

AE-NeRF: Audio Enhanced Neural Radiance Field for Few Shot Talking Head Synthesis

Dongze Li^{1,2}, Kang Zhao³, Wei Wang^{2*}, Bo Peng²,
Yingya Zhang³, Jing Dong², Tieniu Tan^{2,4}

¹School of Artificial Intelligence, University of Chinese Academy of Sciences

²CRIPAC & MAIS, Institute of Automation, Chinese Academy of Sciences

³Alibaba Group

⁴Nanjing University

Abstract

Audio-driven talking head synthesis is a promising topic with wide applications in digital human, film making and virtual reality. Recent NeRF-based approaches have shown superiority in quality and fidelity compared to previous studies. However, when it comes to few-shot talking head generation, a practical scenario where only few seconds of talking video is available for one identity, two limitations emerge: 1) they either have no base model, which serves as a facial prior for fast convergence, or ignore the importance of audio when building the prior; 2) most of them overlook the degree of correlation between different face regions and audio, e.g., mouth is audio related, while ear is audio independent. In this paper, we present Audio Enhanced Neural Radiance Field (AE-NeRF) to tackle the above issues, which can generate realistic portraits of a new speaker with few-shot dataset. Specifically, we introduce an Audio Aware Aggregation module into the feature fusion stage of the reference scheme, where the weight is determined by the similarity of audio between reference and target image. Then, an Audio-Aligned Face Generation strategy is proposed to model the audio related and audio independent regions respectively, with a dual-NeRF framework. Extensive experiments have shown AE-NeRF surpasses the state-of-the-art on image fidelity, audio-lip synchronization, and generalization ability, even in limited training set or training iterations.

Introduction

Audio-driven talking head generation is an essential technique with broad application scenarios such as digital human, film making, video conference and virtual reality. Many literature (Prajwal et al. 2020; Shen et al. 2023; Ye et al. 2023) have been put forward to learn the audio-to-lip mapping by using deep generative models, such as GAN, diffusion model, VAE, etc. Among them, Neural Radiance Field (NeRF) (Mildenhall et al. 2020) based methods (Guo et al. 2021; Liu et al. 2022; Shen et al. 2022) have shown promising results, which map audio features to dynamic neural radiance fields to model a talking head.

However, NeRF-based methods usually adopt identity-specific training, i.e., one needs to train a model from scratch

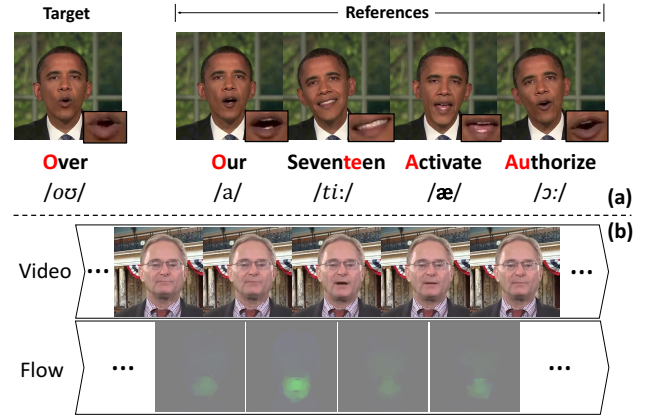


Figure 1: Our observation. (a) It shows the target/reference images and their phonetic symbols. The target is more similar to the first and last reference image because their pronunciation is closer. (b) We calculate the flow of adjacent frames, it can be seen the lower half face varies obviously than other regions.

for each new identity. What's worse, to make model generalize to various mouth shapes, the training set for each identity should be large, which is difficult to be satisfied in practice since the data for one identity are often limited. One-shot talking head generation (Chen et al. 2019; Prajwal et al. 2020; Zhou et al. 2020) may be a solution, which drives the novel identity from one reference image without training. But it sacrifices the fidelity of talking head, especially the teeth consistency and details. To balance data availability, training efficiency and generation quality, we focus on a practical scenario: few-shot talking head synthesis, that is, we need to train a NeRF model rapidly on a short talking video of one identity, which is capable of generating high-fidelity talking head with a given audio. Existing NeRF-based methods suffer from the following limitations when applied to this setting:

Lack of a robust prior. In order to quickly generalize to the few-shot identity, it is necessary to pre-train a base model across multi-identity to provide a basic audio-to-lip translation and implicit facial priors, such as color, shape and texture, which are helpful to restore faithful facial details.

*Corresponding author.

Current methods either lack the prior or construct a less robust prior. The former (Guo et al. 2021; Liu et al. 2022) is trained directly on the dataset of one identity. When the data is limited, the accuracy of rendering will drop significantly. The latter (Shen et al. 2022) ignores the correlation of audio between the target and reference images.

Audio-Face misalignment. In the audio-driven neural radiance field, most methods render the color of each ray conditioned on the audio feature, which means the entire talking head is considered to be *audio related*. Actually, we empirically find there exist many *audio decoupled* regions that have a weak or no correlation with the audio signal, such as hairs, ears and wrinkles. Modeling these two regions (i.e., audio related and decoupled) with one radiance field will cause a misalignment between the visual and audio information, resulting in sub-optimal synthesis results.

In this work, we propose Audio Enhanced NeRF (AE-NeRF) to tackle these two issues. Based on a reference scheme, we introduce *Audio Aware Aggregation* module and *Audio-Aligned Face Generation* strategy, to empower NeRF models with the ability to synthesize high quality talking heads with limited training data.

Specifically, we pre-train a base model on multi-identity dataset first. For each identity, we input several reference images to provide visual information from different poses to help the model render the target image. We find that *the close pronunciations have the similar mouth shapes* (see (a) of Fig. 1). Therefore, when aggregating the visual features from the reference images, whose weight should be higher if its audio is closer to that of the target image. For an audio-driven application, the *Audio Aware Aggregation* module will make the learned prior more robust and accurate.

On the other hand, we employ a dual-NeRF framework to simultaneously model the audio related and audio decoupled regions of a talking head. According to our observations in Fig. 1 (b), the lower half face can be regarded as the audio related part, whose variations have strong correlations with the audio signal. This part is modeled with an audio associated NeRF that conditions on audio features, while the rest parts are modeled by an audio independent NeRF that requires no audio features. Thanks to the disentanglement between different face regions and the audio signal, our *Audio-Aligned Face Generation* strategy brings better audio-to-lip consistency and finer rendering results.

To summarize, three key contributions are made to improve the practical few shot talking head synthesis.

- We propose an Audio Aware Aggregation module based on a reference scheme, which takes full advantage of the audio visual relationships between target and reference images and yields a strong prior.
- We introduce an Audio-Aligned Face Generation strategy to decouple the face modeling into audio associated NeRF and audio independent NeRF, achieving better audio-lip synchronization and facial details.
- Sufficient experiments have proved the superiority of our AE-NeRF over state-of-the-art on image fidelity, audio-lip synchronization, and generalization ability.

Related Work

Audio-driven Talking Head Generation

Audio-driven talking head generation aims to animate a speaker according to input audios. Image based methods (Prajwal et al. 2020; Zakharov et al. 2019) utilize GANs (Goodfellow et al. 2020) and Auto-encoders (Kingma and Welling 2013) to generate talking faces with audio signals as conditional inputs. Model based methods (Chen et al. 2019; Thies et al. 2020; Das et al. 2020; Wang, Mallya, and Liu 2021; Zhou et al. 2020; Song et al. 2022) leverage structural information such as 2D landmarks or 3DMM parameters for better face modeling. For instance, (Chen et al. 2019) and (Thies et al. 2020) generate faces with predicted facial landmarks or 3DMM expression coefficients. These methods can quickly adapt to an unseen identity. However, the prediction error of the representations may lead to inferior image quality, and they usually require hundreds of videos for training. NeRF based methods (Guo et al. 2021; Liu et al. 2022; Shen et al. 2022) have brought a new trend of talking head synthesis. They perform optimization on the video clip of a single person, and can synthesize pose-controllable faces of any resolution with high fidelity. AD-NeRF (Guo et al. 2021) use two separated NeRFs to model the head and the torso part respectively. SSP-NeRF (Liu et al. 2022) performs rays re-sampling based on the loss magnitude of different semantic regions. Despite the above advantages, the generalization ability of NeRF based methods to new identities still needs to be improved, and they suffer from performance drop when the video clip is relatively short.

Few Shot Neural Rendering

Neural Radiance Fields (NeRFs) (Mildenhall et al. 2020) combines MLPs with differentiable volume rendering and achieves photorealistic view synthesis results. Although impressive results are obtained, the original NeRF needs to be retrained for each new scene, which are both time consuming and computational expensive. Moreover, when only sparse views are available, because of the lack of the prior knowledge between scenes (Yu et al. 2021), the synthesis results can suffer from a large degradation in quality.

Few Shot Neural Rendering (Trevithick and Yang 2021; Yu et al. 2021; Chen et al. 2021; Gu et al. 2021; Xu et al. 2022) are proposed to alleviate these problems with the assistance of different kinds of priors such as 2D image features (Trevithick and Yang 2021; Yu et al. 2021; Wang et al. 2021), trainable latent codes (Jang and Agapito 2021; Gafni et al. 2021) and style inputs (Gu et al. 2021; Chan et al. 2022). Among them, the pixel level feature (Trevithick and Yang 2021; Yu et al. 2021; Wang et al. 2021) from randomly chose 2D reference images is the most commonly used prior to promote the NeRF’s rendering ability when only a few observations are available. When coming up with a new scene, the NeRF can perform quick generalization based on the reference image features from that scene. DFRF (Shen et al. 2022) directly uses the above reference scheme for few shot talking head generation. But it ignores the importance of audio features in talking head rendering.

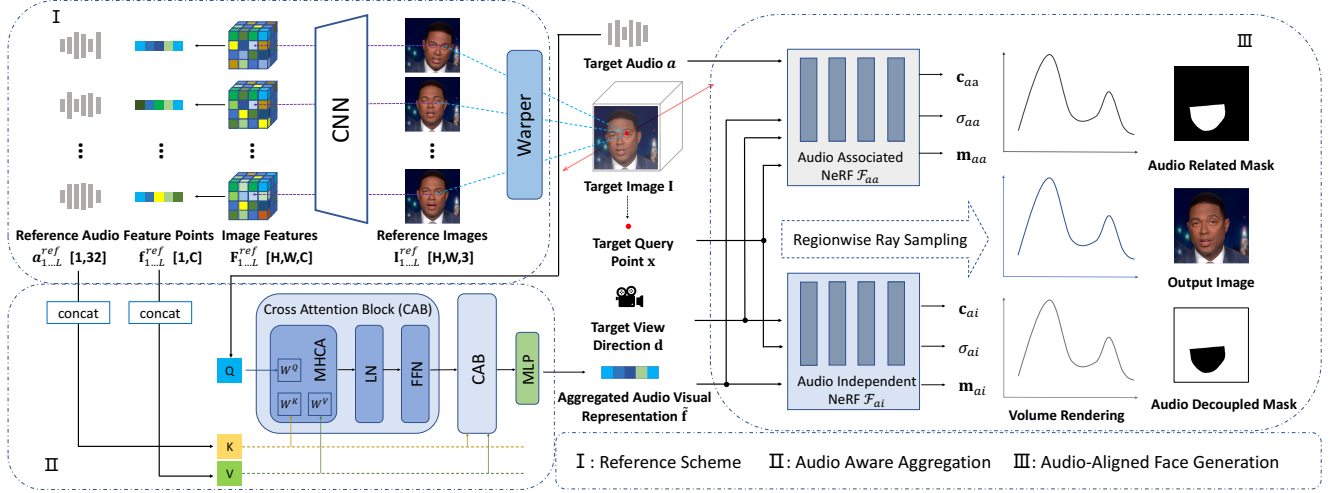


Figure 2: Overview of the proposed AE-NeRF. The reference scheme gathers the audio visual information from reference images and audio features precisely. Audio Aware Aggregation module fuses these features with cross attention and yields a strong representation. Audio-Aligned Face Generation strategy models the whole face region with two separated NeRFs, synthesizing the portrait with high fidelity.

Proposed Method

Overview

The full pipeline of our AE-NeRF is shown in Fig. 2. Both audio features and visual features are extracted as the references. The proposed Audio Aware Aggregation module fuses these features with cross attention and yields a strong prior for fully leveraging the limited data. Then, the audio related and audio decoupled regions are modeled by our Audio-Aligned Face Generation strategy. Finally, the portrait of the speaker and semantic masks are synthesized through volume rendering.

NeRF for Audio Driven Talking Head

The original NeRF encodes a static scene as a continuous volumetric radiance field \mathcal{F} , which is modeled by an MLP. It takes a 3D query point \mathbf{x} and its view direction \mathbf{d} as input, and outputs the corresponding density σ and color \mathbf{c} : $F(\mathbf{x}, \mathbf{d}) = (\mathbf{c}, \sigma)$. When applying NeRF to talking head, one will take the audio feature as an additional input, and the rendering process can be written as $\mathcal{F}(\mathbf{x}, \mathbf{d}, \mathbf{a}) = (\mathbf{c}, \sigma)$.

Reference Scheme. Despite the superior rendering quality, NeRF-based methods have to optimize each identity individually since no prior knowledge is shared between different identities. To improve the generalization ability of NeRFs on few observations, pixel level features from multi-view images (dubbed as reference images) (Yu et al. 2021; Trevithick and Yang 2021) are brought to construct a visual prior. DFRF (Shen et al. 2022) first utilizes this reference scheme in talking head generation, improving rendering ability on few shot datasets to some extent.

Specifically, given L reference images, let $\mathbf{I}_i \in \mathbb{R}^{H_i \times W_i \times 3}$ and $\mathbf{P}_i \in \mathbb{R}^{3 \times 4}$ denote the i -th image and camera projection matrix respectively ($i \in \{0, 1, \dots, L-1\}$).

A shallow convolutional network without downsampling is employed to extract dense features $\mathbf{F}_i \in \mathbb{R}^{H \times W \times D}$ from each image \mathbf{I}_i , where H, W and D are the height, width and the feature channel respectively. To facilitate the rendering of a 3D point \mathbf{x} on target image, we first project \mathbf{x} onto i -th reference image to obtain image features $\mathbf{f}_i^{ref} \in \mathbb{R}^D$. Then all extracted features $\mathbf{f}_{1 \dots L}^{ref}$ are merged as a condition $\tilde{\mathbf{f}}$. So, an audio-driven NeRF model with reference scheme can be formulated as

$$\mathcal{F}(\mathbf{x}, \mathbf{d}, \mathbf{a}, \tilde{\mathbf{f}}) = (\mathbf{c}, \sigma). \quad (1)$$

In addition, we denote the projection coordinate of the 3D point \mathbf{x} to the i -th 2D reference image as $\mathbf{p}_i^{ref} = (\mathbf{u}_i, \mathbf{v}_i)$. Since the talking head is dynamic, directly performing projection may bring some errors. Thus, an image warper is imposed to calibrate the 2D coordinate by predicting its offset $\Delta \mathbf{p}_i^{ref}$ on the feature plane. The calibrated coordinate is denoted by

$$\mathbf{p}_i^{ref'} = \mathbf{p}_i^{ref} + \Delta \mathbf{p}_i^{ref}. \quad (2)$$

Audio Aware Aggregation

In a talking head video, if the speakers in the reference image and the target image have similar speech contents, they tend to have similar mouth shapes, as we mentioned above. Therefore, we introduce Audio Aware Aggregation module into the feature fusion process to make the reference image, whose audio is more similar to target, contribute more. Let $\mathbf{a}, \mathbf{a}_{1 \dots L}^{ref}$ and $\mathbf{f}_{1 \dots L}^{ref}$ be target audio feature, reference audio feature and reference image feature respectively, then we have

$$\tilde{\mathbf{f}} = \text{AAA}(\mathbf{a}, \mathbf{a}_{1 \dots L}^{ref}, \mathbf{f}_{1 \dots L}^{ref}). \quad (3)$$

The Audio Aware Aggregation module utilizes a transformer structure with cross attention blocks to fuse the audio and the visual information. To be more concrete, $\mathbf{a}_{1...L}$, $\mathbf{a}_{1...L}^{ref}$, and $\mathbf{f}_{1...L}^{ref}$ are projected and reshaped to get their corresponding tokens $\mathbf{T}_a^{tar} \in \mathbb{R}^{N \times 128}$, $\mathbf{T}_a^{ref} \in \mathbb{R}^{N \times 128}$ and $\mathbf{T}_f^{ref} \in \mathbb{R}^{N \times 128}$, N is the number of query point samples in a batch. The audio and image tokens are then modeled by the Multi-Head Cross Attention (MHCA), along with Layer Normalization (LN), Residual Connection (RC) and Feed Forward Network (FFN). The cross attention block, with \mathbf{T}_a^{tar} as *query*, \mathbf{T}_a^{ref} as *key*, \mathbf{T}_f^{ref} as *value*, can be formalized as

$$\text{MHCA}(\mathbf{T}_a^{tar}, \mathbf{T}_a^{ref}, \mathbf{T}_f^{ref}) = \text{Softmax} \left[\frac{\mathbf{T}_a^{tar} \mathbf{W}^Q (\mathbf{T}_a^{ref} \mathbf{W}^K)^T}{\sqrt{d}} \right] \mathbf{T}_f^{ref} \mathbf{W}^V, \quad (4)$$

where $\mathbf{W}^Q \in \mathbb{R}^{128 \times d}$, $\mathbf{W}^K \in \mathbb{R}^{128 \times d}$, $\mathbf{W}^V \in \mathbb{R}^{128 \times d}$ are the projection matrices with hidden dimension d , which is also set to 128. It can be seen the closer the *key* (reference audio) and *query* (target audio) are, the greater the weight of *value* (reference feature) in \mathbf{f} will be. When fitting a new identity, given the reference images and audio features, it can help the NeRF to quickly model the texture and geometry. Two cross attention blocks are involved in the module and their output is passed through two Full Connection layers with a ReLU activation in between, and yields the final aggregated audio visual feature prior $\tilde{\mathbf{f}}$.

We also introduce the audio aware manner into the image warper module. The warper takes the target query point \mathbf{x} and the target audio \mathbf{a} as input, together with the corresponding audio \mathbf{a}_{ref} and ray direction \mathbf{d}_{ref} of the i -th reference, and outputs the coordinate offset, achieving more precise feature extraction:

$$\Delta \mathbf{p}_i^{ref} = (\Delta \mathbf{u}_i, \Delta \mathbf{v}_i) = \text{Warper}(\mathbf{x}, \mathbf{a}, \mathbf{a}_i^{ref}, \mathbf{d}, \mathbf{d}_i^{ref}). \quad (5)$$

Audio-Aligned Face Generation

As stated before, a disentangled modeling of the audio related region and the audio decoupled region is of great significance. Our Audio-Aligned Face Generation strategy uses an audio associated NeRF and an audio independent NeRF to model these two regions separately, and only the audio associated NeRF conditions on audio feature. To merge the rendering results of the two NeRF models, we add an additional parsing branch to predict mask \mathbf{m}_{aa} or \mathbf{m}_{ai} , where \mathbf{m}_{aa} has 1 in audio related region (i.e., the lower half face), and 0 in audio independent region, \mathbf{m}_{ai} is otherwise. Then we have the following two formulations:

$$\begin{aligned} \mathcal{F}_{aa}(\mathbf{x}, \mathbf{d}, \mathbf{a}, \tilde{\mathbf{f}}) &= (\mathbf{c}_{aa}, \sigma_{aa}, \mathbf{m}_{aa}) \\ \mathcal{F}_{ai}(\mathbf{x}, \mathbf{d}, \tilde{\mathbf{f}}) &= (\mathbf{c}_{ai}, \sigma_{ai}, \mathbf{m}_{ai}). \end{aligned} \quad (6)$$

For a query point \mathbf{x} , we can input it into two NeRFs and use the predicted mask to blend the colors or densities of both outputs, like (Ma et al. 2023). However, it will double the training time because a batch of rays has to go through the two NeRFs simultaneously.

Regionwise Ray Sampling. Consequently, we elaborate a Regionwise Ray Sampling mechanism to sample different

rays in different sub-regions, to mitigate the computational overhead caused by the dual NeRF. In this mechanism, each NeRF takes as input only the rays from its own corresponding regions and an overlapping region, improving the training speed without damaging the rendering quality.

Concretely, for a set of sampling rays Ω , we use $\epsilon \times \Omega$ to represent a new set that has $\epsilon \times |\Omega|$ rays randomly sampled from Ω ($|\Omega|$ means the number of rays in Ω). The set of the rays from the audio related region and the audio decoupled region are denoted as Ω_{ar} and Ω_{ad} . An overlap region is defined as $\Omega_{overlap} = (\epsilon \times \Omega_{ar}) \cup (\epsilon \times \Omega_{ad})$, where rays are fed into two NeRFs simultaneously. While the remaining parts $\Omega_{aa} = \Omega_{ar} \setminus \Omega_{overlap}$ and $\Omega_{ai} = \Omega_{ad} \setminus \Omega_{overlap}$ are rendered by their corresponding NeRFs separately. (\cup and \setminus denote sets union and subtraction operations). In practice, ϵ is set to 0.4 for the best lip generation result. More effects of this Regionwise Ray Sampling mechanism can be found in the supplementary material.

Volume Rendering. During training, for rays from Ω_{aa} or Ω_{ai} , the color density and the occupancy are obtained directly from their corresponding NeRFs. For a pixel lying in the overlapping region $\Omega_{overlap}$, its color density and the occupancy become the mixup of the two NeRFs:

$$\begin{aligned} \mathbf{c} &= \begin{cases} \mathbf{m}_{aa} \cdot \mathbf{c}_{aa} + \mathbf{m}_{ai} \cdot \mathbf{c}_{ai}, & \mathbf{r} \in \Omega_{overlap} \\ \mathbf{c}_{aa}, & \mathbf{r} \in \Omega_{aa} \\ \mathbf{c}_{ai}, & \mathbf{r} \in \Omega_{ai} \end{cases} \\ \sigma &= \begin{cases} \sigma_{aa} + \sigma_{ai}, & \mathbf{r} \in \Omega_{overlap} \\ \sigma_{aa}, & \mathbf{r} \in \Omega_{aa} \\ \sigma_{ai}, & \mathbf{r} \in \Omega_{ai} \end{cases} \end{aligned} \quad (7)$$

During inference, the whole image are regarded as the overlap region, and rendered by two NeRFs at the same time, since there is no ground truth mask available. To get the predicted RGB pixel and the mask, we utilize classical volume rendering to accumulate the samples on the ray:

$$\begin{aligned} \hat{\mathbf{C}}(\mathbf{r}) &= \int_{t_n}^{t_f} T(t) \sigma(\mathbf{r}(t), \mathbf{a}, \tilde{\mathbf{f}}) \mathbf{c}(\mathbf{r}(t), \mathbf{d}, \mathbf{a}, \tilde{\mathbf{f}}) dt, \\ \hat{\mathbf{m}}(\mathbf{r}) &= \int_{t_n}^{t_f} T(t) \sigma(\mathbf{r}(t), \mathbf{a}, \tilde{\mathbf{f}}) \mathbf{m}(\mathbf{r}(t), \mathbf{d}, \mathbf{a}, \tilde{\mathbf{f}}) dt, \end{aligned} \quad (8)$$

where $T(t) = \exp\left(-\int_{t_n}^t \sigma(s) ds\right)$ denotes for the accumulated transmittance along the ray from t_n to t , t_n and t_f are the lower and the upper bound of depth respectively. Eq. (8) shows the rendering process of the audio associated NeRF, where both the audio signal and the aggregated audio visual feature take part in the volume rendering process. For audio independent NeRF, audio signal should be removed.

Network Training

Following the original NeRF (Mildenhall et al. 2020), we use a reconstruction loss term to optimize the coarse and the fine network (we still take audio associated NeRF as example if not specified), which can be written as

$$\begin{aligned} \mathcal{L}_p &= \sum_{\mathbf{r} \in \mathcal{R}} \left[\left\| \hat{\mathbf{C}}_c(\mathbf{r}) - \mathbf{C}(\mathbf{r}) \right\|_2^2 + \left\| \hat{\mathbf{C}}_f(\mathbf{r}) - \mathbf{C}(\mathbf{r}) \right\|_2^2 \right], \\ \mathcal{L}_m &= \sum_{\mathbf{r} \in \mathcal{R}} \left[\left\| \hat{\mathbf{m}}_c(\mathbf{r}) - \mathbf{m}(\mathbf{r}) \right\|_2^2 + \left\| \hat{\mathbf{m}}_f(\mathbf{r}) - \mathbf{m}(\mathbf{r}) \right\|_2^2 \right], \end{aligned} \quad (9)$$

where $\hat{\mathbf{C}}_c(\mathbf{r})$ and $\hat{\mathbf{C}}_f(\mathbf{r})$ are the predicted pixels from the coarse and the fine model respectively, \mathcal{R} denotes for a batch of rays, and $\mathbf{C}(\mathbf{r})$ is the ground truth pixel color corresponding to each sampled ray. Similarly, $\hat{\mathbf{m}}(\mathbf{r})$ and $\mathbf{m}(\mathbf{r})$ are the predicted mask and ground truth. For the audio associated NeRF, the ground truth mask is defined as

$$\mathbf{m}(\mathbf{r}) = \begin{cases} 1, & \mathbf{r} \in \Omega_{ar} \\ 0, & \mathbf{r} \in \Omega_{ad} \end{cases} \quad (10)$$

while the ground truth mask for the audio independent NeRF is the opposite. Besides, we use an l_2 loss term to regularize the magnitude of the predicted offset of the warper, which can be written as

$$\mathcal{L}_o = \frac{1}{L \cdot |\mathcal{P}|} \sum_{i=1}^L \sum_{\mathbf{x} \in \mathcal{P}} \sqrt{\Delta \mathbf{u}_i^2 + \Delta \mathbf{v}_i^2}, \quad (11)$$

where \mathcal{P} is the collection of all the 3D query points.

Our final loss term can be given as

$$\mathcal{L} = \mathcal{L}_p + \lambda_m \mathcal{L}_m + \lambda_o \mathcal{L}_o, \quad (12)$$

where λ_m and λ_o are weight parameters.

Experiments

Experimental Setup

Dataset Preparation. We use the videos provided by AD-NeRF (Guo et al. 2021), DFRF (Shen et al. 2022), and HDTF dataset (Zhang et al. 2021) to conduct our experiments. Videos are all resampled to 25 fps and resized to a resolution of 512×512 . For each video, the first half of it is used for training and the second half is used for inference.

Baseline Methods. We compare our method with one image based method Wav2Lip (Prajwal et al. 2020), two model based methods ATVG (Chen et al. 2019) and MakeItTalk (Zhou et al. 2020), and two NeRF based methods AD-NeRF (Guo et al. 2021) and DFRF (Shen et al. 2022). For the first three methods, we use their official code and provided pre-trained models. For AD-NeRF and DFRF, we retrain them on each video on the same number of iterations as our method for fair comparison. DFRF also has a base model like our method, and we have tried to pre-train a base model for AD-NeRF like DFRF and our method, but it fails to generate plausible results for the lack of generalization ability. Comparison with another SOTA NeRF-based method SSP-NeRF (Liu et al. 2022) whose training code is not provided, more implementation details, the limitation of our method and the ethical consideration can be found in the supplementary material.

Evaluation Metrics. We employ evaluation metrics that have been previously used in talking head generation. We use PSNR and SSIM to evaluate the image level quality of generated results, LPIPS (Zhang et al. 2018) to evaluate the feature level quality. We also use Landmark Distance (LMD) (Chen et al. 2018) and SyncNet Confidence (Chung and Zisserman 2017) to further measure the mouth shapes and the audio visual synchronization.

	Testset A					Testset B	Testset C
Methods	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	LMD \downarrow	Sync \uparrow	Sync \uparrow	Sync \uparrow
gt	∞	1	0	0	8.545	8.406	8.873
ATVG	19.12	0.646	0.523	2.591	5.657	4.726	6.315
Wav2Lip	29.64	0.843	0.423	2.612	9.750	7.824	10.715
MakeItTalk	22.28	0.655	0.480	10.720	5.945	4.378	5.556
AD-NeRF	27.73	<u>0.881</u>	0.202	<u>2.603</u>	4.274	4.230	4.656
DFRF	<u>32.30</u>	0.949	<u>0.080</u>	3.023	5.219	4.859	5.321
AE-NeRF	32.63	0.949	0.078	2.425	<u>6.904</u>	<u>6.217</u>	<u>6.690</u>

Table 1: Method comparison under self-driven (Testset A) and cross-driven (Testset B and C) setting. The best and the second results are in bold and underlined respectively.

Face Quality Comparison

To compare the quality of the generated talking head thoroughly, two different settings are taken into account: Self driven setting, where the video and the audio are from the same person. Cross driven setting, where the audio from one person is used to drive another identity. Each video is about two minutes in length.

Results under Self-driven Setting. Key frames generated according to the Obama video in (Guo et al. 2021) and the broadcaster videos in (Shen et al. 2022) are shown in Fig.3. NeRF based methods have shown superiority image quality against image based methods and model based methods, and have managed to generate high fidelity synthesis results. But AD-NeRF suffers from head-torso misalignment as the red arrows pointed out, while DFRF fails to generate some face details correctly. Besides, both AD-NeRF and DFRF tend to generate lips misaligned from the ground truth. Our AE-NeRF has shown the best lip-alignment with the ground truth frames, as well as the facial details.

Quantitative comparison results on AD-NeRF and DFRF videos are shown in Testset A part of Tab.1. Wav2Lip (Prajwal et al. 2020) uses a pretrained SyncNet (Chung and Zisserman 2017) as the optimization objective and it achieves a SyncNet score even better than the ground truth. However, the quality of its generated images is relatively low. NeRF-based methods have shown their superiority not only at the pixel level (PSNR and SSIM) but also at the feature level (LPIPS). AD-NeRF performs slightly worse than DFRF and our method in image quality metrics due to the head-torso misalignment. It can be seen that our AE-NeRF surpasses baseline methods by a large margin in SyncNet confidence and LMD, which indicates the effect of learning aggregated audio visual features for lip synthesis.

Results under Cross-driven Setting. Cross driven results are shown in Testset B and Testset C of Tab.1. Audios from HDTF dataset are used to drive other identities. We only calculate the SyncNet score since there is no ground truth for other metrics. AD-NeRF and DFRF fail to synthesize accurate lip shapes according to audios from different speakers. We attribute this to the lack of aggregated audio visual information. Our AE-NeRF have shown competitive performance in audio-lip consistency.

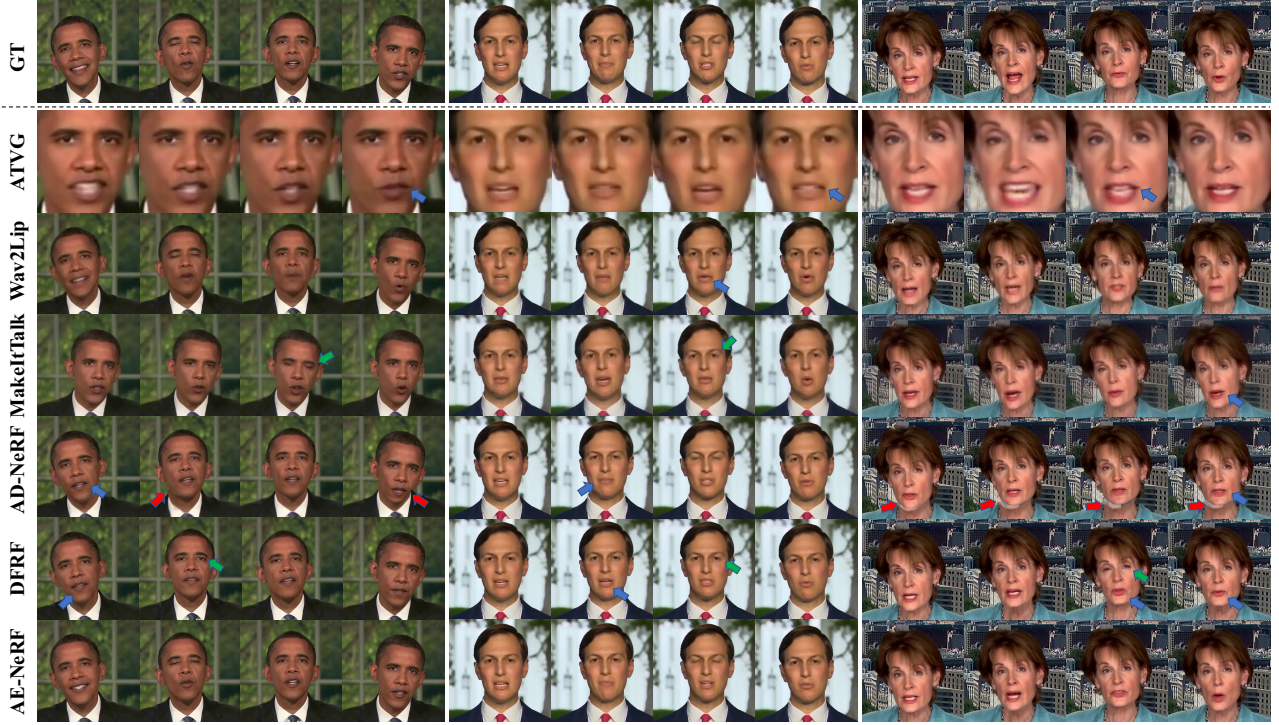


Figure 3: Qualitative comparison with other baseline methods for generated key frame results. We use the blue arrow to denote the inaccurate lip synthesis results like incorrect shape or blurred lips, the red arrow to denote the head-torso inconsistency, and the green arrow to represent inaccurate expression synthesis results.

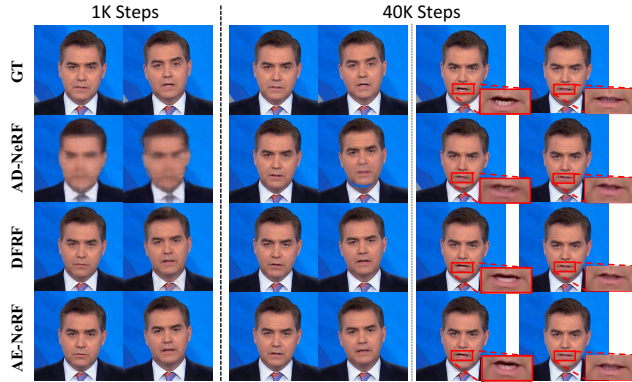


Figure 4: Qualitative comparison against other methods with 1k and 40k finetuning steps. 15s video clip is available for training.

Few Shot Talking Head Synthesis

We compare our AE-NeRF with other NeRF based talking head methods under a more challenging setting, few shot talking head synthesis, which further validate the generation ability of our method.

Synthesizing Talking Head with Short Videos. Firstly, we show the performance of different NeRF based methods on very short videos, with each model from a different method

Method	Length	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	LMD \downarrow	Sync \uparrow
AD-NeRF	10s	22.83	0.846	0.156	2.427	1.587
	15s	23.37	0.867	0.138	2.116	3.528
	20s	23.01	0.856	0.142	1.684	4.102
DFRF	10s	28.87	0.926	0.076	1.971	3.512
	15s	29.60	0.938	0.066	1.804	3.688
	20s	31.32	0.942	0.069	1.84	4.459
AE-NeRF	10s	28.93	0.930	0.072	1.944	6.102
	15s	29.52	0.938	0.067	1.766	6.217
	20s	31.49	0.946	0.064	1.528	6.743

Table 2: Method comparison with different training data length under 40k iterations.

being fine-tuned by 40k steps. Metrics calculated with training data lengths of 10s, 15s, and 20s are shown in Tab. 2. It can be seen that when training under a few shot setting, our method still maintains high image generation quality and audio visual alignment. Comparison of different faical details are shown in Fig. 4. Our method keeps the best of the original facial details, especially at the mouth region, while AD-NeRF and DFRF sometimes generate incorrect lip shape. We attribute this phenomenon to the lack of feature prior and the naive visual feature fusion process without considering the audio information. This further validates the face modeling ability of our AE-NeRF.






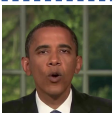
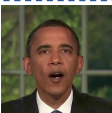
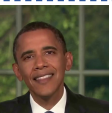
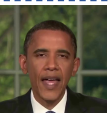

Target	References				
					
Mess	Challenge	Happening	Middle	World	
DFRF	0.2446	0.2857	0.2177	0.2175	
AE-NeRF	0.7583	0.9792	0.0425	0.0652	
					
Over	Our	Seventeen	Activate	Authorize	
DFRF	0.2612	0.2461	0.2463	0.2463	
AE-NeRF	0.9088	0.0071	0.0119	0.9547	

Figure 5: Attention score between the target feature and different reference features in the cross attention block. Similar audio visual contents bring higher attention scores.

Methods		PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	LMD \downarrow	Sync \uparrow
ground truth		∞	1	0	0	8.065
AD-NeRF	1k	22.13	0.777	0.305	9.158	0.291
	10k	23.72	0.868	0.143	2.148	3.697
	40k	23.37	0.867	0.138	2.116	3.528
DFRF	1k	29.39	0.931	0.086	1.918	2.995
	10k	29.47	0.936	0.073	1.905	2.985
	40k	29.60	0.936	0.070	1.804	3.688
AE-NeRF	1k	29.18	0.930	0.087	1.893	5.552
	10k	29.32	0.937	0.072	1.765	6.044
	40k	29.52	0.938	0.067	1.766	6.217

Table 3: Method comparison with 15s training clip under different training iterations.

Synthesizing Talking Head with Few Iterations. We further explore the generation ability of different methods with few training iterations. We utilize video clips released from DFRF (Shen et al. 2022) to carry out our experiment. Each video is 30s in length. We compare the portraits generated after 1k and 40k training steps to further show the synthesis results of different methods, which is also shown in Fig. 4. Within 1k steps, DFRF and our method can fit the new identity, which AD-NeRF fails to generalize on. In 40k steps, our method can generate portraits with higher fidelity than DFRF. Quantitatively, results with different training steps are shown in Tab. 3, where our AE-NeRF achieves similar image quality metrics with DFRF under the same training step, but it has shown superiority in SyncNet confidence and LMD, indicating better lip synthesis results.

Ablation Study

An ablation study is conducted to show the function of each proposed module. We replace our Audio Aware Aggregation with a slot attention module (Locatello et al. 2020) which simply fuses the visual feature without considering audio signals, denoted as w/o AAA. We also test the per-

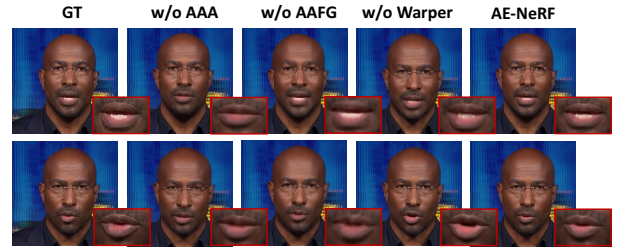


Figure 6: Qualitative ablation study

Methods	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	LMD \downarrow	Sync \uparrow
GT	∞	1	0	0	8.545
w/o AAA	31.25	0.929	0.092	3.646	4.570
w/o AAFG	32.30	0.940	0.080	3.938	6.282
w/o Warper	32.49	0.949	0.076	2.779	6.952
AE-NeRF	32.64	0.950	0.076	2.613	7.813

Table 4: Quantitative ablation study on different modules.

formance of the model without Audio-Aligned Face Generation, where an audio associated NeRF is used to model the whole face area, denoted as w/o AAFG. The model trained without the warper is denoted as w/o Warper. The qualitative and quantitative results are shown in Fig. 6 and Tab. 4. The Audio Aware Aggregation module brings better quality in both pixel and feature level. Model obtained without this module has less awareness of the audio signal, resulting in inferior image quality. Our Audio-Aligned Face Generation strategy brings more correct lip shapes and more natural expressions. The warper can bring more accurate visual feature points, which can also affect the audio lip synchronization.

To further study the effect of the Audio Aware Aggregation module, we calculate the average attention scores between different references and the target feature in the first attention block of our Audio Aware Aggregation module, shown in Fig. 5. As a comparison, scores in the feature aggregation module used in DFRF are also taken into account. The inner product between target-reference pairs with similar audio visual contents is significantly higher than those with distinct contents in our cross attention block. This proves that the audio visual interaction process assists the rendering of the target pixel, resulting in better generation results.

Conclusion

This work presents Audio Enhanced Neural Radiance Field (AE-NeRF) for few shot talking head synthesis. Our method consists of an Audio Aware Aggregation module which learns a strong prior for improving the generalization ability and an Audio-Aligned Face Generation strategy to better model the audio related and the audio decoupled face regions. Comparisons between SOTA methods confirm that our AE-NeRF achieves better image quality and fidelity under the custom scenario and a few shot setting.

Ethical Statement

Our AE-NeRF is capable of generating vivid speech portraits with high fidelity, and can be applied to various situations such as virtual human, digital games and film making. On the other hand, the misuse of the talking head synthesis technique can lead to moral and legal issues, such as crafting malicious DeepFake videos. We are committed to fighting against this kind of egregious behavior and use our code and models in the development of the DeepFake detection models.

Acknowledgments

This work is supported by the National Key Research and Development Program of China under Grant No. 2021YFC3320103, the National Natural Science Foundation of China (NSFC) under Grants 62372452, U19B2038, and by Alibaba Group through Alibaba Innovative Research Program.

References

- Chan, E. R.; Lin, C. Z.; Chan, M. A.; Nagano, K.; Pan, B.; De Mello, S.; Gallo, O.; Guibas, L. J.; Tremblay, J.; Khamis, S.; et al. 2022. Efficient geometry-aware 3D generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16123–16133.
- Chen, A.; Xu, Z.; Zhao, F.; Zhang, X.; Xiang, F.; Yu, J.; and Su, H. 2021. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14124–14133.
- Chen, L.; Li, Z.; Maddox, R. K.; Duan, Z.; and Xu, C. 2018. Lip movements generation at a glance. In *Proceedings of the European conference on computer vision (ECCV)*, 520–535.
- Chen, L.; Maddox, R. K.; Duan, Z.; and Xu, C. 2019. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7832–7841.
- Chung, J. S.; and Zisserman, A. 2017. Out of time: automated lip sync in the wild. In *Computer Vision—ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20–24, 2016, Revised Selected Papers, Part II* 13, 251–263. Springer.
- Das, D.; Biswas, S.; Sinha, S.; and Bhowmick, B. 2020. Speech-driven facial animation using cascaded gans for learning of motion and texture. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX* 16, 408–424. Springer.
- Gafni, G.; Thies, J.; Zollhofer, M.; and Nießner, M. 2021. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8649–8658.
- Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.
- Gu, J.; Liu, L.; Wang, P.; and Theobalt, C. 2021. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. *arXiv preprint arXiv:2110.08985*.
- Guo, Y.; Chen, K.; Liang, S.; Liu, Y.-J.; Bao, H.; and Zhang, J. 2021. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5784–5794.
- Jang, W.; and Agapito, L. 2021. Codenerf: Disentangled neural radiance fields for object categories. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12949–12958.
- Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.
- Liu, X.; Xu, Y.; Wu, Q.; Zhou, H.; Wu, W.; and Zhou, B. 2022. Semantic-aware implicit neural audio-driven video portrait generation. In *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*, 106–125. Springer.
- Locatello, F.; Weissenborn, D.; Unterthiner, T.; Mahendran, A.; Heigold, G.; Uszkoreit, J.; Dosovitskiy, A.; and Kipf, T. 2020. Object-centric learning with slot attention. *Advances in Neural Information Processing Systems*, 33: 11525–11538.
- Ma, T.; Li, B.; He, Q.; Dong, J.; and Tan, T. 2023. Semantic 3D-aware Portrait Synthesis and Manipulation Based on Compositional Neural Radiance Field. *arXiv preprint arXiv:2302.01579*.
- Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *ECCV*.
- Prajwal, K.; Mukhopadhyay, R.; Namboodiri, V. P.; and Jawahar, C. 2020. 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*, 484–492.
- Shen, S.; Li, W.; Zhu, Z.; Duan, Y.; Zhou, J.; and Lu, J. 2022. Learning dynamic facial radiance fields for few-shot talking head synthesis. In *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XII*, 666–682. Springer.
- Shen, S.; Zhao, W.; Meng, Z.; Li, W.; Zhu, Z.; Zhou, J.; and Lu, J. 2023. DiffTalk: Crafting Diffusion Models for Generalized Audio-Driven Portraits Animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1982–1991.
- Song, L.; Wu, W.; Qian, C.; He, R.; and Loy, C. C. 2022. Everybody’s talkin’: Let me talk as you want. *IEEE Transactions on Information Forensics and Security*, 17: 585–598.
- Thies, J.; Elgharib, M.; Tewari, A.; Theobalt, C.; and Nießner, M. 2020. 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, 716–731. Springer.
- Trevithick, A.; and Yang, B. 2021. Grf: Learning a general radiance field for 3d representation and rendering. In

Proceedings of the IEEE/CVF International Conference on Computer Vision, 15182–15192.

Wang, Q.; Wang, Z.; Genova, K.; Srinivasan, P. P.; Zhou, H.; Barron, J. T.; Martin-Brualla, R.; Snavely, N.; and Funkhouser, T. 2021. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4690–4699.

Wang, T.-C.; Mallya, A.; and Liu, M.-Y. 2021. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10039–10049.

Xu, Q.; Xu, Z.; Philip, J.; Bi, S.; Shu, Z.; Sunkavalli, K.; and Neumann, U. 2022. Point-nerf: Point-based neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5438–5448.

Ye, Z.; Jiang, Z.; Ren, Y.; Liu, J.; He, J.; and Zhao, Z. 2023. GeneFace: Generalized and High-Fidelity Audio-Driven 3D Talking Face Synthesis. *arXiv preprint arXiv:2301.13430*.

Yu, A.; Ye, V.; Tancik, M.; and Kanazawa, A. 2021. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4578–4587.

Zakharov, E.; Shysheya, A.; Burkov, E.; and Lempitsky, V. 2019. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9459–9468.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhang, Z.; Li, L.; Ding, Y.; and Fan, C. 2021. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3661–3670.

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