Pacific Graphics (2023)
R. Chaine, Z. Deng, and M. H. Kim (Editors)

Short Paper

Generalizable Dynamic Radiance Fields For Talking Head Synthesis With Few-shot

R. Dang , S. Wang and H. Wang

Tsinghua Shenzhen International Graduate School, Tsinghua University

Abstract

Audio-driven talking head generation has wide applications in virtual games, hosts, online meetings, etc. Recently, great achievements have been made in synthesizing talking heads based on neural radiance fields. However, the existing few-shot talking head synthesis methods still suffer from inaccurate deformation and lack of visual consistency. Therefore, we propose a Generalizable Dynamic Radiance Field (GDRF), which can rapidly generalize to unseen identities with few-shot. We introduce a warping module with 3D constraints to act in feature volume space, which is identity adaptive and exhibits excellent shape-shifting abilities. Our method can generate more accurately deformed and view consistent target images compared to previous methods. Furthermore, we map the audio signal to 3DMM parameters by applying an LSTM network, which helps get long-term context and generate more continuous and natural video. Extensive experiments demonstrate the superiority of our proposed method.

CCS Concepts

• ***Computing methodologies*** → ***Reconstruction; Animation;*** *Shape representations;*

1. Introduction

Audio-driven talking head generation has wide application and development prospects in VR/AR, virtual hosting, online video conferences, etc. Synthesizing continuous and high-fidelity audio-driven talking head videos is highly challenging due to the difficulty of using audio signals to control facial expressions and poses. Currently, many works have been proposed to synthesize talking heads. With the rise of Generative Adversarial Networks (GAN) [GPAM*20], the GANs-based driven face method [KLA19, DBSB20, TET*20a] has been widely used. However, these 2D-based approaches cannot generate realistic and vivid talking heads due to the lack of 3D structure of the head.

Recently, the neural radiance field (NeRF) [MST*21], which models the geometry and appearance of a specific person as a function, has shown excellent performance in synthetic photo-realistic talking head [GCL*21, YZY*22]. However, these methods struggle to generalize to new scenes, requiring a large number of input frames and training for a long time, which limits the application in practice. There are some few-shot NeRFs [WWG*21, YYTK21, CXZ*21] that generalize to new scenes with few inputs by extracting pixel-aligned features from the reference images as prior information of the radiance field. However, these methods are only suitable for synthesizing static scenes. On the other hand, some works [GTZN21, PSB*21, ZAB*22] have been proposed to construct dynamic radiance fields by introducing 3D deformation fields, and thousands of frames of images are still required.

Figure 1: *Synthetic results of talking head on different methods for fine-tuning 5k iterations with only 5s video input. (a) NeRF [MST*21]. (b) ADNeRF [GCL*21]. (c) DFRF [SLZ*22]. (d) Ours.*

Therefore, it is contradictory to construct a 3D neural deformation field with few-shot. As shown in Figure 1, the faces synthesized by [MST*21] and [GCL*21] are incomplete and blurred due to few input frames and short optimization time. While DFRF [SLZ*22] generalizes novel identities with 15s clips by proposing a 2D warp-

DOI: 10.2312/pg.20231274

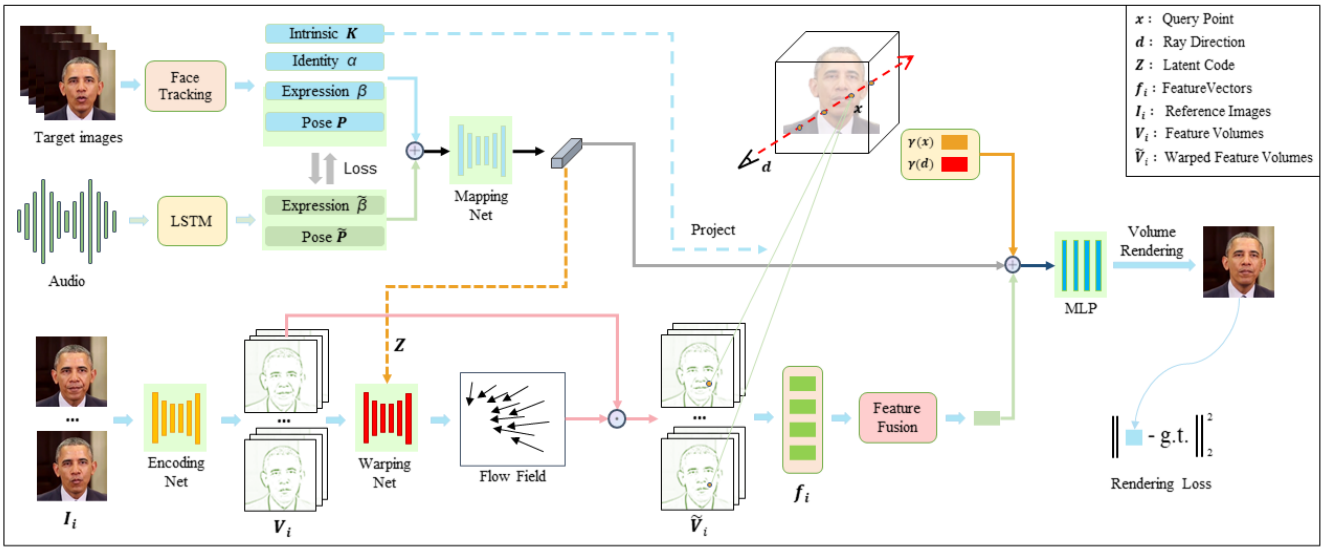


Figure 2: *The framework of our proposed generalizable dynamic radiance field (GDRF).*

ing module at the image level, the synthesized image still has artefacts such as inconsistency and blurriness. In contrast, we introduce a warping network with 3D constraints to conduct warping in the deep feature space instead of the image space, which exhibits a more robust and precise deformation performance. As shown in Figure 1, our proposed warping strategy synthesizes sharper facial details (such as teeth, wrinkles, and mouth contours) than DFRF. In addition, prior works [GCL*21, SLZ*22] directly extract audio features to implicitly drive the target face, which cannot learn long-term context, so the generated video lacks continuity between frames and has a time delay relative to the input video.

To address the abovementioned problems, we propose a generalizable dynamic radiance field, synthesizing the high-fidelity audio-driven talking head with a few-shot. Our method achieves more accurate lip movement and higher-quality generated images and videos than previous methods. First, We extract feature volumes from reference images as priors on the radiance field for fast generalization to different identities. Then, we introduce a warping network with 3D constraints that perform warping in the feature volume space. This helps infer precise lip movement and generate visually continuous images with only a 5s clip. Unlike previous 2D warping methods [RLC*21, HZL22], our warping module is constrained by the 3D geometry in the radiance field, which makes the warped frame view consistent. Besides, our deformation model has lower degrees of freedom, making it easier to infer deformation with few-shot than the 3D deformation field. More importantly, the warping module acts on the feature volume space instead of the image space, making it faster to build the deformation field of the new identity and improve the quality of generated images [ZLW*22]. Subsequently, based on an LSTM module, we map audio features to expressions and poses as the warping network's motion descriptors. This helps generate view-consistent and natural target talking head videos [YYZ*20]. Finally, a multi-layer perceptron (MLP) is exploited to estimate attributes of spatial query points, and the classic volume rendering method is used to generate the target image. Experimental visualization results and quantitative indicators verify the superiority of the proposed method.

In general, our work mainly makes the following contributions:

- We propose a generalizable dynamic radiance field for synthesizing visually continuous and high-precision audio-driven talking head with only 5s video input;
- We introduce an audio-guided warping module with 3D constraints performing on the feature volume, which is identity adaptive and shows accurate deformation in short iterations with few-shot;
- We utilize an LSTM module to map audio into expression and pose to guide further the warping network, which contains long-term context, making the facial movement more continuous and natural.

2. Related Work

Speeching head generation driven by audio. Audio-driven speeching head generation requires convincing facial motion learned from audio signals. There are some works [ZLL*19, SLZ*22] learn lip motion directly from audio features, while Prajwal et al. [PMNJ20] focus on lip movements by utilizing the Wav2Lip model. Some works [GZZ*23, YZC*22, ZZH*21, ZNF*21, ZSW*21] utilize GAN-like architecture to generate more generalizable interim results to reinforce overall quality. Transformer-based models [WQZ*23, MWH*23, WLDY22, SZW*22] also perform greatly benefited from well-designed architecture and the perception ability of cross-attention. Besides, Ji et al. [JZW*21] focus on disentangling audio into emotion-related parts and content-related parts to achieve better results, especially around the mouth area. Nevertheless, because most of the abovementioned methods abandon parametric representation, these approaches face challenges and lack long-term context.

In order to avoid the interference of irrelevant factors in the audio and enhance the continuity of the generated target video, several

works [TET*20a, CCL*20, LCC21, ZLDF21, ZCW*23, XZZ*23, ZFC*23] explicitly control the target face synthesis by mapping audio to face parameters, such as landmarks and 3DMMs [BV99]. Wang et al. [WLD*21] generate videos with spatial and temporal continuity by introducing key points estimated by motion field. Huang et al. [HZL22] learn more continuous and precise lip movements by 3DMM parameters learned from a transformer model. Therefore, smooth and realistic videos are easier to synthesize by utilizing 3DMM parameters.

Neural radiance field with few-shot. Recently, much work has been done on reconstructing static scenes from a few images. For instance, PixelNeRF [YYTK21] extracts the pixel-aligned feature from one or few reference images as a prior condition and synthesizes new perspectives. IBRNet [WWG*21] realizes the generalization of NeRF for unseen scenes through a generic view interpolation function. What's more, MVSNeRF [CXZ*21] performs 3D convolution on the cost volume constructed by the MVS-Net [YLL*18] and then obtains an encoding volume containing per-voxel neural features, thereby realizing the generalization of different scenes with only three images. Besides, there are also some works [HPX*22, ZZSC22] proposing parametric nerf models to achieve photo-realistic face synthesis with free-view images. However, these works are only suitable for reconstructing static scenes and cannot achieve the generalization of dynamic scenes.

Dynamic radiance field. NeRF is only applicable to the reconstruction of static scenes. In order to extend NeRF to dynamic scenes, [PSB*21, PSH*21, PCPMMN21, LXW*22] deforms the query point from the observation space to the canonical space by optimizing a backward 3D deformation field. NerFace [GTZN21] and RigNeRF [AXS*22] control the expression and pose of the human face through a morphable model to model the dynamics of the face. Moreover, SNARF [CZB*21] proposes forward deformation fields that improve generalization and show better deformation performance. However, these mentioned methods all require many input frames, so building a dynamic field suitable for a small number of inputs is challenging.

3. Method

The overall framework of our proposed model is shown in Figure 2. Unlike previous methods, we map the audio features into 3DMM [BV99] parameters based on the LSTM module, which is essential to generate precise and continuous lip movements. Subsequently, we introduce a warping network with 3D constraints guided by the 3DMM parameters to warp the feature volumes. Below are the implementation details.

3.1. Audio Mapping Network

Learning precise lip movements directly from audio features requires many frames and lacks long-term context, resulting in discontinuous and delayed video generation. In order to extract time dependencies and reduce latency, we use an LSTM module to learn the facial parameters from the audio and then use the parameters to drive the target image, which helps generate a more natural and smooth audio-driven video.

Based on 3DMM, each face can be represented as a linear combination of shape and texture vector, and the face shapes S can be represented as:

$$\mathbf{S} = \overline{\mathbf{S}} + \alpha \mathbf{B}_{id} + \beta \mathbf{B}_{exp} \tag{1}$$

where $\overline{\mathrm{S}}$ represents the average face, and B_{id}, B_{exp} are the PCA basis for identity and expression seperately. Where $\alpha \in \mathrm{R}^{80}$ and $\beta \in \mathrm{R}^{64}$ are the coefficient of identity and expression. Besides, the head poses $P \in \mathrm{R}^6$ is the inverse of camera pose $[R, T]$, where R and T are the rotation and translation matrix. In this paper, we utilize the face tracking method [TZS*16] to estimate the facial parameters of input frames.

Firstly, to obtain the time dependence, we use a fixed-length window to extract the Mel-frequency cepstral coefficients (MFCC) features of adjacent audio frames as the inputs of the LSTM network. Subsequently, we input the MFCC feature sequence of each frame into the LSTM network and then output the predicted facial expression and pose parameters sequence, which can be formulated as follows:

$$\begin{cases} [\tilde{\beta}_t, \tilde{P}_t, h_t, c_t] = L\left(E\left(a_t\right), h_{t-1}, c_{t-1}\right) \\ \mathbf{a} = \{a_1, a_2, \ldots, a_T\} \\ \tilde{\beta} = \left\{\tilde{\beta}_1, \tilde{\beta}_2, \ldots, \tilde{\beta}_T\right\} \\ \tilde{\mathbf{P}} = \left\{\tilde{P}_1, \tilde{P}_2, \ldots, \tilde{P}_T\right\} \end{cases} \tag{2}$$

where E represents the module that encodes the MFCC features and L represents the LSTM network. Besides, each time, c_t and h_t represent the cell and hidden states. Where a, $\tilde{\beta}$ and $\tilde{P}$ represent the input MFCC feature sequence, predicted expression coefficient sequence and pose sequence by LSTM separately, and T is the length of the window.

In this paper, we fine-tune the mapping from audio features to facial parameters based on a pre-trained LSTM model trained on a large video corpus. In addition, we use a mean squared error loss between the ground truth and the predicted parameters to constrain the training process, which can be denoted as:

$$\mathcal{L}_{\beta,P} = \|\beta - \tilde{\beta}\|^2 + \lambda_1 \|P - \tilde{P}\|^2 \tag{3}$$

where β and P are the ground-truth expression and pose parameters estimated by face2face [TZS*16]. We set the weight factor $\lambda_1 = 0.2$, and extra loss constraints are added between consecutive frames to enhance the continuity between frames:

$$\begin{aligned} \mathcal{L}_c = \lambda_2 \left[\sum_{t=1}^{T-1} \left(\tilde{\beta}_{t+1} - \tilde{\beta}_t\right)^2\right] \\ + \lambda_3 \left[\sum_{t=1}^{T-1} \left(\tilde{P}_{t+1} - \tilde{P}_t\right)^2\right] \end{aligned} \tag{4}$$

where we set $\lambda_2 = 0.01$ and $\lambda_3 = 0.001$. Therefore, the total loss function of the LSTM network is formulated as follows:

$$\mathcal{L}_l = \mathcal{L}_{\beta,P} + \mathcal{L}_c \tag{5}$$

3.2. Warping Network with 3D Constraints

Feature extraction. Firstly, randomly select N frames from the input video as reference images. After that, a feature encoding net-

Table 1: *Performance comparison of different methods on videos of different lengths.*

Method	NeRF			AD-NeRF			DFRF			Ours		
	3s	5s	10s	3s	5s	10s	3s	5s	10s	3s	5s	10s
PSNR ↑	14.36	15.80	14.28	24.76	23.38	23.89	28.61	29.18	29.28	29.18	29.26	**29.32**
SSIM ↑	0.337	0.377	0.337	0.807	0.767	0.795	0.909	0.912	0.912	0.912	0.915	**0.916**
LPIPS ↓	0.691	0.540	0.607	0.145	0.176	0.164	0.067	0.065	0.064	0.062	**0.060**	**0.060**
SyncNet ↑	-	-	-	0.918	0.780	0.724	2.760	3.273	3.325	3.153	3.900	**4.037**

work E_c is used to extract the feature volume V_i of each reference image I_i, which can be formulated as:

$$V_i = E_c(I_i), (i = 1, 2, \cdots, N) \tag{6}$$

Semantic mapping. It is recognized that audio has a strong correlation with lip movements. However, it is also necessary to consider head movements to make the talking head more natural. According to Eq.(1), the position of the face mesh points is determined by the expression and identity coefficients, which represent identity and expression coefficients associated with lip movements. Besides, the identity parameters are related to the person's speaking style, so the identity parameter can help synthesize high-quality images. So we concatenate the relevant parameters and utilize a semantic mapping network M to map them into high-dimensional motion descriptor z:

$$z = M(\alpha \oplus \tilde{\beta} \oplus \tilde{P}) \tag{7}$$

where α, $\tilde{\beta}$ and $\tilde{P}$ are the identity parameter, predicted expression parameter, and head pose, respectively. Where $\oplus$ denotes the concatenation operation. The descriptor z is further used to guide the warping network by AdaIN [HB17] operator:

$$\text{AdaIN}(m_i; z) = \gamma^z \left(\frac{m_i - \mu(m_i)}{\sigma(m_i)} \right) + \beta^z \tag{8}$$

where m_i is the feature map of each convolutional layer, and $\mu(\cdot)$, $\sigma(\cdot)$ are the average and variance calculation of the statistic m_i separately. Where γ^z and β^z represent the affine parameter, calculated by performing an affine transformation on the latent code z, respectively.

Warping network. Afterwards, we employ a warping network W to estimate the deformations between feature volumes V_i and target feature volumes $\tilde{V}_i$. Guided by the latent code z, the warping module generates a flow field $w = W(V_i, z)$, which denotes the coordinate offset between the source and the target. Based on w, we obtain the warped feature volumes $\tilde{V}_i$ by interpolated sampling S:

$$\tilde{V}_i = S(V_i, W(V_i, z)) \tag{9}$$

3.3. Modeling Facial Radiance Field

Subsequently, we construct a radiance field to reconstruct the target image based on pixel-aligned feature vector [YYTK21].

Feature aggregation. For the sampling point x, we project it into the reference image coordinate system through the intrinsic matrix K and camera pose $[R, T]$. After that, the warped pixel-aligned feature vector f_i corresponding to the reference image is calculated by bilinear interpolation:

$$f_i = B_{il}(P_{roj}(x, K_i, R_i, T_i), \tilde{V}_i), (i = 1, 2, \cdots, N) \tag{10}$$

where P_{roj} represents the projection from the world coordinate system to the reference image coordinate system, and B_{il} represents the bilinear interpolation operator. Then we aggregate the N feature vectors $\{f_1, f_2 \ldots, f_N\}$ with an attention model [LWU*20] to form the final warped feature vector f.

Volume properties prediction. Finally, the facial radiance field f_θ is realized as a multi-layer perceptron (MLP). In addition to the spatial coordinates x and the viewing direction d of the query point, we further add the latent code z and the aggregated feature vector f to the radiance field. Then we estimate density σ and color c of the query point by f_θ:

$$(\sigma, c) = f_\theta(\gamma(x), \gamma(d), z, f) \tag{11}$$

where $\gamma(\cdot)$ is the position encoding employed by NeRF.

3.4. Volume Rendering

Color prediction. After that, we calculate the color of each ray $r(t) = o + td$ by volume rendering:

$$\begin{aligned} C(r) &= \int_{d_n}^{d_f} c(r(t), d)\sigma(r(t))T(t)dt, \\ T(t) &= \exp\left(- \int_{d_n}^{t} \sigma(r(s))ds \right) \end{aligned} \tag{12}$$

where o and d are the starting point and direction of the ray. Where $T(t)$, d_n, and d_f represent the cumulative transmittance along the camera ray and the near and far boundaries of the ray, respectively. Following DFRF, we render the background, neck, and shoulders as the image's background, making the talking head more natural.

Loss Function. We use the mean squared error between the ground truth colour $C_g(r)$ and the rendered colour $C(r)$ to compute the loss:

$$\mathcal{L}_r = \sum_{r \in R} \| C_g(r) - C(r) \|_2^2 \tag{13}$$

Where r represents the collection of training rays for each batch. Finally, the total loss function is as follows:

$$\mathcal{L} = \mathcal{L}_r + \lambda_l \cdot \mathcal{L}_l \tag{14}$$

4. Experiments

4.1. Implementation Details

In this section, we present the experimental details and demonstrate the superiority of the proposed method.

Dataset. We select face videos with a duration between 40s and 50s, containing different languages. Then, process the video to a

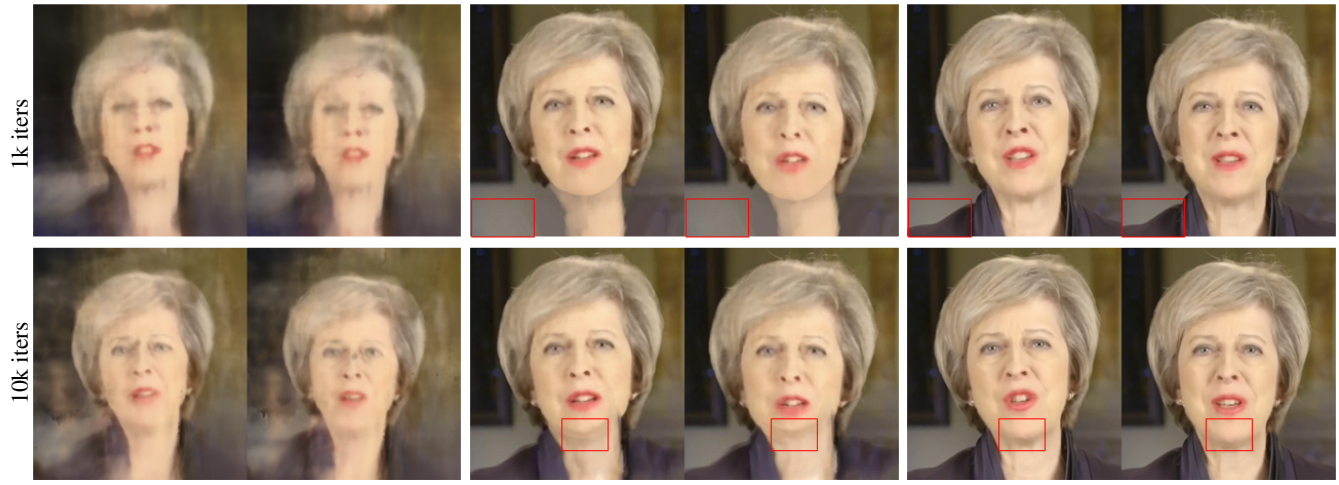


Figure 3: *Comparison of visualization results of different methods for fine-tuning 1k and 10k iterations with 5s input.*

resolution of 512×512 and a frame rate of 25 FPS. We select several videos from the dataset to train a basic model, including different languages. To verify the generalization, we divide the video of unseen identities into 3s, 5s, and 10s, and the remaining video length is used for testing.

Table 2: *Performance comparison of different methods on different iterations.*

Method		PSNR↑	SSIM↑	LPIPS↓	SyncNet↑
NeRF	1k	17.03	0.647	0.566	-
	5k	17.28	0.640	0.528	-
	10k	17.05	0.627	0.526	-
ADNeRF	1k	22.36	0.825	0.234	0.110
	5k	22.07	0.822	0.231	0.399
	10k	24.55	0.824	0.215	0.732
DFRF	1k	28.45	0.885	0.122	1.989
	5k	28.74	0.886	0.113	2.650
	10k	28.85	0.894	0.106	3.165
Ours	1k	28.59	0.889	0.090	2.664
	5k	29.04	0.902	0.083	3.548
	10k	**29.15**	**0.903**	**0.076**	**3.765**

Table 3: *Metric comparisons with non-NeRF methods Synthesizing Obama [SSKS17] and NVP [TET*20b] on two test sets (A and B) collected from the demos of these two methods, respectively.*

Method	SyncNet↑	
	Test set A	Test set B
Synthesizing Obama	-	4.713
NVP	4.286	-
Ours	5.230	5.441

Setup. The experiment is conducted on NVIDIA Tesla V100 GPUs and trained end-to-end. We use the Adam optimizer [KB14] to train the basic model for 300k steps with an initial learning rate of 0.0001. We set the number of reference images to 4 frames so that in the case of limited computing resources, the whole model has the best performance, which has been verified in the ablation experiments.

Evaluation metrics. To evaluate the quality of synthesized images, we utilize peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM), and learned perceptual image patch similarity (LPIPS) as the evaluation metric. In addition, SyncNet [CZ16] is used to evaluate the synchronization quality between audio and visual.

4.2. Overall Comparison Results

In this section, we choose NeRF and NeRF-based audio-driven talking head synthesis methods as the baseline. Nerface is the method of reconstructing 4D facial avatars by constructing dynamic radiance fields, but it requires long-time videos to train for new identities. ADNeRF and DFRF are the methods of constructing the audio-driven radiance fields, and DFRF is able to generalize to unseen identities with few-shot.

Different training video lengths. To verify the performance of our proposed method with few-shot , we divide the unseen identity face video into 3s,5s, and 10s. Then, we fine-tune the pre-trained models for 15k iterations with videos of different lengths separately. To keep it fair, we use the same data to optimize the same steps based on the basic model released by DFRF. Since NeRF and ADNeRF cannot be generalized to different identities, the same data is directly used for training. The quantitative results of the experiment are shown in Table 1. From the results in Table 1, we know that due to the small number of input frames, NeRF and ADNeRF have poor synthesis performance. Compared with the DFRF, our proposed method performs better on the input of each length. Since the PSNR score favours blurry images [PSB*21] and cannot represent the sensory quality, we pay more attention to the LPIPS score and Syncnet confidence.

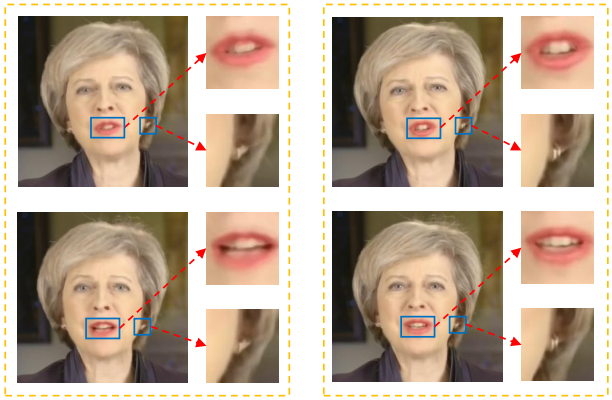
Different iteration steps. In this section, to compare the impact of different optimization times, we choose another 5s training data to iterate 1k, 5k, and 10k steps on different methods separately. The quantification results on the test set are shown in Table 2. Quantitative results demonstrate the superior performance of our method in short iterations. Especially in LPIPS and SyncNet scores, there

is a significant improvement compared to DFRF. In Table 3, we follow AD-NeRF and DFRF to compare with non-NeRF methods NVP [TET*20b] and Synthesizing Obama [SSKS17] to show our superiority. We train our model for 40k iterations. Notice that both of them are not few-shot methods. According to Figure 3, NeRF and ADNeRF methods fail to generate complete and clear images due to insufficient input images and short training time, especially in the torso and face contour. In contrast, our proposed method generates acceptable images for 1k iterations and high-quality images for 10k iterations. What's more, We zoom in on the details of the DFRF and our visualization results for 10k iterations, as shown in Figure 4. Figure 4 shows that our proposed method synthesizes sharper high-frequency details such as tooth edges and earring lines. In contrast, DFRF has poor synthesis quality at high-frequency details.

Speech-driven experiments in different languages. To verify our method's performance in different languages, we also selected Chinese and French test sets for verification. Similarly, a 5s video is taken from the test sets as input. In Table 4, we show the SyncNet [CZ16] scores for target videos in different languages driven by source audio in different languages. The same id in the second column indicates that the source and target are from the same identity. The source and target in other columns come from different identities. The quantitative results show that our proposed method has excellent performance among different language generalizations.

Table 4: *Comparison of driver performance in different languages.*

Target \ Source	Same id	English	Chinese	French
English	4.221	4.089	3.804	2.997
Chinese	3.821	3.007	3.545	3.129
French	3.478	3.209	2.866	3.187



(a) DFRF (b) Ours

Figure 4: *Detailed comparison of generation images on DFRF and our method. It can be seen that our proposed method has sharper generated images compared to DFRF.*

4.3. Ablation Study

The impact of the number of reference images. In this part, to choose the appropriate number of reference images, we set the number of reference images to 2, 4, 6 and 8 frames, respectively, and fine-tune 10k steps on the same test set. The quantitative comparison is shown in Table 5. The results show that our model is stable when the number of reference images is 4. Although there are higher PSNR and SSIM scores when the number of reference images is set to 6, we choose to set it to 4, considering computation consumption.

Table 5: *Comparison of results with different numbers of reference image inputs.*

Number	2	4	6	8
PSNR ↑	29.08	29.15	**29.16**	29.14
SSIM ↑	0.901	0.903	**0.904**	0.902
LPIPS ↓	0.077	**0.076**	**0.076**	0.077

The impact of different modules. To realize the face's deformation, we introduce a deformation network with 3D constraints performing on the feature volume. In addition, to improve the quality of the generated video, we apply an LSTM network to map audio features into 3DMM parameters to get long-term context. To verify the effectiveness of each module, we leverage the same data set to train three different basic models, namely "WarpNet", "WrapNet+LS", and our complete model. After that, we use these three basic models to fine-tune 15k steps for the same 5s video, and the experimental results are shown in Table 6.

In Table 6, the DFRF method is the baseline for comparison. The "WarpNet" donates using the same audio feature as DFRF to guide our proposed warping network instead of the 3DMM parameters learned from the LSTM module. The experimental results in the first and second rows demonstrate the superiority of the proposed warping module. Obviously, our proposed warping module outperforms DFRF on every measurement index. Besides, the "WrapNet+LS" means introducing the LSTM module to map features into expression and pose parameters and then using expression and pose parameters as the motion descriptor to guide the warping network. Compared with the "WarpNet" model, the significant improvement of the SyncNet score shows that LSTM helps generate smoother and more natural videos. Our complete model denotes that the "WarpNet+LS" model incorporates the identity parameters, which show the best performance and verify the importance of identity parameters for generating target faces.

Table 6: *Performance comparison of different models.*

Method	PSNR↑	SSIM↑	LPIPS↓	SyncNet↑
DFRF	29.18	0.912	0.065	3.273
WarpNet	29.25	0.914	0.062	3.342
WarpNet+LS	**29.28**	0.914	**0.060**	3.858
Ours	29.26	**0.915**	**0.060**	**3.900**

5. CONCLUSION

In this paper, we have proposed a generalizable dynamic radiance field for talking head synthesis with few-shot. We apply the warping network with 3D constraints to warp the feature volume extracted from reference images to eliminate the mismatch between

the target and source images. By inputting warped feature vectors and speech information as priors, the radiance field can well generalize to different identities. Further, based on the LSTM network, we map the audio signal to the 3DMM parameters to get long-term audio context, which makes the generated video more continuous and natural. A series of experiments demonstrate the superiority of our method in synthesizing talking heads with few-shot.

Applications and broader impact. Our work achieves generalizable dynamic neural radiance field reconstruction for speeching heads with only a few inputs, making it a reality to widely put speeching head synthesis technology into practical applications, including digital avatars, virtual anchors, games, online conferences, etc. However, we should also use this technology within the scope of legal ethics to prevent the misuse and disclosure of personal information.

Acknowledgment. This research was funded through National Key Research and Development Program of China (Project No. 2022YFB36066), in part by the Shenzhen Science and Technology Project under Grant (JCYJ20220818101001004, JSGG20210802153150005).

References

[AXS*22] ATHAR S., XU Z., SUNKAVALLI K., SHECHTMAN E., SHU Z.: Rignerf: Fully controllable neural 3d portraits. In *CVPR* (2022). 3

[BV99] BLANZ V., VETTER T.: A morphable model for the synthesis of 3d faces. In *SIGGRAPH* (1999). 3

[CCL*20] CHEN L., CUI G., LIU C., LI Z., KOU Z., XU Y., XU C.: Talking-head generation with rhythmic head motion. In *ECCV* (2020), Springer. 3

[CXZ*21] CHEN A., XU Z., ZHAO F., ZHANG X., XIANG F., YU J., SU H.: Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *ICCV* (2021). 1, 3

[CZ16] CHUNG J. S., ZISSERMAN A.: Out of time: automated lip sync in the wild. In *ACCV* (2016), Springer. 5, 6

[CZB*21] CHEN X., ZHENG Y., BLACK M. J., HILLIGES O., GEIGER A.: Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *ICCV* (2021). 3

[DBSB20] DAS D., BISWAS S., SINHA S., BHOWMICK B.: Speech-driven facial animation using cascaded gans for learning of motion and texture. In *ECCV* (2020), Springer. 1

[GCL*21] 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 *ICCV* (2021). 1, 2

[GPAM*20] GOODFELLOW I., POUGET-ABADIE J., MIRZA M., XU B., WARDE-FARLEY D., OZAIR S., COURVILLE A., BENGIO Y.: Generative adversarial networks. *CACM* (2020). 1

[GTZN21] GAFNI G., THIES J., ZOLLHOFER M., NIESSNER M.: Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *CVPR* (2021). 1, 3

[GZZ*23] GUAN J., ZHANG Z., ZHOU H., HU T., WANG K., HE D., FENG H., LIU J., DING E., LIU Z., ET AL.: Stylesync: High-fidelity generalized and personalized lip sync in style-based generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 1505–1515. 2

[HB17] HUANG X., BELONGIE S.: Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV* (2017). 4

[HPX*22] HONG Y., PENG B., XIAO H., LIU L., ZHANG J.: Headnerf: A real-time nerf-based parametric head model. In *CVPR* (2022). 3

[HZL22] HUANG R., ZHONG W., LI G.: Audio-driven talking head generation with transformer and 3d morphable model. In *ACM MM* (2022). 2, 3

[JZW*21] JI X., ZHOU H., WANG K., WU W., LOY C. C., CAO X., XU F.: Audio-driven emotional video portraits. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2021), pp. 14080–14089. 2

[KB14] KINGMA D. P., BA J.: Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014). 5

[KLA19] KARRAS T., LAINE S., AILA T.: A style-based generator architecture for generative adversarial networks. In *CVPR* (2019). 1

[LCC21] LU Y., CHAI J., CAO X.: Live speech portraits: real-time photorealistic talking-head animation. *TOG* (2021). 3

[LWU*20] LOCATELLO F., WEISSENBORN D., UNTERTHINER T., MAHENDRAN A., HEIGOLD G., USZKOREIT J., DOSOVITSKIY A., KIPF T.: Object-centric learning with slot attention. *NeurIPS* (2020). 4

[LXW*22] LIU X., XU Y., WU Q., ZHOU H., WU W., ZHOU B.: Semantic-aware implicit neural audio-driven video portrait generation. In *European Conference on Computer Vision* (2022), Springer, pp. 106–125. 3

[MST*21] MILDENHALL B., SRINIVASAN P. P., TANCIK M., BARRON J. T., RAMAMOORTHI R., NG R.: Nerf: Representing scenes as neural radiance fields for view synthesis. *CACM* (2021). 1

[MWH*23] 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. *arXiv preprint arXiv:2301.01081* (2023). 2

[PCPMMN21] PUMAROLA A., CORONA E., PONS-MOLL G., MORENO-NOGUER F.: D-nerf: Neural radiance fields for dynamic scenes. In *CVPR* (2021). 3

[PMNJ20] 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* (2020), pp. 484–492. 2

[PSB*21] 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). 1, 3, 5

[PSH*21] PARK K., SINHA U., HEDMAN P., BARRON J. T., BOUAZIZ S., GOLDMAN D. B., MARTIN-BRUALLA R., SEITZ S. M.: Hypernerf: a higher-dimensional representation for topologically varying neural radiance fields. *TOG* (2021). 3

[RLC*21] REN Y., LI G., CHEN Y., LI T. H., LIU S.: Pirenderer: Controllable portrait image generation via semantic neural rendering. In *ICCV* (2021). 2

[SLZ*22] SHEN S., LI W., ZHU Z., DUAN Y., ZHOU J., LU J.: Learning dynamic facial radiance fields for few-shot talking head synthesis. In *ECCV* (2022), Springer. 1, 2

[SSKS17] SUWAJANAKORN S., SEITZ S. M., KEMELMACHER-SHLIZERMAN I.: Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (ToG) 36*, 4 (2017), 1–13. 5, 6

[SZW*22] SUN Y., ZHOU H., WANG K., WU Q., HONG Z., LIU J., DING E., WANG J., LIU Z., HIDEKI K.: Masked lip-sync prediction by audio-visual contextual exploitation in transformers. In *SIGGRAPH Asia 2022 Conference Papers* (2022), pp. 1–9. 2

[TET*20a] THIES J., ELGHARIB M., TEWARI A., THEOBALT C., NIESSNER M.: Neural voice puppetry: Audio-driven facial reenactment. In *ECCV* (2020), Springer. 1, 3

[TET*20b] 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* (2020), Springer, pp. 716–731. 5, 6

[TZS*16] THIES J., ZOLLHOFER M., STAMMINGER M., THEOBALT C., NIESSNER M.: Face2face: Real-time face capture and reenactment of rgb videos. In *CVPR* (2016). 3

[WLD*21] WANG S., LI L., DING Y., FAN C., YU X.: Audio2head: Audio-driven one-shot talking-head generation with natural head motion. *arXiv preprint arXiv:2107.09293* (2021). 3

[WLDY22] 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* (2022), vol. 36, pp. 2531–2539. 2

[WQZ*23] WANG J., QIAN X., ZHANG M., TAN R. T., LI H.: Seeing what you said: Talking face generation guided by a lip reading expert. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 14653–14662. 2

[WWG*21] 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). 1, 3

[XZZ*23] XU C., ZHU J., ZHANG J., HAN Y., CHU W., TAI Y., WANG C., XIE Z., LIU Y.: High-fidelity generalized emotional talking face generation with multi-modal emotion space learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 6609–6619. 3

[YLL*18] YAO Y., LUO Z., LI S., FANG T., QUAN L.: Mvsnet: Depth inference for unstructured multi-view stereo. In *ECCV* (2018). 3

[YYTK21] YU A., YE V., TANCIK M., KANAZAWA A.: pixelnerf: Neural radiance fields from one or few images. In *CVPR* (2021). 1, 3, 4

[YYZ*20] 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). 2

[YZC*22] YIN F., ZHANG Y., CUN X., CAO M., FAN Y., WANG X., BAI Q., WU B., WANG J., YANG Y.: Styleheat: One-shot high-resolution editable talking face generation via pre-trained stylegan. In *European conference on computer vision* (2022), Springer, pp. 85–101. 2

[YZY*22] YAO S., ZHONG R., YAN Y., ZHAI G., YANG X.: Dfa-nerf: Personalized talking head generation via disentangled face attributes neural rendering. *arXiv preprint arXiv:2201.00791* (2022). 1

[ZAB*22] ZHENG Y., ABREVAYA V. F., BÜHLER M. C., CHEN X., BLACK M. J., HILLIGES O.: Im avatar: Implicit morphable head avatars from videos. In *CVPR* (2022). 1

[ZCW*23] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 8652–8661. 3

[ZFC*23] ZHONG W., FANG C., CAI Y., WEI P., ZHAO G., LIN L., LI G.: Identity-preserving talking face generation with landmark and appearance priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 9729–9738. 3

[ZLDF21] ZHANG Z., LI L., DING Y., FAN C.: Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *CVPR* (2021). 3

[ZLL*19] ZHOU H., LIU Y., LIU Z., LUO P., WANG X.: Talking face generation by adversarially disentangled audio-visual representation. In *AAAI* (2019). 2

[ZLW*22] ZHANG J., LI X., WAN Z., WANG C., LIAO J.: Fdnerf: Few-shot dynamic neural radiance fields for face reconstruction and expression editing. In *SIGGRAPH* (2022). 2

[ZNF*21] ZHANG C., NI S., FAN Z., LI H., ZENG M., BUDAGAVI M., GUO X.: 3d talking face with personalized pose dynamics. *IEEE Transactions on Visualization and Computer Graphics* (2021). 2

[ZSW*21] 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/CVF conference on computer vision and pattern recognition* (2021), pp. 4176–4186. 2

[ZZH*21] ZHANG C., ZHAO Y., HUANG Y., ZENG M., NI S., BUDAGAVI M., GUO X.: Facial: Synthesizing dynamic talking face with implicit attribute learning. In *Proceedings of the IEEE/CVF international conference on computer vision* (2021), pp. 3867–3876. 2

[ZZSC22] ZHUANG Y., ZHU H., SUN X., CAO X.: Mofanerf: Morphable facial neural radiance field. In *ECCV* (2022), Springer. 3