AED-NeRF: Audio-Driven and Emotion-Editing Dynamic Neural Radiance Fields for Expressive Talking Face Avatar

Lu Ping[1,2], Song Li[3], Shi Wenzhe[1,2], Lin Zonghao[3], Ling Jun[3]

(1. State Key Laboratory of Mobile Network and Mobile Multimedia Technology, Shenzhen 518055, China;
2. ZTE Corporation, Shenzhen 518057, China;
3. Shanghai Jiao Tong University, Shanghai 200240, China)

DOI: 10.12142/ZTECOM.202601010

https://kns.cnki.net/kcms/detail/34.1294.TN.20260228.1556.002.html, published online February 28, 2025

Manuscript received: 2024-09-06

Abstract: While neural radiance field (NeRF) methods have shown promising results in generating talking faces, existing studies primarily focus on the correlation between avatars and driving sources. However, these studies often overlook emotion modeling, resulting in the generation of emotionless or unnatural facial animations. In response, this paper introduces an audio-driven and emotion-editing dynamic NeRF (AED-NeRF) approach, designed for the real-time generation of expressive talking face avatars driven by audio inputs. Specifically, we integrate audio features into a grid-based NeRF to compensate for the lack of a deformation channel, successfully capturing lip dynamics and enabling end-to-end generation from audio-driven sources to talking face avatars. Emotion labels, comprising emotion categories and intensity levels, guide the proposed NeRF framework to implicitly model visual emotions, allowing for explicit control and editing of facial expressions. Extensive qualitative and quantitative experiments validate the effectiveness and advantages of our proposed method, demonstrating its ability to achieve real-time, photo-realistic talking face avatar generation across different audio and emotion scenarios.

Keywords: talking face avatar; neural radiance fields; AED-NeRF

Citation (Format 1): Lu P, Song L, Shi W Z, et al. AED-NeRF: audio-driven and emotion-editing dynamic neural radiance fields for expressive talking face avatar [J]. ZTE Communications, 2026, 24(1): 72 - 80. DOI: 10.12142/ZTECOM.202601010
Citation (Format 2): P. Lu, L. Song, W. Z. Shi, et al., "AED-NeRF: audio-driven and emotion-editing dynamic neural radiance fields for expressive talking face avatar," *ZTE Communications*, vol. 24, no. 1, pp. 72 - 80, Mar. 2026. doi: 10.12142/ZTECOM.202601010.

1 Introduction

With the rapid evolution of deep learning[1] and generative modeling[2], talking face avatars are undergoing unprecedented development[3 - 8] and have been progressively integrated into our visual experiences, such as virtual video conferences, film redubbing, and digital human representation. Despite these advancements, talking face generation encounters numerous challenges in practical applications.

Image-based methods generate talking face avatars by employing techniques including image-to-image translation[9 - 11] and generative adversarial networks (GANs) [12 - 14]. Nevertheless, the absence of 3D perception tends to result in flat and unrealistic visual results. Model-based approaches explicitly construct 3D talking faces based on intermediate representations such as facial landmarks[15], coefficients[16] and vertices[4]. While leveraging 3D face modeling produces higher-quality results, cumulative errors and information loss during the intermediate representation prediction can lead to semantic mismatches between lip movements and audio cues.

Recently, the emergence of neural radiance fields (NeRF)[17] has provided a novel framework for talking face generation. NeRF-based methods can render realistic talking face avatars at high resolutions from novel views with reduced training data[5, 18 - 19]. However, the inference speed of the vanilla NeRF is insufficient to meet the real-time requirements of audio-driven talking face avatars in practical applications. Moreover, existing works fail to fully implement emotional modeling for talking face avatars, resulting in emotionless or unnatural human faces[20 - 22].

In this paper, we propose an audio-driven and emotion-editing dynamic NeRF (AED-NeRF) for real-time and expressive talking face avatar generation. Our method consists of three processing modules and two NeRF models. In the processing pipeline, we introduce an audio processing module, an emotion encoder, and a pose estimation module, where the audio encoder extracts features from audio sequences, and the emotion encoder encodes explicit emotion labels based on

This work was supported by ZTE Industry-University-Institute Cooperation Funds under Grant No. IA20230921015.

their category and intensity, obtaining emotion features. We employ an off-the-shelf method[23 - 24] to estimate 3D head pose for additional spatial control. Besides, we employ two NeRF models to render the head and torso separately by taking target identity video sequences, synchronized audio sequences, and emotion labels as inputs. In the modeling parts, AED-NeRF utilizes spatial features, audio features, and emotional features as inputs to neural radiance fields, implicitly modeling identity as volume density and RGB color. During the inference stage, given arbitrary head pose sequences, driving audio, and emotion labels, AED-NeRF performs volume rendering[25] according to the predicted volume density and color learned in the training stage, generating an expressive 3D avatar matching the driving source in real time.

Our contributions are summarized as follows:

• We introduce audio and emotional features to compensate for the lack of a deformation channel in the grid NeRF, implicitly modeling head dynamics and enabling end-to-end talking face avatar generation.

• We consider implicit emotion modeling in talking face avatars with emotion labels for diverse emotional expressions, guiding NeRF to implicitly model facial expressions and explicitly control the emotional editing of talking face avatars during the inference stage.

• Extensive experiments demonstrate that AED-NeRF can generate photorealistic, expressive talking face avatars in real time under different audio and emotional conditions.

2 Related Work

1) Image-based talking face generation. Image-based methods generate 2D talking face avatars using image-to-image translation or GANs. Zhou et al.[26] proposed a disentangled audio-visual system to disentangle identity and audio content through adversarial learning, improving lip synchronization for 2D talking face avatars. Das et al.[27] employed cascaded GANs to separately learn general lip motion and identity-specific texture. Zhou et al.[11] animated a single image with an audio clip by predicting landmark displacement from disentangled audio content and identity. Though these methods work well for stylized facial images, they have difficulty in generating realistic human face avatars due to the lack of 3D perception information.

2) Model-based talking face generation. Model-based methods explicitly generate 3D talking faces based on intermediate facial representations such as landmarks, coefficients and vertices. Kumar et al.[6] utilized long short-term memory (LSTM) to learn the mapping from driving audio sources to lip landmarks, and then generated pixel-to-pixel Obama avatars via UNet. Thies et al.[16] proposed a general audio-to-expression network to predict the expression coefficients of a 3D face model based on audio features, and a UNet-based neural rendering network to render talking face avatars from expression coefficients, thus enabling cross-identity audio driving. Richard et al.[28] designed a categorical latent space using 3D face vertices as the intermediate representation based on cross-modality loss. This space disentangles audio-correlated and audio-uncorrelated information and thus results in reasonable movements in audio-uncorrelated facial regions. Though model-based methods can generate high-quality talking face avatars, they suffer from problems including complex pipelines, expensive data labels, and information loss of driving source.

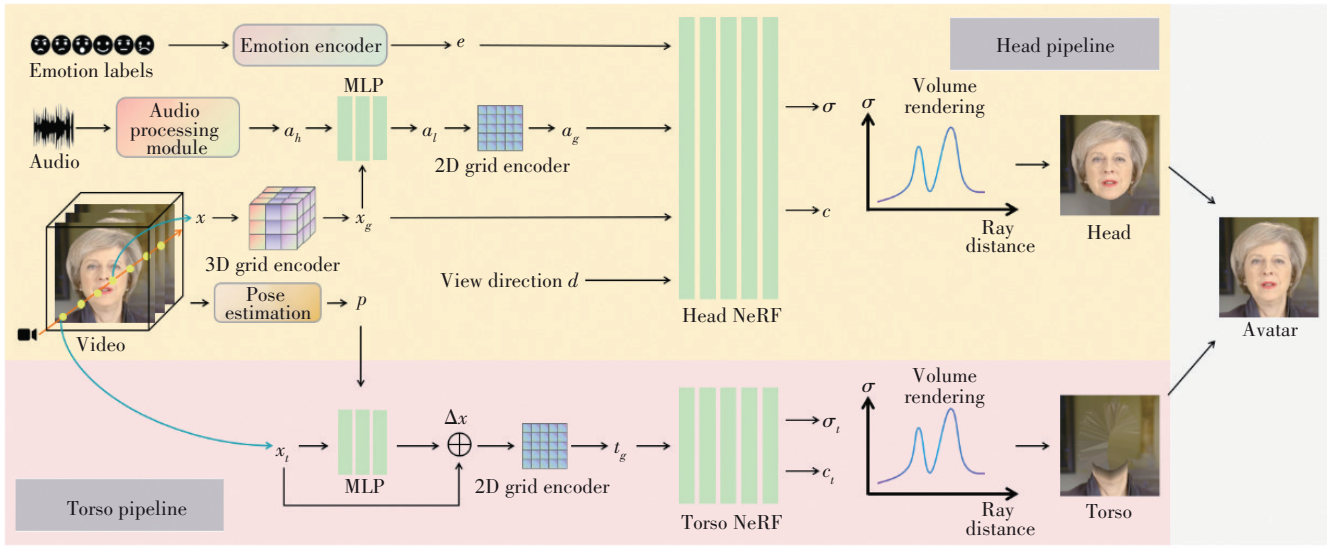
3) NeRF-based talking face generation. Neural radiance fields[17] have achieved great success in natural rendering and provide a new implementation for end-to-end talking face avatar generation. Gafni et al.[29] introduced NeRF to the field of talking face generation and proposed the first dynamic face radiance fields. They trained multilayer perceptron (MLP) conditioned on latent codes based on face coefficients and camera poses reconstructed by a face tracker, realizing face reconstruction and pose control of talking face avatar. Guo et al.[5] proposed audio-driven neural radiance fields which take DeepSpeech[30] audio features as conditional input to MLP and individually model head part and torso part of talking face avatars. Hong et al.[31] designed a parametrized general model for representing faces under different views, expressions and lighting, and significantly improved the rendering speed of NeRF via integrating a 2D neural rendering strategy into it. Shen et al.[18] conditioned NeRF on 2D face images to learn the face prior and fine-tune the face radiance fields with few identity images to generalize to a new identity rapidly. However, aforementioned NeRF-based methods struggle to meet real-time requirements in practical application due to the slow rendering speed.

4) Emotional talking face generation. Though talking face avatar generation has made significant progress in recent years, most works focus on the correlation between digital avatars and driving sources (e.g., audio, text, etc.) while neglecting emotion modeling, which leads to emotionless or unnatural human faces. Therefore, some researchers have turned to the study of emotion modeling for expressive talking face avatars. Eskimez et al.[10] improved emotional expressions by encoding emotion categories and designing an emotion discriminator to supervise the training. Ji et al.[32] proposed an implicit emotion displacement learner to modify facial dynamics for realistic emotion patterns. Tan et al.[33] extracted emotion embeddings from audio as queries and utilized a memory network to retrieve the best-matching expressions for talking face avatars. However, NeRF-based emotion modeling has not been fully explored. We show a simple but effective method to integrate emotion into NeRF modeling in Section 3.

3 Methods

3.1 Overview

In this section, we present our AED-NeRF framework in Fig. 1. The inputs to the network include target identity video sequences, synchronized audio sequences, and emotion la-



MLP: multilayer perceptron NeRF: neural radiance fields

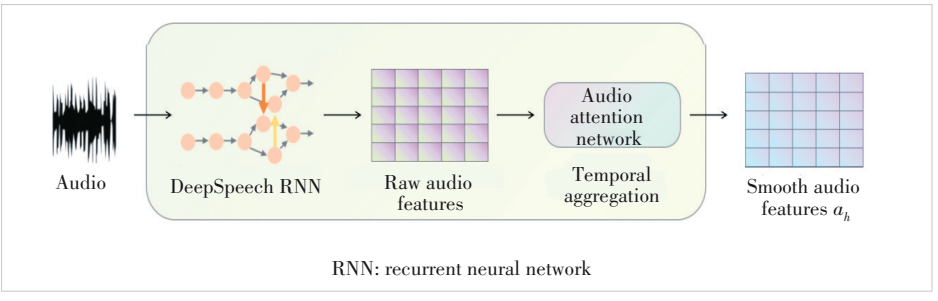
Figure 1. An overview of AED-NeRF framework

bels. We utilize an off-the-shelf method[23 - 24] to estimate 3D human poses for further spatial features. We adopt an audio encoder to extract audio features from the audio sequences and a one-hot encoder to encode explicit emotion labels according to their category and intensity to obtain emotion features. Then, AED-NeRF takes spatial, audio, and emotion features as the inputs to the neural radiance fields and implicitly models the identity, which is represented by volume density and RGB colors. In the inference stage, given an arbitrary reference of head pose sequences, driving audio and emotion labels, AED-NeRF performs volume rendering based on the volume density and color predicted in the training stage, and generates an expressive 3D digital human that matches the driving source in real time.

3.2 Audio Processing Module

The vanilla NeRF[17] is only suitable for static scene modeling. To apply it to dynamic talking faces, we introduce audio features to compensate for the deformation channel of NeRF to model dynamic lip motions. Our audio processing module is illustrated in Fig. 2. The module first extracts the corresponding DeepSpeech[30] audio features from the input audio for each frame using a pre-trained recurrent neural network (RNN) model. Then, an audio attention network aggregates DeepSpeech audio features of neighboring frames in a self-attention manner to obtain smooth high-dimensional audio features a_h.

Previous NeRF-based methods[5, 18] directly concatenate high-dimensional audio features with spatial features and feed them into NeRF. However, this leads to high-dimensional inputs for the MLP, significantly increasing computational cost and resulting in slow training and rendering. To meet the real-time demand of digital human applications, we adopt the grid NeRF[34] to replace a portion of MLP forward propagation to query the spatial and audio features with linear interpolation. It compresses the size of MLP and accelerates the rendering speed effectively. Specifically, for any point x in the dynamic scene, it is firstly encoded as spatial grid features x_g by a 3D spatial grid encoder E^3_{spatial}. Then, high-dimensional audio features a_h are fused with the spatial grid features x_g and compressed to 2D audio features a_l by an MLP. This explicitly conditions audio features on the spatial position to ensure that the effect of audio sequences is constrained to the facial region only, rather than the torso or background. Finally, 2D audio features a_l are encoded as audio grid features a_g by a 2D audio grid encoder E^2_{audio} and then fed into NeRF.



RNN: recurrent neural network

Figure 2. Audio processing module

3.3 Emotion Modeling and Editing

Diverse emotional expressions effectively represent the emotional state of a digital human and contribute to a more realistic and vivid talking face avatars. We propose an explicit method for emotion control and editing based on emotion labels. Specifically, we set five basic emotion categories (i. e., neutral, angry, fearful, happy, and sad) and three levels of emotion intensity (i.e., weak, medium, and strong), and the combination of a specific emotion category and intensity is regarded as an emotion label. As illustrated in Fig. 3, given an emotion label as input, the emotion category and intensity are encoded into category features e_c and intensity features e_i by one-hot encoders, respectively. We concatenate category features e_c and intensity features e_i as the emotion feature $e = (e_c, e_i)$, which is fed into NeRF as the guidance of expression modeling. In the inference stage, facial expressions can be explicitly controlled and edited by combining different category features e_c and intensity features e_i. Experiments demonstrate that emotion labels and the simple but effective emotion encoder are enough for NeRF to model the dynamics of facial expressions through MSE loss and generate expressive talking face avatars.

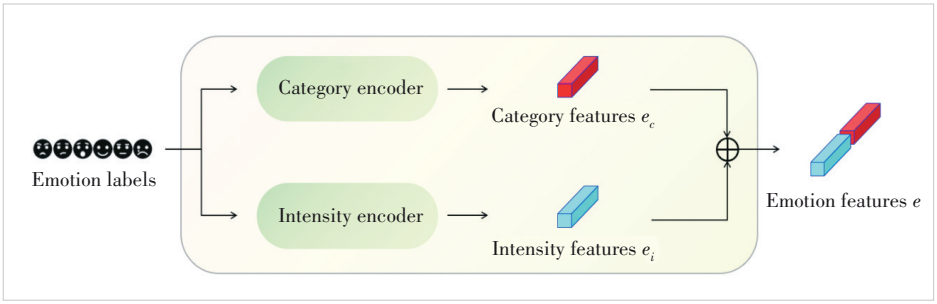


Figure 3. Illustration of emotion encoder

To achieve this, our head NeRF $\mathcal{F}_\Theta^{\text{head}}$ takes spatial grid features x_g , view direction d, audio grid features a_g and emotion features e as inputs to predict density σ and RGB color c of samples in camera rays. This can be formulated as:

$$\mathcal{F}_\Theta^{\text{head}}: (x_g, d, a_g, e) \rightarrow (\sigma, c) \tag{1}$$

3.4 Torso Modeling

Compared with the head part, torso movements are relatively slight and weakly correlated with our driving source. Thus, we follow SSP-NeRF[35] and design a deformation-based NeRF for torso modeling. Specifically, given a point x_t in the 2D image space, we condition it on the head pose p to predict the deformation of torso movements Δx via an MLP. This ensures that torso movements are synchronized with the head to avoid mismatched results caused by independent modeling of the head and torso. Then, the deformation Δx is added to the initial position x_t and fed to the 2D torso grid encoder E^2_{torso} to obtain torso grid features t_g. Finally, we feed grid features t_g into our torso NeRF to predict the density σ_t and RGB color c_t of the torso part. This can be formulated as:

$$\mathcal{F}_\Theta^{\text{torso}}: (t_g) \rightarrow (\sigma_t, c_t) \tag{2}$$

3.5 Implementation Details

1) Volume rendering. Given the density σ and RGB color c, we follow the rendering process of vanilla NeRF[17]. Specifically, we accumulate the density and RGB color of samples along the camera rays cast through each pixel to compute the output color under the specific view direction for talking face avatars. Given the camera center o and view direction d, the camera ray is represented as $r(t) = o + td$. Let near and far bounds be t_n and t_f , and the expected output color $\mathcal{C}$ is:

$$\mathcal{C}(r; \Theta, a_g, e) = \int_{t_n}^{t_f} T(t)\sigma(r(t))c(r(t), d)\,\mathrm{d}t \tag{3}$$

where $T(t)$ denotes the accumulated transmittance along the ray from t_n to t:

$$T(t) = \exp\left(-\int_{t_n}^{t}\sigma(r(s))\,\mathrm{d}s\right) \tag{4}$$

2) Loss function. We utilize the mean squared error (MSE) loss to minimize pixel-level reconstruction error:

$$L_{\text{MSE}} = \left\| \mathcal{C} - \mathcal{C}_{gt} \right\|_2^2 \tag{5}$$

where $\mathcal{C}$ is the rendered color and $\mathcal{C}_{gt}$ is the ground truth. As lip synchronization is crucial for talking face avatars, the pixel-level loss struggles to learn the complex semantic mapping from audio features to lip movements. Therefore, an additional learned perceptual image patch similarity (LPIPS)[36] loss is introduced for fine-tuning in the lip region:

$$L_{\text{LPIPS}} = \text{LPIPS}(\mathcal{P}, \mathcal{P}_{gt}) \tag{6}$$

where $\mathcal{P}$ is the rendered lip patch and $\mathcal{P}_{gt}$ is the ground truth.

Besides, for more accurate rendered results, we introduce the entropy regularization loss to encourage transmittance to be closer to 0 or 1:

$$L_\alpha = -\sum(\log\alpha + (1-\alpha)\log(1-\alpha)) \tag{7}$$

where α is the transparency of each rendered pixel.

We assume that the driving source only affects the facial region rather than the torso or background. Therefore, we adopt an L_1 regularization loss:

$$L_{\text{aud}} = \sum_{a_l \in \bar{R}_{\text{face}}} |a_l| \quad (8),$$

where $\bar{R}_{\text{face}}$ denotes the non-facial region. This encourages a_l to be 0 and avoids artifacts in the non-facial region.

Finally, the overall loss function can be formulated as:

$$L = L_{\text{MSE}} + \lambda_{\text{LPIPS}} L_{\text{LPIPS}} + \lambda_{\alpha} L_{\alpha} + \lambda_{\text{aud}} L_{\text{aud}} \quad (9).$$

3) Training details. We set the window size to 16 for the DeepSpeech RNN and 8 for the audio attention network. Our NeRF is composed of a 5-layer MLP with 64 hidden dimensions. For a specific identity, we first train the head NeRF for 20 000 epochs and fine-tune lip regions for 5 000 epochs. Then, torso NeRF is trained for 20 000 epochs. At each epoch, we randomly sample 256×256 camera rays and 16 samples for each ray and utilize the Adam optimizer with a learning rate of 0.000 5 to optimize the loss function. The loss coefficients are set to 0.01 for λ_{LPIPS}, 0.001 for λ_{α}, and 0.1 for λ_{aud}. For a 4-minute 25-fps video with resolution 512×512, the training time for the head NeRF and torso NeRF in a single RTX4090 is 5 h and 2 h, respectively.

4 Experiments

4.1 Experimental Settings

1) Datasets. For audio driving, we follow previous studies[5, 18] to collect several public speech videos of different celebrities to construct the celebrity dataset. The average video length is about 5 min, and both the recording camera and background are kept static. For emotion editing, we select MEAD[37] as the experimental data. MEAD is a large-scale audio-visual dataset, including abundant 3 – 5 s video clips of 60 actors speaking with 8 different emotions at 3 different intensity levels. We collect front-view video clips of 5 basic emotion categories (neutral, angry, fearful, happy, and sad) and 3 levels of emotion intensity (weak, medium, and strong) from MEAD for emotion editing experiments.

2) Data preprocessing. We first employ a face detection algorithm to locate the facial regions in all videos. Based on these facial regions, we crop the videos to a resolution of 512×512 with the faces in the center and resample them to 25 fps. Since a single video clip from MEAD is not enough for training, we combine the video clips conditioned on the same emotion category and intensity level from the same identity into a 90 – 120 s video as the data for the corresponding emotion label. Finally, we utilize a face parsing algorithm to annotate the head, torso, and background regions, and extract each part individually for each frame.

3) Metrics. We adopt peak signal-to-noise ratio (PSNR) and LPIPS[36] as image quality metrics. Note that PSNR only takes pixel-level differences into account and cannot faithfully reflect human perception of image quality. In comparison, LPIPS captures semantic information and structural similarity of images and is more consistent with subjective perception. For audio-visual synchronization evaluation, we adopt landmark distance (LMD)[38], SyncNet confidence (Sync-C), and SyncNet distance (Sync-D)[39] as metrics. LMD measures the distance between lip landmarks and the ground truth. Sync-C and Sync-D measure alignment and misalignment between audio and video streams via SyncNet, respectively.

4.2 Quantitative Comparisons

We first evaluate the audio-driving performance of our AED-NeRF under self-driven and cross-driven settings and compare it with non-NeRF-based methods, e. g., Wav2Lip[3] and live speech portraits (LSP)[40], and NeRF-based methods, e.g., audio driven NeRF (AD-NeRF)[5] and Dynamic Facial Radiance Fields (DFRF)[18] baselines. The self-driven results are shown in Table 1. PSNR, LPIPS, and LMD for LSP are not reported since LSP cannot generate the same poses as the ground truth. Our method performs best in most metrics with

Table 1. Quantitative comparison under the self-driven setting

Methods	Image Quality		Audio-Visual Synchronization			Rendering Speed	
	PSNR ↑	LPIPS ↓	LMD ↓	Sync-C ↑	Sync-D ↓	Training time/h ↓	Inference speed/fps ↑
GT	∞	0	0	8.897	6.325	/	/
Wav2Lip	**30.90**	0.139	3.311	**7.898**	**6.694**	/	15
LSP	/	/	/	5.181	8.637	/	25
AD-NeRF	28.79	0.101	3.245	3.944	10.603	36	0.09
DFRF	28.85	0.118	3.815	4.184	10.396	72	0.06
AED-NeRF	28.81	**0.088**	**2.826**	6.786	8.252	**7**	**45**

AD-NeRF: audio driven neural radiance fields
AED-NeRF: audio-driven and emotion-editing dynamic neural radiance fields
DFRF: dynamic facial radiance fields
LMD: landmark distance
LPIPS: learned perceptual image patch similarity
LSP: live speech portraits
PSNR: peak signal-to-noise ratio

real-time rendering speed. Specifically, we consider LPIPS as a more informative image quality metric, and AED-NeRF generates higher-quality talking face avatars compared with both existing non-NeRF-based and NeRF-based methods. In terms of audio-visual synchronization, AED-NeRF achieves optimal or sub-optimal scores in all metrics. Since Wav2Lip directly uses SyncNet as a loss term for supervision during training, its SyncNet scores are much better than other methods, even surpassing the ground truth. Our AED-NeRF achieves satisfactory SyncNet scores while significantly outperforming other methods in LMD, indicating that our method can generate synchronized lip movements with the audio source. In addition, AED-NeRF saves 80% – 90% training time and infers 500 – 750 times faster than NeRF-based baselines, enabling real-time applications. The cross-driven results in Table 2 demonstrate that the audio-visual synchronization performance of our AED-NeRF is second only to Wav2Lip but superior to other baselines, indicating that our method can still generate reasonable lip movements under the cross-driven setting.

Table 2. Quantitative comparison under the cross-driven setting

Methods	ID A		ID B	
	Sync-C ↑	Sync-D ↓	Sync-C ↑	Sync-D ↓
Wav2Lip	**8.748**	**7.623**	**8.208**	**7.193**
LSP	3.979	9.656	5.097	8.477
AD-NeRF	3.259	10.123	3.037	10.526
DFRF	4.607	9.235	4.245	10.083
AED-NeRF	6.624	8.799	6.074	8.075

AD-NeRF: audio driven neural radiance fields
AED-NeRF: audio-driven and emotion-editing dynamic neural radiance fields
DFRF: dynamic facial radiance fields
LSP: live speech portraits

AD-NeRF: audio driven neural radiance fields
AED-NeRF: audio-driven and emotion-editing dynamic neural radiance fields
DFRF: dynamic facial radiance fields

Figure 4. Qualitative comparison under the self-driven setting

AD-NeRF: audio driven neural radiance fields
AED-NeRF: audio-driven and emotion-editing dynamic neural radiance fields
DFRF: dynamic facial radiance fields
LSP: live speech portraits

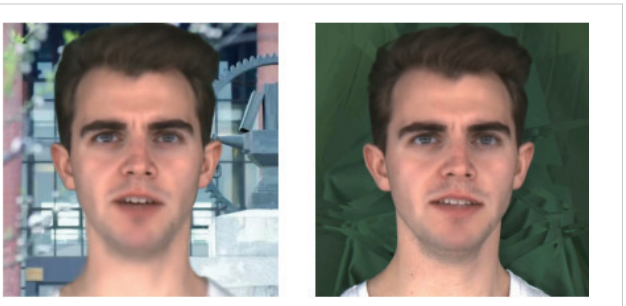
Figure 5. Qualitative comparison under the cross-driven setting

4.3 Qualitative Comparisons

Quantitative metrics have limitations in visual quality assessment and sometimes exhibit inconsistencies with subjective human perception. Therefore, we further conduct a qualitative evaluation of audio driving and emotion editing.

The self-driven results are illustrated in Fig. 4. In terms of image quality, Wav2Lip exhibits skin color distortion and obvious artifacts in the lip region; AD-NeRF loses some high-frequency information of images and suffers from head-torso separation when moving heavily; in contrast, our AED-NeRF faithfully reconstructs the talking face avatar of the reference identity. As for audio-visual synchronization, Wav2Lip, AD-NeRF, and DFRF deviate significantly from the ground truth, while our AED-NeRF synthesizes reasonable and accurate lip movements. The cross-driven results are illustrated in Fig. 5. By analyzing the lip synchronization, our AED-NeRF is capable of robustly synthesizing audio-visual synchronized lip movements even in challenging situations such as the pronunciation of the vowel /o/.

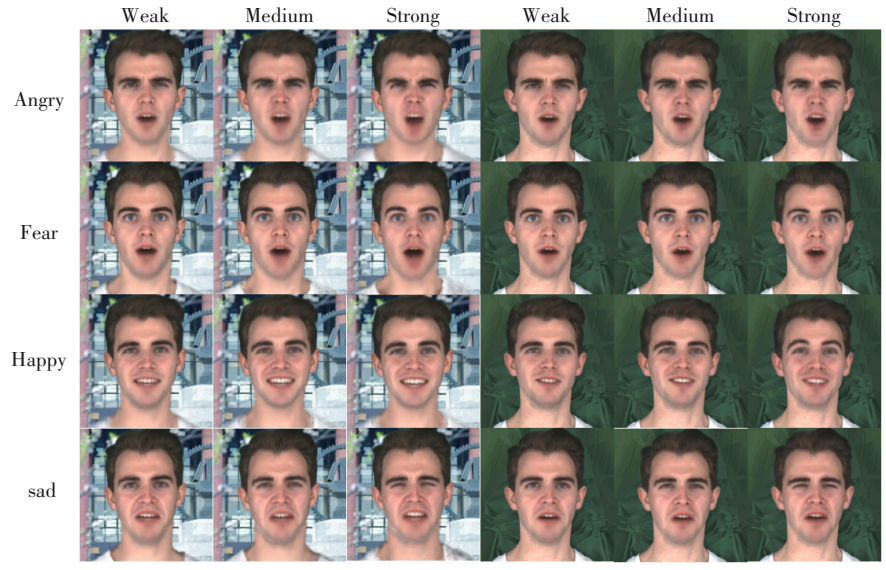
Since none of the baselines can generate corresponding expressions from emotion labels, we train AD-NeRF on each emotion-labeled video individually for comparison with our AED-NeRF. The neutral reference and generated emotion comparison are illustrated in Figs. 6 and 7, respectively. We replace the background with natural scenery in AD-NeRF and keep the green screen in our AED-NeRF for easier visual com-



(a) AD-NeRF (b) AED-NeRF

AD-NeRF: audio driven neural radiance fields
AED-NeRF: audio-driven and emotion-editing dynamic neural radiance fields

Figure 6. Neutral reference generated by (a) AD-NeRF and (b) AED-NeRF



(a) AD-NeRF (b) AED-NeRF

AD-NeRF: audio driven neural radiance fields
AED-NeRF: audio-driven and emotion-editing dynamic neural radiance fields

Figure 7. Qualitative comparison of emotion editing

Figure 8. Benefiting from NeRF architecture and background disentanglement, our AED-NeRF can synthesize talking face avatars in novel views and support background editing

parison. Note that our AED-NeRF can generate different emotional expressions by editing emotion labels within a single model. By comparing the facial expressions in the same frames, it can be found that our AED-NeRF warps facial regions (e. g., forehead, eyes, and mouth) corresponding to the emotion labels, and presents the desired expression. The warping becomes more pronounced as the intensity level increases, which reflects the differences among the three intensity levels.

Besides, AED-NeRF can synthesize talking face avatars in novel views and support background editing, which benefits from NeRF architecture and background disentanglement during data preprocessing.

4.4 Ablation Study

We conduct an ablation study to verify the effect of emotion modeling under the self-driven setting in Fig. 9. Without the guidance of emotion labels, the network has to model the visual emotion based only on audio conditions, which leads to a neutral or mismatched face even driven by extremely emotional audio. Our AED-NeRF generates expressive talking face avatars with matched facial expressions and more precise lip movements benefiting from emotion modeling.

5 Limitations

We have demonstrated that our AED-NeRF can generate realistic audio-driven expressive talking face avatars in real time. However, several limitations remain for future work. Lips may be jittering or unsynchronized with audio under the cross-driven setting, as an English automatic speech recognition model is employed to extract audio features which is not accurate for other languages. Therefore, how to extract general audio features across different languages is significant for further audio-visual synchronization improvement. Though our AED-NeRF supports emotion editing to generate expressive talking face avatars, the range of editing is limited to our provided emotion labels; that is, it cannot generalize beyond training data. A possible solution is to design an emotion recognition module to automatically classify emotion categories and intensity lev-

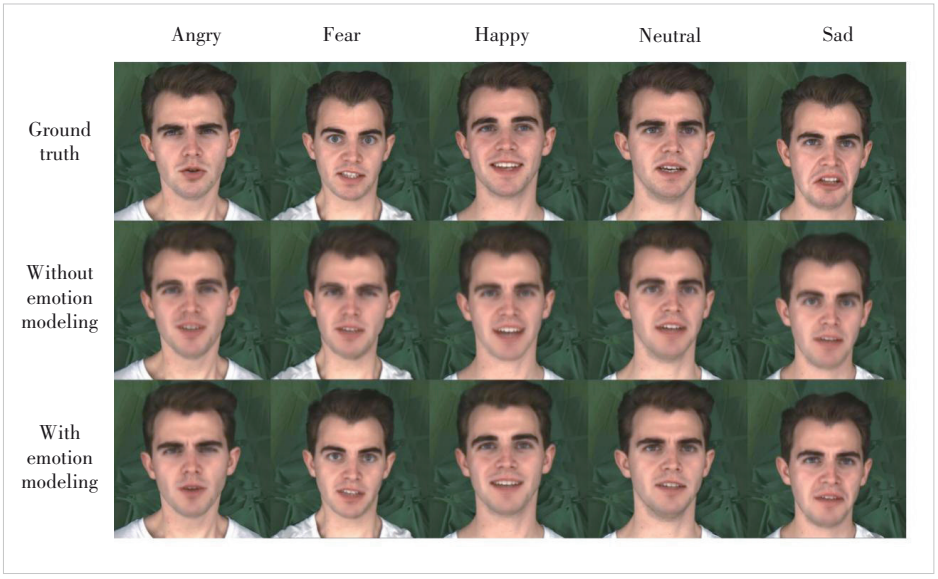


Figure 9. Ablation study on emotion modeling

els from videos, and learn a latent space for emotion embedding. Besides, slight variations, such as camera parameters, lighting and clothing, can affect modeling of one identity and lead to unideal generated results due to the nature of vanilla NeRF. For improved robustness to these variations, we will refer to works like NeRF in the wild[41] to disentangle environmental variations from facial dynamics in the future.

6 Conclusions

We propose AED-NeRF for the real-time generation of audio-driven, expressive talking face avatars. Audio and emotion features are introduced as the deformation channel for NeRF to implicitly model facial dynamics. Emotion labels composed of categories and intensity levels are encoded as guidance for emotion modeling of talking face avatars. Extensive experiments demonstrate that our AED-NeRF can generate photo-realistic expressive talking face avatars under different audio inputs and emotion settings in real time.

Ethical consideration: AED-NeRF can generate photorealistic expressive talking face avatars in real time under different audio and emotional conditions. However, talking face synthesis techniques could be misused. We restrict our AED-NeRF for research purposes only and support the development of deepfake detection methods.

References

[1] LeCun Y, Bengio Y, Hinton G. Deep learning [J]. Nature, 2015, 521 (7553): 436 – 444. DOI: 10.1038/nature14539
[2] Creswell A, White T, Dumoulin V, et al. Generative adversarial networks: an overview [J]. IEEE signal processing magazine, 2018, 35(1): 53 – 65. DOI: 10.1109/MSP.2017.2765202
[3] Prajwal K R, Mukhopadhyay R, Namboodiri V P, et al. A lip sync expert is all you need for speech to lip generation in the wild [C]//The 28th ACM International Conference on Multimedia. ACM, 2020: 484 – 492. DOI: 10.1145/3394171.3413532
[4] Thies J, Elgharib M, Tewari A, et al. Neural voice puppetry: Audio-driven facial reenactment [PP/OL]. arxiv (2020-07-29) [2024-09-06]. https://arxiv.org/abs/1912.05566
[5] Guo Y D, Chen K Y, Liang S, et al. AD-NeRF: audio driven neural radiance fields for talking head synthesis [C]//International Conference on Computer Vision. IEEE, 2021: 5764 – 5774. DOI: 10.1109/ICCV48922.2021.00573
[6] Kumar R, Sotelo J, Kumar K, et al. Obamanet: photo-realistic lip-sync from text [PP/OL]. arxiv (2017-12-06) [2024-09-06]. https://arxiv.org/abs/1801.01442
[7] Wang J D, Qian X Y, Zhang M L, et al. Seeing what you said: talking face generation guided by a lip reading expert [C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2023: 14653 – 14662. DOI: 10.1109/CVPR52729.2023.01408
[8] Zhong W Z, Fang C W, Cai Y Q, et al. Identity-preserving talking face generation with landmark and appearance priors [C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2023: 9729 – 9738. DOI: 10.1109/CVPR52729.2023.00938
[9] Isola P, Zhu J Y, Zhou T H, et al. Image-to-image translation with conditional adversarial networks [C]//Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017: 5967 – 5976. DOI: 10.1109/CVPR.2017.632
[10] Eskimez S E, Zhang Y, Duan Z Y. Speech driven talking face generation from a single image and an emotion condition [J]. IEEE transactions on multimedia, 2022, 24: 3480 – 3490. DOI: 10.1109/TMM.2021.3099900
[11] Zhou Y, Han X T, Shechtman E, et al. MakeItTalk: speaker-aware talking-head animation [J]. ACM transactions on graphics, 2020, 39(6): 1 – 15. DOI: 10.1145/3414685.3417774
[12] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets [C]//The 28th International Conference on Neural Information Processing Systems. ACM, 2014: 2672 – 2680. DOI: 10.5555/2969033.2969125
[13] Yin F, Zhang Y, Cun X D, et al. StyleHEAT: one-shot high-resolution editable talking face generation via Pretrained StyleGAN [C]//European Conference on Computer Vision (ECCV). ECVA, 2022: 85 – 101. DOI: 10.1007/978-3-031-19790-1_6
[14] Doukas M C, Zafeiriou S, Sharmanska V. Headgan: video-and-audio-driven talking head synthesis: Vol. 1 [PP/OL]. arxiv (2021-08-23) [2024-09-06]. https://arxiv.org/abs/2012.08261
[15] Chen L L, Maddox R K, Duan Z Y, et al. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss [C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2019: 7824 – 7833. DOI: 10.1109/CVPR.2019.00802
[16] Thies J, Elgharib M, Tewari A, et al. Neural voice puppetry: audio-driven facial reenactment [C]//European conference on computer vision (ECCV). ECVA, 2020: 716 – 731. DOI: 10.1007/978-3-030-58517-4_42
[17] Mildenhall B, Srinivasan P P, Tancik M, et al. NeRF: representing scenes as neural radiance fields for view synthesis [J]. Communications of the ACM, 2021, 65(1): 99 – 106. DOI: 10.1145/3503250
[18] Shen S, Li W H, Zhu Z, et al. Learning dynamic facial radiance fields for Few-shot talking head synthesis [C]//European Conference on Computer Vision (ECCV). ECVA, 2022: 666 – 682. DOI: 10.1007/978-3-031-19775-8_39
[19] Yao S Y, Zhong R Z, Yan Y C, et al. Dfa-NeRF: personalized talking

head generation via disentangled face attributes neural rendering [PP/OL]. arxiv (2022-01-03) [2024-09-16]. https://arxiv.org/abs/2201.00791
[20] Liu X, Xu Y H, Wu Q Y, et al. Semantic-aware implicit neural audio-driven video portrait generation [C]//European Conference on Computer Vision (ECCV). ECVA, 2022: 106 – 125. DOI: 10.1007/978-3-031-19836-6_7
[21] Ye Z H, He J Z, Jiang Z Y, et al. Geneface++: generalized and stable real-time audio-driven 3d talking face generation [PP/OL]. arxiv (2023-05-01) [2024-09-06]. https://arxiv.org/abs/2305.00787
[22] Yu Z T, Yin Z X, Zhou D Y, et al. Talking head generation with probabilistic audio-to-visual diffusion priors [C]//International Conference on Computer Vision (ICCV). IEEE, 2023: 7611 – 7621. DOI: 10.1109/ICCV51070.2023.00703
[23] Garg R, Roussos A, Agapito L. A variational approach to video registration with subspace constraints [J]. International journal of computer vision, 2013, 104(3): 286 – 314. DOI: 10.1007/s11263-012-0607-7
[24] Andrew A M. Multiple view geometry in computer vision [J]. Kybernetes, 2001, 30(9/10): 1333 – 1341. DOI: 10.1108/k.2001.30.9_10.1333.1
[25] Kajiya J T, Von Herzen B P. Ray tracing volume densities [J]. ACM SIGGRAPH computer graphics, 1984, 18(3): 165 – 174. DOI: 10.1145/964965.808594
[26] Zhou H, Liu Y, Liu Z W, et al. Talking face generation by adversarially disentangled audio-visual representation [J]. Proceedings of the AAAI conference on artificial intelligence, 2019, 33(1): 9299 – 9306. DOI: 10.1609/aaai.v33i01.33019299
[27] Das D, Biswas S, Sinha S, et al. Speech-driven facial animation using cascaded GANs for learning of motion and texture [C]//European Conference on Computer Vision. ECVA, 2020: 408 – 424. DOI: 10.1007/978-3-030-58577-8_25
[28] Richard A, Zollhöfer M, Wen Y D, et al. MeshTalk: 3D face animation from speech using cross-modality disentanglement [C]//IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2021: 1153 – 1162. DOI: 10.1109/ICCV48922.2021.00121
[29] Gafni G, Thies J, Zollhofer M, et al. Dynamic neural radiance fields for monocular 4D facial avatar reconstruction [C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2021: 8645 – 8654. DOI: 10.1109/cvpr46437.2021.00854
[30] Hannun A, Case C, Casper J, et al. Deep Speech: scaling up end-to-end speech recognition [PP/OL]. arxiv (2014-12-19) [2024-09-06]. https://arxiv.org/abs/1412.5567
[31] Hong Y, Peng B, Xiao H Y, et al. HeadNeRF: a realtime NeRF-based parametric head model [C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2022: 20342 – 20352. DOI: 10.1109/CVPR52688.2022.01973
[32] Ji X Y, Zhou H, Wang K, et al. EAMM: one-shot emotional talking face via audio-based emotion-aware motion model [C]//ACM SIGGRAPH 2022 Conference Proceedings. ACM, 2022: 1 – 10. DOI: 10.1145/3528233.3530745
[33] Tan S, Ji B, Pan Y. EMMN: emotional motion memory network for audio-driven emotional talking face generation [C]//International Conference on Computer Vision (ICCV). IEEE, 2023: 22089 – 22099. DOI: 10.1109/ICCV51070.2023.02024
[34] Müller T, Evans A, Schied C, et al. Instant neural graphics primitives with a multiresolution hash encoding [J]. ACM transactions on graphics, 2022, 41(4): 1 – 15. DOI: 10.1145/3528223.3530127
[35] Liu X, Xu Y H, Wu Q Y, et al. Semantic-aware implicit neural audio-driven video portrait generation [C]//European Conference on Computer Vision (ECCV). ECVA, 2022: 106 – 125. DOI: 10.1007/978-3-031-19836-6_7
[36] Zhang R, Isola P, Efros A A, et al. The unreasonable effectiveness of deep features as a perceptual metric [C]//Conference on Computer Vision and Pattern Recognition. IEEE, 2018: 586 – 595. DOI: 10.1109/CVPR.2018.00068
[37] Wang K, Wu Q Y, Song L S, et al. MEAD: a large-scale audio-visual dataset for emotional talking-face generation [C]//European conference on computer vision (ECCV). ECVA, 2020: 700 – 717. DOI: 10.1007/978-3-030-58589-1_42
[38] Chen L L, Li Z H, Maddox R K, et al. Lip movements generation at a glance [C]//European conference on computer vision. ECVA, 2018: 538 – 553. DOI: 10.1007/978-3-030-01234-2_32
[39] Chung J S, Zisserman A. Out of time: automated lip sync in the wild [EB/OL]. [2024-09-06]. https://www.robots.ox.ac.uk/~vgg/publications/2016/Chung16a/chung16a.pdf. DOI: 10.1007/978-3-319-54427-4_19
[40] Lu Y X, Chai J X, Cao X. Live speech portraits: real–time photorealistic talking-head animation [J]. ACM transactions on graphics, 2021, 40(6): 1 – 17. DOI: 10.1145/3478513.3480484
[41] Martin-Brualla R, Radwan N, Sajjadi M S M, et al. NeRF in the wild: neural radiance fields for unconstrained photo collections [C]//Computer Vision and Pattern Recognition. IEEE, 2021: 7206 – 7215. DOI: 10.1109/cvpr46437.2021.00713

Biographies

Lu Ping is the Vice President of ZTE Corporation, Director of the R&D Project of the Technology Planning Department, and Deputy Executive Director of the National Key Laboratory of Mobile Network and Mobile Multimedia Technology. His research fields include immersive communication, cloud computing, big data, augmented reality, and multimedia service technologies. He has supported and participated in major national science and technology projects as well as national science and technology support projects, and has published numerous academic papers in related fields.

Song Li (song_li@sjtu.edu.cn) received his BE and MS degrees in engineering in 1997 and 2000, respectively, and his PhD degree in electrical engineering from Shanghai Jiao Tong University (SJTU), China in 2005. He then joined SJTU as a faculty member and is currently a Full Professor at the Department of Electronic Engineering. He was also a Visiting Professor with Santa Clara University, USA from 2011 to 2012. He has more than 200 publications, obtained over 40 granted patents, and proposed 18 standard technical proposals in video coding and image processing. He has been serving as an Associate Editor for *Multidimensional Systems and Signal Processing* since 2012 and a Guest Editor for a special issue on "Quality of Experience for Advanced Broadcast Services" in 2018 in the *IEEE Transactions on Broadcasting*.

Shi Wenzhe is a strategy planning engineer with ZTE Corporation, a member of the National Key Laboratory of Mobile Network and Mobile Multimedia Technology, China, and a planning engineer of XRExplore platform products. His research objects include immersive communication, indoor visual AR navigation, SFM 3D reconstruction, visual SLAM, real-time cloud rendering, VR, and spatial perception.

Lin Zonghao received his BE degree in information engineering from Shanghai Jiao Tong University, China in 2023. He is currently pursuing his master's degree at the Department of Electronic Engineering, Shanghai Jiao Tong University, China. His research interests include image synthesis and talking face generation.

Ling Jun received his master's degree in electronic engineering and information science from University of Science and Technology of China in 2018. He is currently pursuing his PhD degree at the Department of Electronic Engineering, Shanghai Jiao Tong University, China. His research interests include image animation, talking face generation, and deep generative modeling.