AD-NeRF: Audio Driven Neural Radiance Fields for Talking Head Synthesis

Yudong Guo[*1,2] Keyu Chen[1,2] Sen Liang[3] Yong-Jin Liu[4] Hujun Bao[3] Juyong Zhang[†1]

[1]University of Science and Technology of China [2]Beijing Dilusense Technology Corporation
[3]Zhejiang University [4]Tsinghua University

Abstract

Generating high-fidelity talking head video by fitting with the input audio sequence is a challenging problem that receives considerable attentions recently. In this paper, we address this problem with the aid of neural scene representation networks. Our method is completely different from existing methods that rely on intermediate representations like 2D landmarks or 3D face models to bridge the gap between audio input and video output. Specifically, the feature of input audio signal is directly fed into a conditional implicit function to generate a dynamic neural radiance field, from which a high-fidelity talking-head video corresponding to the audio signal is synthesized using volume rendering. Another advantage of our framework is that not only the head (with hair) region is synthesized as previous methods did, but also the upper body is generated via two individual neural radiance fields. Experimental results demonstrate that our novel framework can (1) produce high-fidelity and natural results, and (2) support free adjustment of audio signals, viewing directions, and background images. Code is available at https://github.com/YudongGuo/AD-NeRF.

1. Introduction

Synthesizing high-fidelity audio-driven facial video sequences is an important and challenging problem in many applications like digital humans, chatting robots, and virtual video conferences. Regarding the talking-head generation process as a cross-modal mapping from audio to visual faces, the synthesized facial images are expected to perform natural speaking styles while synchronizing photo-realistic streaming results as same as the original videos.

Currently, a wide range of approaches have been proposed for this task. Earlier methods built upon professional artist modelling [12, 60] or complicated motion capture system [6, 54] are limited in high-end areas of the movie and game industry. Recently, many deep-learning-based techniques [35, 42, 10, 58, 7, 43, 48, 59, 21, 57] are proposed to learn the audio-to-face translation by generative adversarial networks (GANs). However, resolving such a problem is highly challenging because it is non-trivial to faithfully relate the audio signals and face deformations, including expressions and lip motions. Therefore, most of these methods utilize some intermediate face representations including reconstructing explicit 3D face shapes [55] and regressing expression coefficients [43] or 2D landmarks [41, 47]. Due to the information loss caused by the intermediate representation, it might lead to semantic mismatches between original audio signals and the learned face deformations. Moreover, existing audio-driven methods suffer from several limitations, such as only rendering the mouth part [41, 43] or fixed by static head pose [35, 42, 10, 7], thus are not suitable for advanced talking head editing tasks, like pose-manipulation and background-replacement.

To address these issues of existing talking head methods, we turn attention to recent developed neural radiance fields (NeRF). We present AD-NeRF, an audio-driven neural radiance fields model that can handle the cross-modal mapping problem without introducing extra intermediate representation. Different from existing methods which rely on 3D face shape, expression coefficient or 2D landmarks to encode the facial image, we adopt the neural radiance field (NeRF) [30] to represent the scenes of talking heads. Inspired by dynamic NeRF [16] for modeling appearance and dynamics of a human face, we directly map the corresponding audio features to dynamic neural radiance fields to represent the target dynamic subject. Thanks to the neural rendering techniques which enable a powerful ray dispatching strategy, our model can well represent some fine-scale facial components like teeth and hair, and achieves better image qualities than existing GAN-based methods. Moreover, the volumetric representation provides a natural way to freely adjust the global deformation of the animated speakers, which can not be achieved by traditional 2D image generation methods. Furthermore, our method takes the head pose and upper body movement into consideration and is capable of producing vivid talking-head results for real-

[*]This work was done when Yudong Guo and Keyu Chen were intern at Dilusense.

[†]Corresponding author: juyong@ustc.edu.cn.

world applications.

Specifically, our method takes a short video sequence, including the video and audio track of a target speaking person as input. Given the audio features extracted via *DeepSpeech* [1] model and the face parsing maps, we aim to construct an audio-conditional implicit function that stores the neural radiance fields for talking head scene representations. As the movement of the head part is not consistent with that of the upper body part, we further split the neural radiance field representation into two components, one for the foreground face and the other for the foreground torso. In this way, we can generate natural talking-head sequences from collected training data. Please refer to the supplementary video for better visualization of our results.

In summary, the contributions of our proposed talking-head synthesis method contain three main aspects:

- We present an audio-driven talking head method that directly maps the audio features to dynamic neural radiance fields for portraits rendering, without any intermediate modalities that may cause information loss. Ablation studies show that such direct mapping has better capability in producing accurate lip motion results with training data of a short video.
- We decompose the neural radiance fields of human portrait scenes into two branches to model the head and torso deformation respectively, which helps to generate more natural talking head results.
- With the help of audio-driven NeRF, our method enables talking head video editings like pose-manipulation and background-replacement, which are valuable for potential virtual reality applications.

2. Related Work

Audio-driven Facial Animation. The goal of audio-driven facial animation is to reenact a specific person in sync with arbitrary input speech sequences. Based on the applied targets and techniques, it can be categorized into two classes: model-based and data-driven methods. The model-based approaches [39, 12, 60] require expertise works to establish the relationships between audio semantics and lip motions, such as phoneme-viseme mapping [14]. Therefore, they are inconvenient for general applications except for advanced digital creations like movie and game avatars. With the rise of deep learning techniques, many data-driven methods are proposed to generate photo-realistic talking-head results. Earlier methods try to synthesize the lip motions that fulfill the training data of a still facial image [5, 13, 8, 53, 7, 46]. Later it is improved to generate full image frames for President Obama by using quantities of his addressing videos [41]. Based on the developed 3D face reconstruction [19, 11, 50] and generative adversarial networks, more and more approaches are proposed by intermediately estimating 3D face shapes [22, 43, 55] or facial landmarks [56, 47]. In contrast to our method, they require more training data due to the latent modalities, i.e., prior parametric models or low-dimensional landmarks.

Video-driven Facial Animation. Video-driven facial animation is the process of transferring facial pose and expression from a source actor to a target. Most approaches on this task rely on model-based facial performance capture [44, 45, 24, 23]. Thies *et al*. [44] track dynamic 3D faces with RGB-D cameras and then transfer facial expressions from the source actor to the target. Thies *et al*. [45] further improve the pipeline by using RGB cameras only. Kim *et al*. [24] utilize a generative adversarial network to synthesize photo-realistic skin texture that can handle skin deformations conditioned on renderings. Kim *et al*. [23] analyze the notion of style for facial expressions and show its importance for video-based dubbing.

Implicit Neural Scene Networks. Neural scene representation is the use of neural networks for representing the shape and appearance of scenes. The neural scene representation networks (SRNs) was first introduced by Sitzmann *et al*. [40], in which the geometry and appearance of an object is represented as a neural network that can be sampled at points in space. Since from last year, Neural Radiance Fields (NeRF) [30] has gained a lot of attention for neural rendering and neural reconstruction tasks. The underlying implicit representation of the shape and appearance of 3D objects can be transformed into volumetric ray sampling results. Follow-up works extend this idea by using in-the-wild training data including appearance interpolation [29], introducing deformable neural radiance fields to represent non-rigidly moving objects [31, 36], and optimizing NeRF without pre-computed camera parameters [52].

Neural Rendering for Human. Neural rendering for human heads and bodies have also attracted many attentions [15, 28, 27]. With recent implicit neural scene representations [38, 20], Wang *et al*. [51] present a compositional 3D scene representation for learning high-quality dynamic neural radiance fields for upper body. Raj *et al*. [37] adopt pixel-aligned features [38] in NeRF to generalize to unseen identities at test time. Gao *et al*. [17] present a meta-learning framework for estimating neural radiance fields form a single portrait image. Gafni *et al*. [16] propose dynamic neural radiance fields for modeling the dynamics of a human face. Peng *et al*. [33] integrate observations across video frames to enable novel view synthesis for human body from a sparse multi-view video.

3. Method

3.1. Overview

Our talking-head synthesis framework (Fig. 1) is trained on a short video sequence along with the audio track of a target person. Based on the neural rendering idea, we implicitly model the deformed human heads and upper bodies

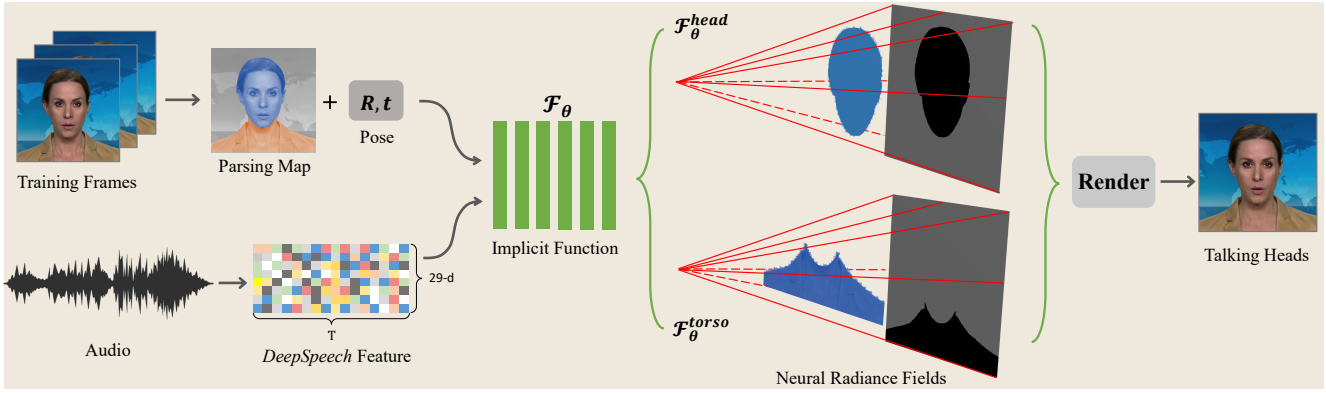

Figure 1. Framework of our proposed talking-head synthesis method. Given a portrait video sequence of a person, we train two neural radiance fields to synthesize high-fidelity talking head with volume rendering.

by neural scene representation, i.e., neural radiance fields. In order to bridge the domain gap between audio signals and visual faces, we extract the semantic audio features and learn a conditional implicit function to map the audio features to neural radiance fields (Sec. 3.2). Finally, visual faces are rendered from the neural radiance fields using volumetric rendering (Sec. 3.3). In the inference stage, we can generate faithful visual features simply from the audio input. Besides, our method can also generate realistic speaking styles of the target person. It is achieved by estimating the neural radiance fields of dynamic heads and upper bodies in a separate manner (Sec. 3.4) with the help of an automatic parsing method [26] for segmenting the head and torso part and extracting a clean background. While we transform the volumetric features into a novel canonical space, the heads and other body parts will be rendered differently with their individual implicit models and thus produce very natural results.

3.2. Neural Radiance Fields for Talking Heads

Based on the standard neural radiance field scene representation [30] and inspired by the dynamic neural radiance fields for facial animation introduced by Gafni *et al.* [16], we present a conditional radiance field of a talking head using a conditional implicit function with an additional audio code as input. Apart from viewing direction $\mathbf{d}$ and 3D location $\mathbf{x}$, the semantic feature of audio $\mathbf{a}$ will be added as another input of the implicit function $\mathcal{F}_\theta$. In practice, $\mathcal{F}_\theta$ is realized by a multi-layer perceptron (MLP). With all concatenated input vectors $(\mathbf{a}, \mathbf{d}, \mathbf{x})$, the network will estimate color values $\mathbf{c}$ accompanied with densities σ along the dispatched rays. The entire implicit function can be formulated as follows:

$$\mathcal{F}_\theta : (\mathbf{a}, \mathbf{d}, \mathbf{x}) \longrightarrow (\mathbf{c}, \sigma). \quad (1)$$

We use the same implicit network structure including positional encoding as NeRF [30].

Semantic Audio Feature. In order to extract the semantically meaningful information from acoustic signals, similar to previous audio-driven methods [10, 43], we employ the popular *DeepSpeech* [1] model to predict a 29-dimensional feature code for each 20ms audio clip. In our implementation, several continuous frames of audio features are jointly sent into a temporal convolutional network to eliminate noisy signals from raw input. Specifically, we use the features $\mathbf{a} \in \mathbb{R}^{16\times 29}$ from the sixteen neighboring frames to represent the current state of audio modality. The usage of audio features instead of regressed expression coefficients [43] or facial landmarks [49] is beneficial for alleviating the training cost of intermediate translation network and preventing potential semantic mismatching issue between audio and visual signals.

3.3. Volume Rendering with Radiance Fields

With the color $\mathbf{c}$ and density σ predicted by the implicit model $\mathcal{F}_\theta$ mentioned above, we can employ the volume rendering process by accumulating the sampled density and RGB values along the rays casted through each pixel to compute the output color for image rendering results. Like NeRF [30], the expected color $\mathcal{C}$ of a camera ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ with camera center $\mathbf{o}$, viewing direction $\mathbf{d}$ and near bound t_n and far bound t_f is evaluated as:

$$\mathcal{C}(\mathbf{r}; \theta, \Pi, \mathbf{a}) = \int_{t_n}^{t_f} \sigma_\theta(\mathbf{r}(t)) \cdot \mathbf{c}_\theta(\mathbf{r}(t), \mathbf{d}) \cdot T(t) dt, \quad (2)$$

where $\mathbf{c}_\theta(\cdot)$ and $\sigma_\theta(\cdot)$ are the outputs of the implicit function $\mathcal{F}_\theta$ described above. $T(t)$ is the accumulated transmittance along the ray from t_n to t:

$$T(t) = exp\left(-\int_{t_n}^{t} \sigma(\mathbf{r}(s)) ds\right). \quad (3)$$

Π is the estimated rigid pose parameters of the face, represented by a rotation matrix $R \in \mathbb{R}^{3\times 3}$ and a translation vector $t \in \mathbb{R}^{3\times 1}$, i.e., $\Pi = \{R, t\}$. Similar to Gafni *et al.* [16], Π is used to transform the sampling points to the canonical space. Note that during the training stage, we only use

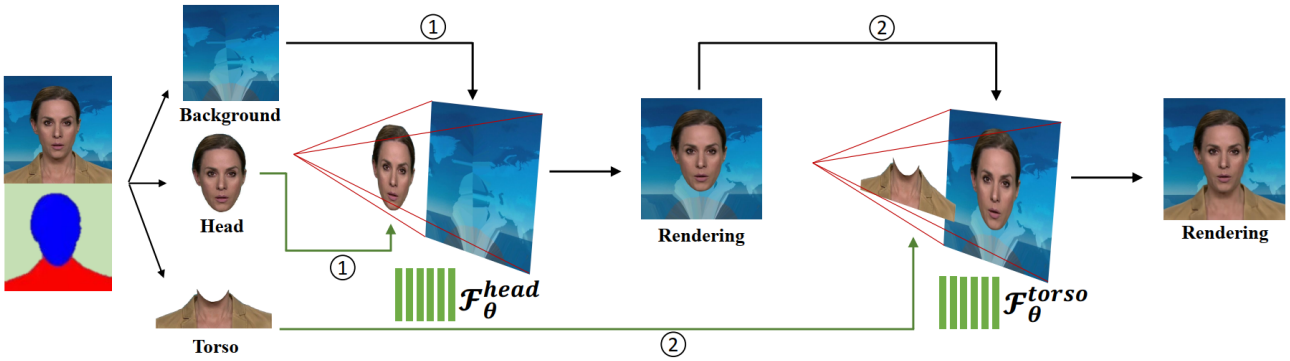

Figure 2. Training process of the two neural radiance fields. We reconstruct the head part and upper-body with Head-NeRF (Step 1) and Torso-NeRF (Step 2) respectively.

the head pose information instead of any 3D face shapes for our network. We use the two-stage integration strategy introduced by Mildenhall *et al*. [30]. Specifically, we first use a coarse network to predict densities along a ray, and then sample more points in areas with high density in the fine network.

3.4. Individual NeRFs Representation

The reason of taking head pose into account for the rendering process is that, compared to the static background, the human body parts (including head and torso) are dynamically moving from frame to frame. Therefore, it is essential to transform the deformed points from camera space to canonical space for radiance fields training. Gafni *et al*. [16] try to handle the dynamic movements by decoupling the foreground and background based on the automatic predicted density, i.e., for dispatched rays passing through the foreground pixels, the human parts will be predicted with high densities while the background images will be ignored with low densities. However, there exist some ambiguities to transform the torso region into canonical space. Since the movement of the head part is not consistent with the movement of the torso part and the pose parameters Π are estimated for the face shape only, applying the same rigid transformation to both the head and torso region together would result unsatisfactory rendering results in the upper body. To tackle this issue, we model these two parts with two individual neural radiance fields: one for the head part and the other for the torso part.

As illustrated in Fig. 2, we initially leverage an automatic face parsing method [26] to divide the training image into three parts: static background, head and torso. We first train the implicit function for the head part $\mathcal{F}_\theta^{head}$. During this step, we regard the head region determined by the parsing map as the foreground and the rest to be background. The head pose Π is applied to the sampled points along the ray casted through each pixel. The last sample on the ray is assumed to lie on the background with a fixed color, namely, the color of the pixel corresponding to the ray, from the background image. Then we convert the rendering image of $\mathcal{F}_\theta^{head}$ to be the new background and make the torso part to be the foreground. Next we continue to train the second implicit model $\mathcal{F}_\theta^{torso}$. In this stage, there are no available pose parameters for the torso region. So we assume all points live in canonical space (i.e., without transforming them with head pose Π) and add the face pose Π to be another input condition (combined with point location $\mathbf{x}$, viewing direction $\mathbf{d}$ and audio feature $\mathbf{a}$) for radiance fields prediction. In other words, we implicitly treat the head pose Π as an additional input, instead of using Π for explicit transformation within $\mathcal{F}_\theta^{torso}$.

In the inference stage, both the head part model $\mathcal{F}_\theta^{head}$ and the torso part model $\mathcal{F}_\theta^{torso}$ accept the same input parameters, including the audio conditional code $\mathbf{a}$ and the pose coefficients Π. The volume rendering process will first go through the head model accumulating the sampled density and RGB values for all pixels. The rendered image is expected to cover the foreground head area on a static background. Then the torso model will fill the missing body part by predicting foreground pixels in the torso region. In general, such an individual neural radiance field representation design is helpful to model the inconsistent head and upper body movements and to produce natural talking head results.

3.5. Editing of Talking Head Video

Since both neural radiance fields take semantic audio feature and pose coefficients as input to control the speaking content and the movement of talking head, our method could enable audio-driven and pose-manipulated talking head video generation by replacing the audio input and adjusting pose coefficients, respectively. Moreover, similar to Gafni *et al*. [16], since we use the corresponding pixel on the background image as the last sample for each ray, the implicit networks learn to predict low density values for the foreground samples if the ray is passing through a background pixel, and high density values for foreground pixels. In this way, our method decouples foreground-background

regions and enables background editing simply by replacing the background image. We further demonstrate all these editing applications in Sec. 4.4.

3.6. Training Details

Dataset. For each target person, we collect a short video sequence with audio track for training. The average video length is 3-5 minutes and all in 25 fps. The recording camera and background are both assumed to be static. In testing, our method allows arbitrary audio input such as speech from different identities, gender and language.

Training Data Preprocessing. There are three main steps to preprocess the training dataset: (1) We adopt an automatic parsing method [26] to label the different semantic regions for each frame; (2) We apply the multi-frame optical flow estimation method [18] to get dense correspondences across video frames in near-rigid regions like forehead, ear and hair, and then estimate pose parameters using bundle adjustment [2]. It is worth noting that the estimated poses are only effective for the face part but not the other body regions like neck and shoulders, i.e., the face poses could not represent the entire movements of upper body; (3) We construct a clean background image without person (as shown in Fig. 2) according to all sequential frames. This is achieved by removing the human region from each frame based on the parsing results and then computing the aggregation results of all the background images. For the missing area, we use Poisson Blending [34] to fix the pixels with neighbor information.

Network & Loss Function. In general, our proposed neural radiance field representation network has two main constraints. The first one is the temporal smooth filter. In Sec. 3.2, we mentioned to process the *DeepSpeech* feature with a window size of 16. The 16 continuous audio features will be sent into a 1D convolutional network to regress the per-frame latent code. In order to assure the stability within audio signals, we adopt the self-attention idea [43] to train a temporal filter on the continuous audio code. The filter is implemented by 1D convolution layers with softmax activation. Hence the final audio condition a is given by the temporally filtered latent code.

Second, we constrain the rendering image of our method to be the same as the training groundtruth. Let $I_r \in \mathbb{R}^{W\times H\times 3}$ be the rendered image and $I_g \in \mathbb{R}^{W\times H\times 3}$ to be the groundtruth, the optmization target is to reduce the photo-metric reconstruction error between I_r and I_g. Specifically, the loss function is formulated as:

$$\mathcal{L}_{photo}(\theta) = \sum_{w=0}^{W}\sum_{h=0}^{H} \|I_r(w,h) - I_g(w,h)\|^2, \quad (4)$$
$$I_r(w,h) = \mathcal{C}(r_{w,h};\theta,\Pi,\mathbf{a})$$

4. Experiments

4.1. Implementation Details

We implement our framework in PyTorch [32]. Both networks are trained with Adam [25] solver with initial learning rate 0.0005. We train each model for $400k$ iterations. In each iteration, we randomly sample a batch of 2048 rays through the image pixels. We train the networks with RTX 3090 and train each model for $400k$ iterations. For a 5-minutes video with resolution 450×450, it takes about 36 hours to train two NeRFs and 12 seconds to render a frame.

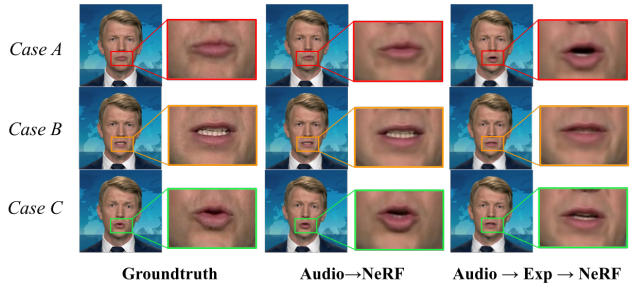

Figure 3. Ablation study on using direct audio or intermediate facial expression representation to condition the NeRF model. It can be observed that direct audio condition has better capability in producing accurate lip motion results.

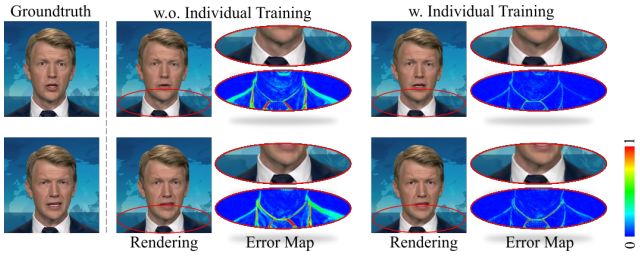

Figure 4. Ablation study on training individual neural radiance field representation for head and torso.

4.2. Ablation Study

We validate two main components adopted in our framework. First, we compare the neural rendering results based on direct audio condition and additional intermediate condition. Second, we explore the beneficial of training separated neural radiance fields for head and torso region.

Audio condition. As aforementioned in Sec. 3.2, our NeRF based talking head model is directly conditioned on audio features to avoid the traininjg cost and information loss within additional intermediate modalities. In Fig. 3, we compare the rendering images generated from audio code and audio-estimated expression code. We use the monocular face tracking method [45] to optimize expression parameters and use the same network structure as Thies *et al.* [43] to estimate expression code from audio. From the illustration results, it can be clearly observed that the audio conditioning is helpful for precise lip synchronization.

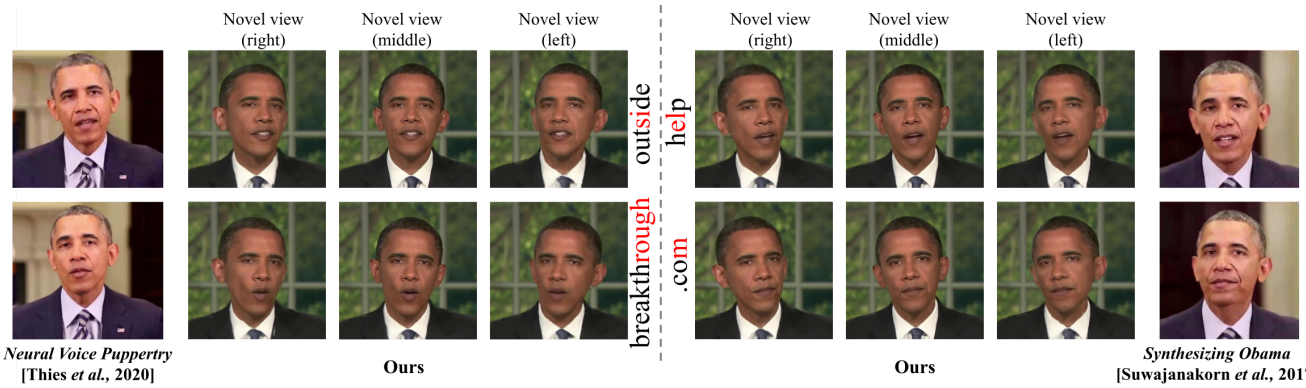

Figure 5. Comparison with model-based methods of Thies *et al*. [43] and Suwajanakorn *et al*. [41]. Our method not only remains the semantics of lip motion, but also supports free adjustment on viewing angles. Please watch our supplementary video for visual results.

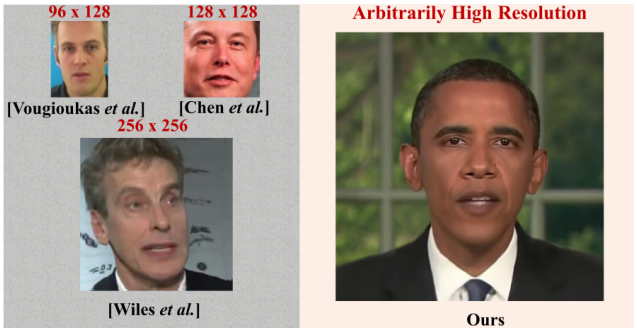

Figure 6. Comparison with image-based methods. The image size decides the image quality of generation results. Please watch our video demo for more results.

Individual training for head and torso region. Another factor we would evaluate is the individual training strategy for head and torso part. To demonstrate the advantages of training two separate neural radiance fields network for these two regions, we conduct an ablative experiment by training a single one NeRF network for the human body movements. In such case, the torso area including neck and shoulders are transformed by the estimated head pose matrices. Therefore there are obviously inaccurate mismatching pixels around the boundary of upper body. We visualize the photo-metric error map of this region for the rendering image and groundtruth. From Fig. 4, the illustrated results prove that our individual training strategy is beneficial for better image reconstruction quality.

We also compute the structural similarity index measure (SSIM) between the generated frames and ground-truth frames on the whole test sequence of 500 frames. The scores are 0.92, 0.88 and 0.87 respectively (higher is better) for our method and the settings of intermediate expression and single NeRF.

4.3. Evaluations

In this section, we compare our method with two categories of talking head synthesis approaches: pure image-based [53, 7, 46] and intermediate model-based [41, 43] ones. We conduct both quantitative and qualitative experiments to evaluate the visualized results generated by each method. In the following, we first summarize the compared methods from two categories and then introduce our designed evaluation metrics.

Comparison with Image-based Method. There are a branch of talking head generation methods [5, 13, 8, 53, 7, 46] entirely lying in the image domain. Recent deep-learning-based approaches are trained for multiple identities and thus can be applied for new target person. However, the limitation of these methods is obvious as they are only capable of producing still face crop images, and differs from our method that produces full-size images with backgrounds and natural taking styles of target person. In Fig. 6, we present the audio-driven facial animation results generated by our method and three competitive methods [53, 7, 46]. It can be clearly observed that the image-based talking head methods are restricted by the input image size and thus could not producing high-resolution imagery as we do.

Comparison with Model-based Method. The model-based method refers to the approach that takes prior information in generating photo-realistic face images. The key component of this categorical methods is the statistical model, e.g., PCA model for mouth textures [41] or 3D morphable model for face shapes [43].

In comparison, we extract the audio from the released demos of the two methods as the input of our framework (we assume the released demos as their best results since both of them did not provide pre-trained model), named as testset A (from *Neural Voice Puppetry* [43]) and testset B (from *SynthesizingObama* [41]). In Fig. 5, we show some selected audio-driven talking head frames from each method. Note that the prior model generally requires large quantities of training data, for example, Suwajanakorn *et al*. [41] reported to use 14 hours high-quality *Obama Addressing* videos for training and Thies *et al*. [43] took more than 3 hours data for training and 2-3 minutes long video for fine-tuning, while our method only requires a short video clip (3-5 minutes) for training. Despite the huge gap of the

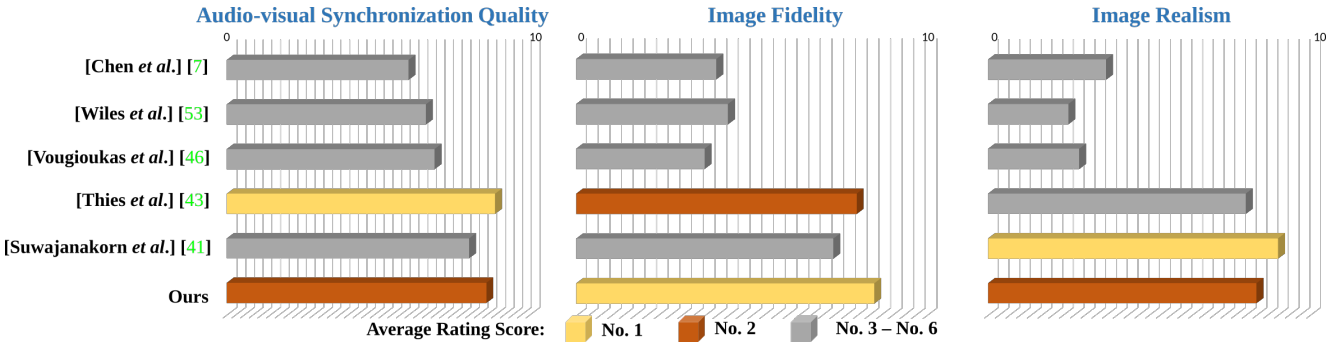

Figure 7. Rating scores from participants. Based on the statics on three different terms, our method achieves comparable results with the other two model-based methods. However, our method only requires a very short video sequence for training, while the other two are trained on multiple and large datasets.

Methods	SyncNet score [9]▲ testset A	SyncNet score [9]▲ testset B	AU error [4]▼ testset A	AU error [4]▼ testset B	Pose	Full-frame	Background
[Chen *et al.*] [7]	6.129	4.388	2.588	3.475	static	×	×
[Wiles *et al.*] [53]	4.257	3.976	3.134	3.127		×	×
[Vougioukas *et al.*] [46]	5.865	6.712	2.156	2.658		×	×
[Thies *et al.*] [43]	4.932	-	1.976	-	copied from	✓	×
[Suwajanakorn *et al.*] [41]	-	5.836	-	2.176	source	✓	×
Ours	5.239	5.411	2.133	2.287	freely adjusted	✓	✓
Original	5.895	6.178	0	0	-	-	-

Table 1. We conduct comparisons on two testsets (A and B) collected from the demos of *Neural Voice Puppetry* [43] and *SynthesizingObama* [41], respectively. ▲ indicates that the confidence value in SyncNet score is better with higher results. ▼ means that AU error is better with smaller numbers. Moreover, our method can synthesize full-frame imagery while enables pose manipulation and background replacement thanks to the audio-driven neural radiance fields.

Figure 8. Comparison with the video-driven method of Kim *et al*. [23]. On the right are the saying words.

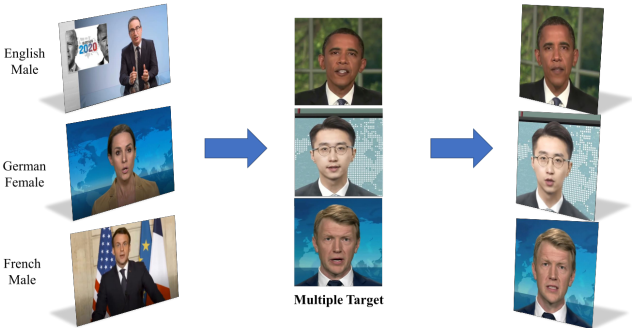

Figure 9. Our method allows arbitrary audio input from different identity, gender and language. For the audio-driven results, please refer to our supplementary video.

training dataset size, our approach is still capable of producing comparable natural results to the other two methods.

Moreover, our method owns the advantage of freely manipulating the viewing directions on the target person, which means that we can freely adjust head poses within the range of training data. We further demonstrate the free-viewing-direction results in Fig. 10 and our supplementary video.

Comparison with Video-driven Method. Besides audio-driven methods, another category of talking head generation methods lie in video-driven, namely driving the target person from a source portrait video. We compare our audio-driven method with a recent style-based video-driven method [23] in Fig. 8. We can see that both methods produce high-fidelity talking head results. Note that the method of Kim et al. [23] takes the video frames as input while our method takes the corresponding audio as input.

Metrics. We employ multiple evaluation metrics to demonstrate the superiority of our method to the others. As an audio-driven talking head generation work, the synchronized visual faces are expected to be consistent with audio input while remaining high image fidelity and realistics. To this end, we propose a combined evaluation design, including SyncNet [9] scores for audio-visual syn-

Figure 10. Our method can generate talking head frames with freely adjusted viewing directions and various background images. Each row from left to right: original frames from a video, reconstructed results with audio and pose from the original video, two samples of background-replacement results, two samples of pose-manipulation results.

chronization quality, Action Units (AU) detection [3] (by OpenFace [4]) for facial action coding consistency between source and generated results, and a diversified user study on image realism, fidelity and synchronization consistency.

SyncNet [9] is commonly used to validate the audio-visual consistency for lip synchronization and facial animation tasks. In this experiment, we use a pretrained SyncNet model to compute the audio-sync offset and confidence of speech-driven facial sequences generated by each comparing method (Tab. 1). Higher confidence values are better.

We employ an action units (AU) detection module by OpenFace [4] to compute the facial action units for the source video that providing audio signals and the corresponding generated results. This metric is aimed at evaluating the muscle activation consistency between the source faces and driven ones. The ideal talking-heads are expected to perform similar facial movements as the source faces. We select the lower face and mouth-related AUs as active subjects and compute the mean errors between source and driven faces. The quantitative results are given in Tab. 1.

Finally, we conduct a user study comparisons with the help of 30 attendees. Each participant is asked to rate the talking-head generation results of 100 video clips (9 from Thies *et al*. [43], 11 from Suwajanakorn *et al*. [41] and 20 from three image-based methods [53, 7, 46] and ours) based on three major aspects: audio-visual synchronization quality, image fidelity and image realism. The head poses for generating results of our method come from a template video clip outside the training set. We collect the rating results within 1 to 10 (the higher the better) and compute the average score that each method gained. The processed statistics are visualized in Fig. 7.

4.4. Applications on Talking Head Editing

As described in Sec. 3.5, our method could enable talking head video editing on audio signal, head movement, and background image. First we show the audio-driven results of the same video with inputs from diverse audio input from different persons in Fig. 9. As we can see, our method produces reasonable results with arbitrary audio input from different identities, gender, and language. Then we show the pose-manipulation and background-replacement results of our method in Fig. 10. We can see that our method allows adjusting viewing directions and various background images replacement for high-fidelity talking portraits synthesis with the trained neural radiance fields. We believe these features would be very exciting for the virtual reality applications like virtual meetings and digital humans.

5. Limitation

We have demonstrated high-fidelity audio-driven talking head synthesis of AD-NeRF. However, our method has limitations. As seen from the supplemental video, for the cross-identity audio-driven results, the synthesized mouth parts sometimes look unnatural due to the inconsistency between the training and driven language. As seen from Fig. 5 and the supplemental video, sometimes the torso parts look blurry due to that the head pose and audio feature cannot totally determine the actual torso movement.

6. Conclusion

We have presented a novel method for high-fidelity talking head synthesis based on neural radiance fields. Using volume rendering on two elaborately designed NeRFs, our method is able to directly synthesize human head and upper body from audio signal without relying on intermediate representations. Our trained model allows arbitrary audio input from different identity, gender and language and supports free head pose manipulation, which are highly demanded features in virtual meetings and digital humans.

Acknowledgement This work was supported by the NSFC (62122071, 61725204), the Youth Innovation Promotion Association CAS (No. 2018495) and "the Fundamental Research Funds for the Central Universities"(No. WK3470000021).

References

[1] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *ICML*, 2016. 2, 3

[2] Alex M Andrew. Multiple view geometry in computer vision. *Kybernetes*, 2001. 5

[3] T. Baltrušaitis, M. Mahmoud, and P. Robinson. Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 2015. 8

[4] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L. Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 59–66, 2018. 7, 8

[5] Christoph Bregler, Michele Covell, and Malcolm Slaney. Video rewrite: driving visual speech with audio. In *SIGGRAPH*, 1997. 2, 6

[6] Yong Cao, Wen C Tien, Petros Faloutsos, and Frédéric Pighin. Expressive speech-driven facial animation. *ACM Transactions on Graphics*, 24(4), 2005. 1

[7] Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *CVPR*, 2019. 1, 2, 6, 7, 8

[8] Joon Son Chung, Amir Jamaludin, and Andrew Zisserman. You said that? In *BMVC*, 2017. 2, 6

[9] J. S. Chung and A. Zisserman. Out of time: automated lip sync in the wild. In *Workshop on Multi-view Lip-reading, ACCV*, 2016. 7, 8

[10] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J Black. Capture, learning, and synthesis of 3d speaking styles. In *CVPR*, 2019. 1, 3

[11] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *CVPRW*, 2019. 2

[12] Pif Edwards, Chris Landreth, Eugene Fiume, and Karan Singh. Jali: an animator-centric viseme model for expressive lip synchronization. *ACM Transactions on Graphics*, 35(4), 2016. 1, 2

[13] Tony Ezzat, Gadi Geiger, and Tomaso Poggio. Trainable videorealistic speech animation. *ACM Transactions on Graphics*, 21(3), 2002. 2, 6

[14] Cletus G Fisher. Confusions among visually perceived consonants. *Journal of speech and hearing research*, 11(4), 1968. 2

[15] Ohad Fried, Ayush Tewari, Michael Zollhöfer, Adam Finkelstein, Eli Shechtman, Dan B Goldman, Kyle Genova, Zeyu Jin, Christian Theobalt, and Maneesh Agrawala. Text-based editing of talking-head video. *ACM Transactions on Graphics (TOG)*, 2019. 2

[16] Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *CVPR*, 2021. 1, 2, 3, 4

[17] Chen Gao, Yichang Shih, Wei-Sheng Lai, Chia-Kai Liang, and Jia-Bin Huang. Portrait neural radiance fields from a single image. *arXiv preprint arXiv:2012.05903*, 2020. 2

[18] Ravi Garg, Anastasios Roussos, and Lourdes Agapito. A variational approach to video registration with subspace constraints. *International journal of computer vision*, 104(3):286–314, 2013. 5

[19] Yudong Guo, Jianfei Cai, Boyi Jiang, Jianmin Zheng, et al. Cnn-based real-time dense face reconstruction with inverse-rendered photo-realistic face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(6), 2018. 2

[20] Yang Hong, Juyong Zhang, Boyi Jiang, Yudong Guo, Ligang Liu, and Hujun Bao. Stereopifu: Depth aware clothed human digitization via stereo vision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[21] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Wayne Wu, Chen Change Loy, Xun Cao, and Feng Xu. Audio-driven emotional video portraits. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14080–14089, 2021. 1

[22] Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics*, 36(4), 2017. 2

[23] Hyeongwoo Kim, Mohamed Elgharib, Hans-Peter Zollöfer, Michael Seidel, Thabo Beeler, Christian Richardt, and Christian Theobalt. Neural style-preserving visual dubbing. *ACM Transactions on Graphics (TOG)*, 2019. 2, 7

[24] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Nießner, Patrick Pérez, Christian Richardt, Michael Zollöfer, and Christian Theobalt. Deep video portraits. *ACM Transactions on Graphics (TOG)*, 37(4):163, 2018. 2

[25] Diederick P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 5

[26] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *CVPR*, 2020. 3, 4, 5

[27] Lincheng Li, Suzhen Wang, Zhimeng Zhang, Yu Ding, Yixing Zheng, Xin Yu, and Changjie Fan. Write-a-speaker: Text-based emotional and rhythmic talking-head generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1911–1920, 2021. 2

[28] Lingjie Liu, Weipeng Xu, Marc Habermann, Michael Zollhöfer, Florian Bernard, Hyeongwoo Kim, Wenping Wang, and Christian Theobalt. Neural human video rendering by learning dynamic textures and rendering-to-video translation. *IEEE Transactions on Visualization and Computer Graphics*, 2020. 2

[29] Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. In *CVPR*, 2021. 2

[30] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 1, 2, 3, 4

[31] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Deformable neural radiance fields. *arXiv preprint arXiv:2011.12948*, 2020. 2

[32] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019. 5

[33] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 2

[34] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. In *ACM SIGGRAPH 2003 Papers*, pages 313–318. 2003. 5

[35] Hai X Pham, Samuel Cheung, and Vladimir Pavlovic. Speech-driven 3d facial animation with implicit emotional awareness: A deep learning approach. In *CVPRW*, 2017. 1

[36] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. pages 10318–10327, June 2021. 2

[37] Amit Raj, Michael Zollhofer, Tomas Simon, Jason Saragih, Shunsuke Saito, James Hays, and Stephen Lombardi. Pixel-aligned volumetric avatars. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 2

[38] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2304–2314, 2019. 2

[39] Oliver Schreer, Roman Englert, Peter Eisert, and Ralf Tanger. Real-time vision and speech driven avatars for multimedia applications. *IEEE Transactions on Multimedia*, 10(3), 2008. 2

[40] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *NeurIPS*, 2019. 2

[41] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics*, 36(4), 2017. 1, 2, 6, 7, 8

[42] Sarah Taylor, Taehwan Kim, Yisong Yue, Moshe Mahler, James Krahe, Anastasio Garcia Rodriguez, Jessica Hodgins, and Iain Matthews. A deep learning approach for generalized speech animation. *ACM Transactions on Graphics*, 36(4), 2017. 1

[43] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural voice puppetry: Audio-driven facial reenactment. In *ECCV*, 2020. 1, 2, 3, 5, 6, 7, 8

[44] Justus Thies, Michael Zollhöfer, Matthias Nießner, Levi Valgaerts, Marc Stamminger, and Christian Theobalt. Real-time expression transfer for facial reenactment. *ACM Trans. Graph.*, 34(6):183–1, 2015. 2

[45] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016. 2, 5

[46] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Realistic speech-driven facial animation with gans. *International Journal of Computer Vision*, pages 1–16, 2019. 2, 6, 7, 8

[47] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *ECCV*, 2020. 1, 2

[48] Suzhen Wang, Lincheng Li, Yu Ding, Changjie Fan, and Xin Yu. Audio2head: Audio-driven one-shot talking-head generation with natural head motion. In *International Joint Conference on Artificial Intelligence*, 2021. 1

[49] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. pages 10039–10049, June 2021. 3

[50] Xueying Wang, Yudong Guo, Bailin Deng, and Juyong Zhang. Lightweight photometric stereo for facial details recovery. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[51] Ziyan Wang, Timur Bagautdinov, Stephen Lombardi, Tomas Simon, Jason Saragih, Jessica Hodgins, and Michael Zollhofer. Learning compositional radiance fields of dynamic human heads, June 2021. 2

[52] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. NeRF−−: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064*, 2021. 2

[53] Olivia Wiles, A Sophia Koepke, and Andrew Zisserman. X2face: A network for controlling face generation using images, audio, and pose codes. In *ECCV*, 2018. 2, 6, 7, 8

[54] Lance Williams. Performance-driven facial animation. In *ACM SIGGRAPH 2006 Courses*. 2006. 1

[55] Ran Yi, Zipeng Ye, Juyong Zhang, Hujun Bao, and Yong-Jin Liu. Audio-driven talking face video generation with natural head pose. *arXiv:2002.10137*, 2020. 1, 2

[56] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *ICCV*, 2019. 2

[57] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. 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*, pages 3661–3670, 2021. 1

[58] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. Talking face generation by adversarially disentangled audio-visual representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9299–9306, 2019. 1

[59] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition, pages 4176–4186, 2021. 1

[60] Yang Zhou, Zhan Xu, Chris Landreth, Evangelos Kalogerakis, Subhransu Maji, and Karan Singh. Visemenet: Audio-driven animator-centric speech animation. *ACM Transactions on Graphics*, 37(4), 2018. 1, 2