking head modeling. The talking head task focuses on the audio-driven face animation. However, the original NeRF is designed for only static scenes. We therefore provide the missing deformation channel by introducing audio condition as shown in the audio stream of Fig. 2. We firstly use a pre-trained RNN-based DeepSpeech [18] module to extract the per-frame audio feature. For inter-frame consistency, a temporal filtering module [39] is further introduced to compute smooth

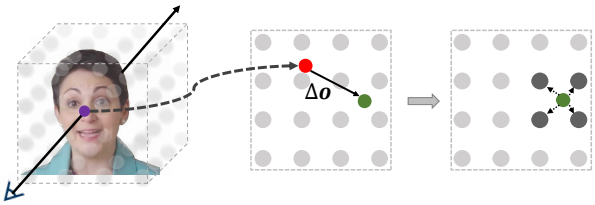

Fig. 3. Visualization of the differentiable face warping. A query 3D point (purple) is projected to the reference image space (red). Then an offset Δo is learned to warp it to the query space (green), where its feature is computed by bilinear interpolation.

audio features A, which can be denoted as the self-attention-based fusion of its neighbor audio features. Taking these audio feature sequences A as the condition, we can learn the audio-lip mapping. This audio-driven facial radiance field can be denoted as $(c, \sigma) = \mathcal{F}_\theta(p, d, A)$.

Since the identity information is implicitly encoded into the facial radiance field, and no explicit identity feature is provided when rendering, this facial radiance field is person specific. For each new identity, it needs to be optimized from scratch on a large dataset. This leads to expensive calculation costs and requires long training videos. To get rid of these restrictions, we design a reference mechanism to empower a well-trained base model to quickly generalize to new person categories, with only a short clip of the target person available. An overview of this reference-based architecture is shown in Fig. 2. Specifically, taken N reference images $M = \left\{M_n \in \mathbb{R}^{H\times W} | 1 \le n \le N\right\}$ and their corresponding camera position $\{T_n\}$ as input, a two-layer convolutional network is used to calculate their pixel aligned image features $F = \left\{F_n \in \mathbb{R}^{H\times W\times D} | 1 \le n \le N\right\}$ without down sampling. Feature dimension D is set as 128 in this work, and H, W indicates the height and width of an image respectively. The use of multiple reference images provides better multi-view informations. For a 3D query point $p = (x, y, z) \in \mathcal{P}$, we project it back to the 2D image spaces of these references using intrinsics $\{K_n\}$ and camera poses $\{R_n, T_n\}$ and get the corresponding 2D coordinate. Using $p_n^{ref} = (u_n, v_n)$ to denote the 2D coordinate in the n-th reference image, this projection can be formulated as:

$$p_n^{ref} = \mathcal{M}(p, K_n, R_n, T_n), \tag{1}$$

where $\mathcal{M}$ is the traditional mapping from world space to image space. These corresponding pixel-level features $\{F_n(u_n, v_n)\} \in \mathbb{R}^{N\times D}$ from N references are then sampled after a rounding operation and fused with an attention-based module [24] to get the final feature $\tilde{F} = Aggregation(\{F_n(u_n, v_n)\}) \in \mathbb{R}^D$. These feature grids contain rich information about identity and appearance. Using them as an additional condition for our facial radiance field makes the model possible to quickly generalize to a new face appearance from a few observed frames. This

dual-driven facial radiance field can be finally formulated as:

$$(c, \sigma) = \mathcal{F}_\theta \left(p, d, A, \tilde{F} \right). \tag{2}$$

3.3 Differentiable Face Warping

In Section 3.2, we project the query 3D point back to the 2D image spaces of these reference images as Eq. (1) to get the conditioned pixel features. This operation bases on the prior knowledge in NeRF that intersecting rays casting from different viewpoints should correspond to the same physical location and thus yield the same color [29]. This strict spatial mapping relationship holds for rigid scenes yet the talking face is dynamic. When speaking, the lip and other facial muscles moves according to the pronunciation. Applying Eq. (1) directly on a deformable talking face may result in the key points mismatch. For example, a 3D point near the corner of the mouth in the standard volume space is mapped back to the pixel space of a reference image. If the reference face shows a different mouth shape, the mapped point may fall away from the desired real mouth corner. Such inaccurate mapping results in incorrect pixel feature conditions from reference images, which further affects the prediction of deformations of talking mouth.

To tackle this limitation, we propose an audio-conditioned and 3D point-wise face warping module $\mathcal{D}_\eta$. It regresses offsets $\Delta o = (\Delta u, \Delta v)$ for every projected point p^{ref} under the specific deformations, just as shown in the image stream of Fig. 2. Specifically, $\mathcal{D}_\eta$ is realized as a deformation field with a three-layer MLP, where η is the learnable parameters. To regress the offset Δo, dynamics differences between the query image and these reference images need to be effectively exploited. The audio information A reflects the dynamics of the query image, while the deformations of the reference images can be seen through image features $\{F_n\}$ implicitly. We therefore take these two parts together with the query 3D point coordinate p as the input for $\mathcal{D}_\eta$. The process to predict the offset with the face warping module $\mathcal{D}_\eta$ can be formulated as:

$$\Delta o_n = \mathcal{D}_\eta(p, A, F_n(u_n, v_n)). \tag{3}$$

The predicted offset o_n is then added to the p_n^{ref} as shown in Fig. 3 to get the exact corresponding coordinate $p_n^{ref'}$ for the 3D query point p,

$$p_n^{ref'} = p_n^{ref} + \Delta o_n = (u_n', v_n'), \tag{4}$$

where $u_n' = u_n + \Delta u_n$ and $v_n' = v_n + \Delta v_n$.

Since the hard index operation $F_n(u_n{}', v_n{}')$ is not differentiable, the gradient cannot be back propagated to this warpping module. We therefore introduce a soft index function to realize the differentiable warpping, where the feature of each pixel is obtained through features interpolation of its surrounding points by bilinear sampling. In this way, the deformation field $\mathcal{D}_\eta$ and the facial radiance field $\mathcal{F}_\theta$ can be jointly optimized end to end. A visualization of this soft index

operation is shown in Fig. 3. For the green point, its pixel feature is computed through the features of its four nearest neighbours by bilinear interpolation. To better constrain the training process of this warping module, we introduce a regularization term L_r to limit the value of predicted offsets in a reasonable range to prevent distortions,

$$L_r = \frac{1}{N \cdot |\mathcal{P}|} \sum_{p \in \mathcal{P}} \sum_{n=1}^{N} \sqrt{\Delta u_n^2 + \Delta v_n^2}, \tag{5}$$

where $\mathcal{P}$ is the collection of all 3D points in the voxel space, and N is the number of reference images. Furthermore, we argue that the points with low density are more likely to be background areas that should have low deformation offset. In these regions, stronger regularization constraints should be imposed. For more reasonable constraint, we change the above L_r as:

$$L_r{}' = (1 - \sigma) \cdot L_r, \tag{6}$$

where σ indicates the density of these points. The dynamic facial radiance field can finally be formulated as:

$$(c, \sigma) = \mathcal{F}_\theta \left(p, d, A, \tilde{F}' \right), \tag{7}$$

where $\tilde{F}' = Aggregation(\{F_n(u_n{}', v_n{}')\})$.

With this face warping module, all reference images can be transformed to the query space for better modeling the talking face deformations. The ablation study in Section 4.2 has proven the effectiveness of this component in producing more accurate and audio-synchronized mouth movements.

3.4 Volume Rendering

The volume rendering is used to integrate the colors c and densities σ from Eq. (7) into face images. We treat the background, torso and neck parts together as the rendering 'background' and restore it frame by frame from the original videos. We set the color of the last point of each ray as the corresponding background pixel to render a natural background including the torso part. Here we follow the setting in the original NeRF, and the accumulated color C of a camera ray r under the condition of audio signal A and image features $\tilde{F}'$ is:

$$C \left(r; \theta, \eta, R, T, A, \tilde{F}' \right) = \int_{z_{near}}^{z_{far}} \sigma(t) \cdot c(t) \cdot T(t)\, dt, \tag{8}$$

where θ and η are the learnable parameters for the facial radiance field $\mathcal{F}_\theta$ and the face warping module $\mathcal{D}_\eta$ respectively. R is the rotation matrix and T is the translation vector. $T(t) = exp\left(- \int_{z_{near}}^{t} \sigma(r(s))\, ds \right)$ is the integral transmittance along camera ray, where z_{near} and z_{far} are the near and far bound of the camera ray. We follow the NeRF to design a MSE loss as $L_{MSE} = \|C - I\|^2$, where I is the ground truth color. Coupled with the regularization term in Eq. (6), the overall loss function can be formulated as:

$$L = L_{MSE} + \lambda \cdot L_r{}'. \tag{9}$$

3.5 Implementation Details

We train only one base radiance field across different identities from coarse to fine. In the coarse training stage, the facial radiance field $\mathcal{F}_\theta$ as Eq. (2) is trained under the supervision of L_{MSE} to grasp the structure of the head and establish a general mapping from audio to lip motions. Then we add the face warping module into training as Eq. (7) to jointly optimize the offset regression network $\mathcal{D}_\eta$ and the $\mathcal{F}_\theta$ end to end with the loss function L in Eq. (9).

For an arbitrary unseen identity with only a short training clip available, we only need tens of seconds of his/her speaking video for fine-tuning based on the well-trained base model. After short iterations of fine-tuning, the personalized mouth pronunciation patterns can be learned, and the rendered image quality is greatly improved. Then this fine-tuned model can be used for inference.

4 Experiments

4.1 Experimental Settings

Dataset. AD-NeRF [17] collects several high-resolution videos in natural scenes to better evaluate the performance in practical application. Following this practice, we collect 12 public videos with an average length of 3 minutes from 11 identities from the YouTube. The protagonists of these videos are all celebrities like news anchors, entrepreneurs or presidents. We resample all videos to 25 FPS and set the resolution as 512×512. We select three videos from different races and languages (English and Chinese), and combine them into a three-minute video to train the base model. For other videos, we split each of them into three training sets of the length of 10s, 15s and 20s. Then the remaining part is used as the test set. There is no overlap between the training set and the test set. All videos and the corresponding identities used in the following experiments are unseen when training the base model. These data will be released for reproduction.

Head Pose. Following the AD-NeRF, we estimate head poses based on Face2Face [40]. To get temporally smooth poses, we further apply the bundle adjustment [1] as a temporal filtering. The camera poses $\{R_n, T_n\}$ are the inverse of head poses, where R is the rotation matrix and T is the translation vector.

Metrics. We conduct performance evaluations through some quantitative metrics and visual results. Peak Signal-to-Noise Ratio (PSNR↑), Structure SIMilarity (SSIM↑) [46] and Learned Perceptual Image Patch Similarity (LPIPS↓) [52] are used as image quality metrics. PSNR tends to give higher scores to blurry images [29]. We therefore recommend the more representative perceptual metrics LPIPS. We further use the SyncNet (offset↓/confidence↑) [8] to measure the audio-visual synchronization. The SyncNet offset is better with smaller absolute value. Here we use the '↓' as a brief indication.

Training Details. Our code is based on PyTorch [30]. All experiments are performed on an RTX 3090. The coefficient λ in Eq. (9) is set as 5e-8. We train the base model with an Adam solver [22] for 300k iterations and then jointly train it with the offset regression network for another 100k iterations.

Reference	1		2		4		6	
Metric	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓
	31.03	**0.019**	31.19	**0.019**	**31.23**	**0.019**	**31.23**	0.020

Table 1. Quantitative comparisons with different numbers of reference images.

Method	NeRF [27]			AD-NeRF [17]			Ours		
	10s	15s	20s	10s	15s	20s	10s	15s	20s
PSNR↑	19.83	19.77	8.02	31.21	**31.32**	30.90	30.95	30.75	30.96
SSIM↑	0.773	0.781	0.003	0.948	0.947	**0.949**	0.948	0.947	**0.949**
LPIPS↓	0.237	0.239	1.058	0.039	0.041	0.040	**0.036**	**0.036**	**0.036**
SyncNet↓↑	-	-	-	15/1.313	-14/0.654	-5/0.932	0/3.447	0/4.105	**0/4.346**

Table 2. Method comparisons when using different lengths of training videos.

4.2 Ablation Study

The Number of Reference Images. In this work, we learn a generic rendering from arbitrary reference face images to talking head with the corresponding appearance (including identity, hairstyle and makeup). Here we perform experiments to investigate the performance gains from various reference face images. We select different numbers of references and fine-tune the base model for 10k iterations on 15s video clip respectively. Quantitative comparisons in Table 1 show that our method is robust to the number of reference images. According to results, we uniformly use four references in the following experiments.

Impact of the Length of Training Data. In this subsection, we investigate the impact of different amounts of training data. We fine-tune the proposed DFRF with 10s, 15s, and 20s training videos for 50k iterations. For fair comparisons, we train NeRF and AD-NeRF with the same data and iterations. It is worth noting that we have tried to pre-train NeRF and AD-NeRF across identities following DFRF. However since they lack the ability to generalize between different identities, such per-training fails to learn the general audio-lip mapping. Experimental results in Table 2 show that tens of seconds of data are insufficient for NeRF training. PSNR tends to give higher scores to blurry images [29], so we recommend LPIPS as more representative metrics for visual quality. In comparison, our method is able to acquire more prior knowledge about the general audio-lip mapping from the base model, thus achieving better audio-visual sync with limited training data. With only a 10s training video, the proposed DFRF can achieve superior 0.036 LPIPS and 3.447 SyncNet confidence, while AD-NeRF struggles in the lip-audio sync.

Effect of Differentiable Face Warping. In DFRF, we propose an audio conditioned differentiable face warping module for better modeling the dynamics of talking face. Here we conduct an ablation study to investigate the contribution of this component. Table 3 shows the generated results with and without warping module on two test sets. All models are fine-tuned on 15s videos for 50k

Method	Test Set A				Test Set B			
	PSNR↑	SSIM↑	LPIPS↓	SyncNet↓↑	PSNR↑	SSIM↑	LPIPS↓	SyncNet↓↑
GT	-	-	-	4/7.762	-	-	-	3/8.947
w/o	29.50	0.907	0.057	-1/4.152	28.98	**0.899**	0.104	-2/2.852
w	**29.66**	**0.911**	**0.053**	**0/4.822**	**29.14**	**0.899**	**0.101**	**0/4.183**

Table 3. Ablation study to investigate the contribution of the proposed differentiable face warping module. 'w' indicates the model equipped with the face warping module.

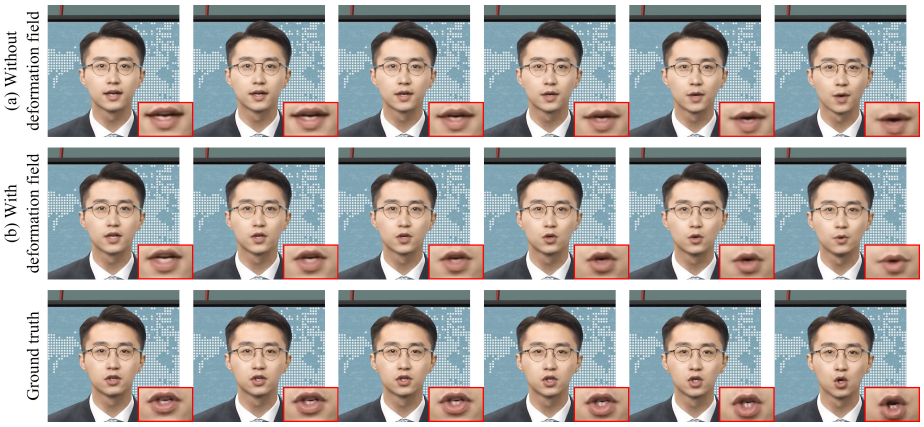

Fig. 4. Ablation study on the proposed face warping module. The ground truth sequence shows a pout-like expression. Generated results from the model equipped with the deformation field reproduce such pronunciation trend well in line (b), while results in line (a) hardly reflect such lip motions.

iterations. Without this module, the query 3D point cannot be mapped to the exact corresponding point in the reference image, especially in some areas with rich dynamics. Therefore, the dynamics of the speaking mouth are affected to some extent, which is reflected in the audio-visual sync (SyncNet score). In contrast, the model equipped with the deformation field can significantly improve the SyncNet confidence and the visual quality also has slight improvement. Fig. 4 further shows some visual results for more intuitive comparisons. In this video sequence, the ground truth shows a pout-like expression. The generated results (b) with the deformation field show such pronunciation trend well, while results in (a) hardly reflect this kind of lip motions.

4.3 Method Comparisons

Method Comparisons in the Few-shot Setting. In this section, we perform method comparisons on two test sets using a 15s training clip for different training iterations. Quantitative results in Table 4 show that our proposed method far

Method		Test Set A				Test Set B			
		PSNR↑	SSIM↑	LPIPS↓	SyncNet↓↑	PSNR↑	SSIM↑	LPIPS↓	SyncNet↓↑
Ground-truth		-	-	-	0/7.217	-	-	-	-1/7.762
NeRF	1k	16.88	0.708	0.198	-	14.69	0.397	0.442	-
	10k	13.98	0.531	0.338	-	15.24	0.396	0.427	-
	40k	15.87	0.556	0.306	-	15.91	0.405	0.394	-
AD-NeRF	1k	27.38	0.901	0.084	-15/0.136	27.61	0.863	0.115	14/0.798
	10k	29.14	0.931	0.057	-14/0.467	30.07	0.905	0.083	-2/0.964
	40k	29.45	0.936	0.039	-14/0.729	**30.72**	0.909	0.059	-2/1.017
Ours	1k	28.96	0.933	0.040	-1/2.996	29.05	0.892	0.076	0/3.157
	10k	29.33	0.935	0.043	0/4.246	29.68	0.905	0.063	0/4.038
	40k	**29.48**	**0.937**	**0.037**	**1/4.431**	30.44	**0.925**	**0.045**	**0/4.951**

Table 4. Method comparisons on two test sets using 15s training clip for different training iterations. More visual results can be seen in Fig. 5 and Fig. 6.

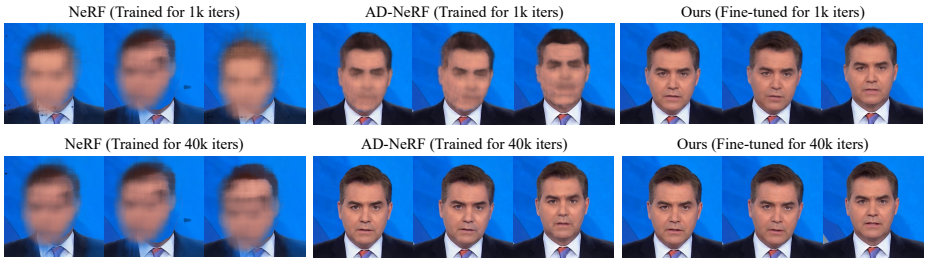

Fig. 5. Visual comparison using 15s training clip for different training iterations.

surpasses NeRF and AD-NeRF in the perceptual image quality metric LPIPS. PSNR tends to give higher scores to blurry images [29] which can be proved in the visualization in Fig. 6, so we recommend LPIPS as more representative metrics for visual quality. We also achieve higher audio-lip synchronization indicated by the SyncNet score while AD-NeRF nearly fails on this indicator. Fig. 5 visualizes the generated frames of the three methods. Under the same 1k training iterations, the visual quality of our method is far superior to others. When training for 40k iterations, the AD-NeRF achieves acceptable visual quality, however some face details are missing. The visual gap with our method can be seen obviously from the zoomed-in details in Fig. 6. We show two generated talking sequences driven by the same audio from our method and AD-NeRF with 15s training clip after 40k iterations in Fig. 6. Compared with the ground truth, our method shows more accurate audio-lip synchronization than AD-NeRF. For example, in the fifth frame, the rendered face from AD-NeRF opens the mouth wrongly. We zoom in some facial details for clearer comparison. It can be seen that our method has generated more realistic details such as sharper hair texture, more obvious wrinkles, brighter pupils and more accurate mouth shape. In

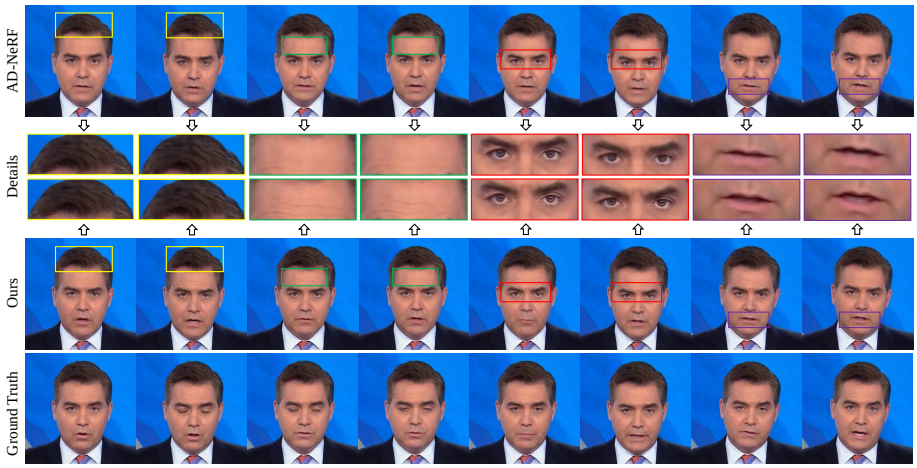

Fig. 6. Comparison with AD-NeRF using the same 15s training clip for 40k training iterations. We zoom in on some facial details for better visual quality comparison.

Method	Test Set A			Test Set B			Few-shot
	PSNR↑	LPIPS↓	SyncNet↓↑	PSNR↑	LPIPS↓	SyncNet↓↑	Method
Suwajanakorn *et al.* [37]	-	-	3/4.301	-	-	-	×
NVP [39]	-	-	-	-	-	-1/4.677	×
AD-NeRF [17]	33.20	0.032	0/5.289	33.85	0.028	0/4.200	×
Ours	**33.28**	**0.029**	**1/5.301**	**34.65**	**0.027**	**1/5.755**	✓

Table 5. Method comparisons with two non-NeRF based methods SO [37] and NVP [39] and the AD-NeRF [17] under the setting with more training data.

our supplementary video, we further add the visual comparison with AD-NeRF when it is trained to convergence (400k iterations).

Method Comparisons with More Training Data. Our DFRF is far superior to others in the few-shot learning setting. For more comprehensive evaluations, we further compare the DFRF with some recent high-performance non-NeRF 3D-based methods [37,39] and the AD-NeRF [17] with more training data (180s training clip). Since the source of [37,39] are not fully open, we follow the AD-NeRF to collect two test sets from the demos of [37,39] for method comparisons, and the results are shown in Table 5. Our method still surpasses others with long training clip up to 180s, since the proposed face warping module better models the talking face dynamics. Moreover, our DFRF is the only method that works in the few-shot learning setting. In the supplementary video, we further include more comparisons with 2D-based (non-NeRF based) methods.

Cross-Language Results. We further verify the performance of our method driven by audios with different languages and genders. We select four models trained with 15s training clips from different languages (source), then conduct

Source-Target	Same Identity	English (Male)	Chinese (Male)	Russian (Male)	French (Female)	Spanish (Female)	German (Female)
English (Male)	-3/5.042	-2/3.805	-2/4.879	-1/4.118	-2/3.019	-2/4.986	-1/4.820
Chinese (Male)	0/4.486	-2/3.029	-2/3.534	-3/4.206	-3/3.931	-3/4.085	-2/4.494
Russian (Male)	-1/4.431	-2/2.831	-2/4.397	-2/5.109	-3/4.307	-1/5.011	-1/5.008
French (Male)	-2/4.132	-2/3.193	-3/3.383	-3/4.088	-2/3.339	-2/3.728	-1/3.529

Table 6. SyncNet scores under the cross language setting.

inference with driven audios cross six languages and different genders (target). We also list the self-driven (source and target are from the same identity) results (the second column) for reference. SyncNet (offset/ confidence)($\downarrow$ / $\uparrow$) scores in Table 6 shows that our method produces reasonable lip-audio synchronization in such cross language setting.

4.4 Applications and Ethical Considerations

The talking head synthesis technique can be used in a variety of practical scenarios, including correcting pronunciation, re-dubbing, virtual avatars, online education, electronic game making and providing speech comprehension for hearing impaired people. However, the talking head technology may bring some potential misuse issues. We are committed to combating these malicious behaviors and advocate more attention to the active application of this technology. We support those organizations that devote themselves to identifying fake defamatory videos, and are willing to provide them with the generated videos to expand the training set for automatic identification technology. Meanwhile, any individual or organization should obtain our permission before using our code, and it is recommended to use a watermark to indicate the generated video.

5 Conclusion

In this paper, we have proposed a dynamic facial radiance field for few-shot talking head synthesis. We employ audio signals coupled with 3D-aware image features as the condition for fast generalizing to novel identities. To better model the mouth motions of talking head, we further learn an audio-conditioned face warping module to deform all reference images to the query space. Extensive experiments show the superiority of our method in generating natural talking videos with limited training data and iterations.

Acknowledgments. This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFA0700802, in part by the National Natural Science Foundation of China under Grant 62125603 and Grant U1813218, in part by a grant from the Beijing Academy of Artificial Intelligence (BAAI).

References

1. Andrew, A.M.: Multiple view geometry in computer vision. Kybernetes (2001)
2. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: Annual Conference on Computer Graphics and Interactive Techniques (1999)
3. Chan, E.R., Monteiro, M., Kellnhofer, P., Wu, J., Wetzstein, G.: pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In: CVPR (2021)
4. Chen, L., Cui, G., Liu, C., Li, Z., Kou, Z., Xu, Y., Xu, C.: Talking-head generation with rhythmic head motion. In: ECCV (2020)
5. Chen, L., Maddox, R.K., Duan, Z., Xu, C.: Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In: CVPR (2019)
6. Christos Doukas, M., Zafeiriou, S., Sharmanska, V.: Headgan: Video-and-audio-driven talking head synthesis. arXiv (2020)
7. Chung, J.S., Jamaludin, A., Zisserman, A.: You said that? In: BMVC (2017)
8. Chung, J.S., Zisserman, A.: Out of time: automated lip sync in the wild. In: ACCV (2016)
9. Cudeiro, D., Bolkart, T., Laidlaw, C., Ranjan, A., Black, M.J.: Capture, learning, and synthesis of 3d speaking styles. In: CVPR (2019)
10. Curless, B., Levoy, M.: A volumetric method for building complex models from range images. In: Annual Conference on Computer Graphics and Interactive Techniques (1996)
11. Das, D., Biswas, S., Sinha, S., Bhowmick, B.: Speech-driven facial animation using cascaded gans for learning of motion and texture. In: ECCV (2020)
12. Eskimez, S.E., Zhang, Y., Duan, Z.: Speech driven talking face generation from a single image and an emotion condition. TMM (2021)
13. Fried, O., Tewari, A., Zollhöfer, M., Finkelstein, A., Shechtman, E., Goldman, D.B., Genova, K., Jin, Z., Theobalt, C., Agrawala, M.: Text-based editing of talking-head video. TOG (2019)
14. Gafni, G., Thies, J., Zollhofer, M., Nießner, M.: Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In: CVPR (2021)
15. Gao, C., Shih, Y., Lai, W.S., Liang, C.K., Huang, J.B.: Portrait neural radiance fields from a single image. arXiv (2020)
16. Gu, K., Zhou, Y., Huang, T.: Flnet: Landmark driven fetching and learning network for faithful talking facial animation synthesis. In: AAAI (2020)
17. Guo, Y., Chen, K., Liang, S., Liu, Y., Bao, H., Zhang, J.: AD-NeRF: Audio driven neural radiance fields for talking head synthesis. In: ECCV (2021)
18. Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., et al.: Deep speech: Scaling up end-to-end speech recognition. arXiv (2014)
19. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. NeurIPS (2015)
20. Ji, X., Zhou, H., Wang, K., Wu, W., Loy, C.C., Cao, X., Xu, F.: Audio-driven emotional video portraits. In: CVPR (2021)
21. Karras, T., Aila, T., Laine, S., Herva, A., Lehtinen, J.: Audio-driven facial animation by joint end-to-end learning of pose and emotion. TOG (2017)
22. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv (2014)
23. Kumar, N., Goel, S., Narang, A., Hasan, M.: Robust one shot audio to video generation. In: CVPRw (2020)

24. Locatello, F., Weissenborn, D., Unterthiner, T., Mahendran, A., Heigold, G., Uszkoreit, J., Dosovitskiy, A., Kipf, T.: Object-centric learning with slot attention. arXiv (2020)
25. Lu, Y., Chai, J., Cao, X.: Live speech portraits: real-time photorealistic talking-head animation. TOG (2021)
26. Meshry, M., Suri, S., Davis, L.S., Shrivastava, A.: Learned spatial representations for few-shot talking-head synthesis. arXiv (2021)
27. 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)
28. Niemeyer, M., Mescheder, L., Oechsle, M., Geiger, A.: Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In: CVPR (2020)
29. Park, K., Sinha, U., Barron, J.T., Bouaziz, S., Goldman, D.B., Seitz, S.M., Martin-Brualla, R.: Nerfies: Deformable neural radiance fields. In: ICCV (2021)
30. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. NeurIPS (2019)
31. 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: ACM MM (2020)
32. Pumarola, A., Corona, E., Pons-Moll, G., Moreno-Noguer, F.: D-nerf: Neural radiance fields for dynamic scenes. In: CVPR (2021)
33. Seitz, S.M., Dyer, C.R.: Photorealistic scene reconstruction by voxel coloring. IJCV (1999)
34. Shang, J., Shen, T., Li, S., Zhou, L., Zhen, M., Fang, T., Quan, L.: Self-supervised monocular 3d face reconstruction by occlusion-aware multi-view geometry consistency. In: ECCV (2020)
35. Sitzmann, V., Zollhöfer, M., Wetzstein, G.: Scene representation networks: Continuous 3d-structure-aware neural scene representations. arXiv (2019)
36. Song, L., Wu, W., Qian, C., He, R., Loy, C.C.: Everybody's talkin': Let me talk as you want. arXiv (2020)
37. Suwajanakorn, S., Seitz, S.M., Kemelmacher-Shlizerman, I.: Synthesizing obama: learning lip sync from audio. TOG (2017)
38. Tewari, A., Fried, O., Thies, J., Sitzmann, V., Lombardi, S., Sunkavalli, K., Martin-Brualla, R., Simon, T., Saragih, J., Nießner, M., et al.: State of the art on neural rendering. In: Computer Graphics Forum (2020)
39. Thies, J., Elgharib, M., Tewari, A., Theobalt, C., Nießner, M.: Neural voice puppetry: Audio-driven facial reenactment. In: ECCV (2020)
40. Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., Nießner, M.: Face2face: Real-time face capture and reenactment of rgb videos. In: CVPR (2016)
41. Tretschk, E., Tewari, A., Golyanik, V., Zollhofer, M., Lassner, C., Theobalt, C.: Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In: ICCV (2021)
42. Trevithick, A., Yang, B.: Grf: Learning a general radiance field for 3d representation and rendering. In: ICCV (2021)
43. Wang, Q., Wang, Z., Genova, K., Srinivasan, P.P., Zhou, H., Barron, J.T., Martin-Brualla, R., Snavely, N., Funkhouser, T.: Ibrnet: Learning multi-view image-based rendering. In: CVPR (2021)
44. Wang, S., Li, L., Ding, Y., Fan, C., Yu, X.: Audio2head: Audio-driven one-shot talking-head generation with natural head motion. arXiv (2021)

45. Wang, T.C., Mallya, A., Liu, M.Y.: One-shot free-view neural talking-head synthesis for video conferencing. In: CVPR (2021)
46. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. TIP (2004)
47. Yao, S., Zhong, R., Yan, Y., Zhai, G., Yang, X.: Dfa-nerf: Personalized talking head generation via disentangled face attributes neural rendering. arXiv (2022)
48. Yenamandra, T., Tewari, A., Bernard, F., Seidel, H.P., Elgharib, M., Cremers, D., Theobalt, C.: i3dmm: Deep implicit 3d morphable model of human heads. In: CVPR (2021)
49. Yi, R., Ye, Z., Zhang, J., Bao, H., Liu, Y.J.: Audio-driven talking face video generation with learning-based personalized head pose. arXiv (2020)
50. Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelnerf: Neural radiance fields from one or few images. In: CVPR (2021)
51. Zakharov, E., Shysheya, A., Burkov, E., Lempitsky, V.: Few-shot adversarial learning of realistic neural talking head models. In: ICCV (2019)
52. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
53. Zhang, X., Wu, X., Zhai, X., Ben, X., Tu, C.: Davd-net: Deep audio-aided video decompression of talking heads. In: CVPR (2020)
54. Zhou, H., Liu, Y., Liu, Z., Luo, P., Wang, X.: Talking face generation by adversarially disentangled audio-visual representation. In: AAAI (2019)
55. Zhou, Y., Han, X., Shechtman, E., Echevarria, J., Kalogerakis, E., Li, D.: Makelttalk: speaker-aware talking-head animation. TOG (2020)
56. Zhu, H., Huang, H., Li, Y., Zheng, A., He, R.: Arbitrary talking face generation via attentional audio-visual coherence learning. IJCAI (2020)
57. Zollhöfer, M., Thies, J., Garrido, P., Bradley, D., Beeler, T., Pérez, P., Stamminger, M., Nießner, M., Theobalt, C.: State of the art on monocular 3d face reconstruction, tracking, and applications. In: Computer Graphics Forum (2018)