Context-Aware Talking-Head Video Editing

Songlin Yang
University of Chinese Academy of Sciences
Beijing, China
yangsonglin2021@ia.ac.cn

Wei Wang*
Institute of Automation, Chinese Academy of Sciences
Beijing, China
wwang@nlpr.ia.ac.cn

Jun Ling
Shanghai Jiao Tong University
Shanghai, China
lingjun@sjtu.edu.cn

Bo Peng
Institute of Automation, Chinese Academy of Sciences
Beijing, China
bo.peng@nlpr.ia.ac.cn

Xu Tan
Microsoft Research Asia
Beijing, China
xuta@microsoft.com

Jing Dong
Institute of Automation, Chinese Academy of Sciences
Beijing, China
jdong@nlpr.ia.ac.cn

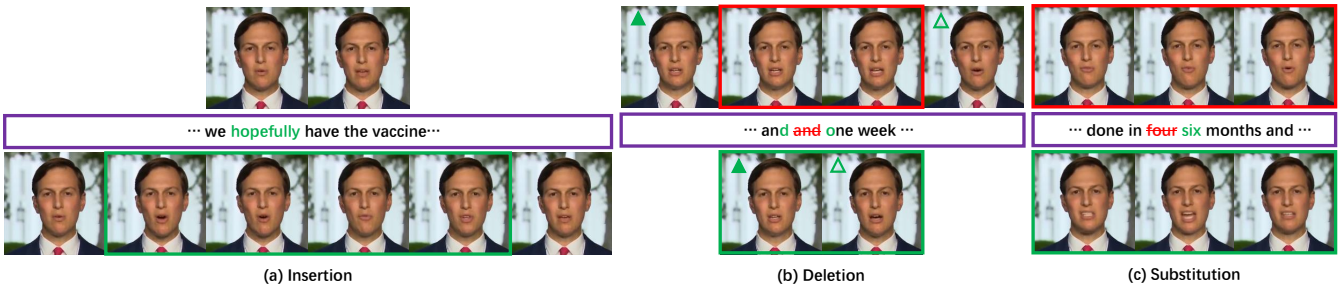

Figure 1: Our results of editing a pre-recorded talking-head video using only few seconds of source video, including insertion, deletion, and substitution. The frames marked as green are generated by our method. For deletion, the smoothing of the two disconnected frames is necessary and we re-generate the disconnected frames to keep the sequential motion smoothness. For insertion and substitution, the edited frames should not only ensure accurate lip synchronization, but also keep the motion smoothness between the edited video clip and the original video frame sequence.

ABSTRACT

Talking-head video editing aims to efficiently insert, delete, and substitute the word of a pre-recorded video through a text transcript editor. The key challenge for this task is obtaining an editing model that generates new talking-head video clips which simultaneously have accurate lip synchronization and motion smoothness. Previous approaches, including 3DMM-based (3D Morphable Model) methods and NeRF-based (Neural Radiance Field) methods, are sub-optimal in that they either require minutes of source videos and days of training time or lack the disentangled control of verbal (e.g., lip motion) and non-verbal (e.g., head pose and expression) representations for video clip insertion. In this work, we fully utilize the video context to design a novel framework for talking-head video editing, which achieves efficiency, disentangled motion control, and sequential smoothness. Specifically, we decompose this framework to motion prediction and motion-conditioned rendering: **(1)** We first design an animation prediction module that efficiently obtains smooth and lip-sync motion sequences conditioned on the driven speech. This module adopts a non-autoregressive network to obtain context prior and improve the prediction efficiency, and it learns a speech-animation mapping prior with better generalization to novel speech from a multi-identity video dataset. **(2)** We then introduce a neural rendering module to synthesize the photo-realistic and full-head video frames given the predicted motion sequence. This module adopts a pre-trained head topology and uses only few frames for efficient fine-tuning to obtain a person-specific rendering model. Extensive experiments demonstrate that our method efficiently achieves smoother editing results with higher image quality and lip accuracy using less data than previous methods.

*Corresponding author.

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3611765

CCS CONCEPTS

• **Information systems** → **Multimedia content creation**; • **Computing methodologies** → *Computer vision*; *Image-based rendering*.

KEYWORDS

Talking face/head synthesis; Video editing; Face animation; Neural radiance field (NeRF)

ACM Reference Format:
Songlin Yang, Wei Wang, Jun Ling, Bo Peng, Xu Tan, and Jing Dong. 2023. Context-Aware Talking-Head Video Editing. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3581783.3611765

1 INTRODUCTION

Are you crazy about the small mistakes in your nearly-perfect presentation videos? Here we are! As shown in Fig. 1 (please find more video examples on this website*), using transcripts to edit talking-head videos [3, 14, 50] provides a user-friendly interface to insert, delete, and substitute words of the pre-recorded videos. Specifically, given a source video, the time-aligned transcript can be first obtained using speech-to-text model [1, 19]. Then, given the starting frame and duration of the to-be-edited words, the off-the-shelf text-to-speech tools [34, 40] and talking-head video editing models [15, 16, 27, 53, 54] can be adopted to synthesize edited speech and video frames. However, compared with speech editing, existing models for talking-head video editing are not satisfactory enough to achieve plausible results, leaving much room for improvement.

Previous methods [14, 50] for text-based talking-head video editing generally require hours of source videos for training. However, in many cases, what a user needs is a tool to edit several words in a short video clip (e.g., several seconds), which means only seconds of video are available for training. It is annoying for users to collect longer videos if they receive unsatisfactory edited results. Therefore, efficiently achieving talking-head video editing with only few video frames and little training time is promising. Moreover, we need an editing model that can not only generate new video clips which have accurate lip synchronization, but also obtain smooth lip motion, expressions, and head poses with the adjacent original frames at the editing point. Unfortunately, this challenge remains.

In order to achieve talking-head video editing, the common strategy is to train a talking-head video generative model to generate video frames that fulfill the editing purpose. As shown in Tab. 1, previous approaches can be split into three categories:

(1) The methods [6, 7, 32, 56, 57] of the first group are often audio-driven, non-NeRF (Neural Radiance Field [29]), and fully-convolution based. While these methods are trained on a large dataset that consists of thousands of different identities, they cannot synthesize visually comparable results with person-specific methods. Besides, the lack of explicit control over the head poses and expressions makes it difficult for them to edit the non-verbal motions to be smoothly inserted into original videos.

(2) The methods [17, 25, 36] in the second group adopt NeRF as the basic rendering model, which can synthesize personalized full-head images when trained on few minutes or even seconds of video data. Albeit the merits of data efficiency, they actually need to sample the pose sequence for face reenactment (i.e., they only can drive the target identity with the original pose sequence and novel audio sequence). If using them for video editing, we have to elaborately sample a pose sequence at a high labor cost, and these methods still fall short in generating motion-smoothing edited results and lacking fine-grained control over the verbal and non-verbal motions.

Method	Needed Data	Time Cost	Motion Smooth	Full -Head
ATVG [7]	-	-	✗	Face
MakeitTalk [57]	-	-	✗	Face
PC-AVS [56]	-	-	✗	Cropped head
Wav2Lip [32]	-	-	✗	Lip
AD-NeRF [17]	5 min	36 h	✗	✔
SSP-NeRF [25]	5 min	36 h	✗	✔
DFRF [36]	15 s	2 h	✗	✔
Fried et al. [14]	1 h	42 h	✗	Lower face
Yao et al. [50]	2.5 min	17 h	✗	Lower face
Ours	**15 s**	**25 min**	✔	✔

Table 1: The comparisons of representative methods related to talking-head video editing.

(3) The third group, text-based talking-head video editing methods [14, 50] are the closest methods to ours. They adopt a 3D face model to disentangle the head pose and expression, and use fully-convolution networks to synthesize face images. However, these methods require at least minutes of source videos and hours of training time, which is impractical in application scenarios.

To tackle these problems, we propose a novel framework for talking-head video editing, which fully utilizes context awareness to improve editing efficiency and smoothness. This framework consists of an animation prediction module and a neural rendering module. The animation prediction module is designed to predict the animation feature sequence corresponding to the edited word for rendering. In particular, this module embeds the animation condition as a latent feature sequence and then adopts the non-autoregressive Transformer [46] architecture to efficiently train on a large video dataset. It fully utilizes the context information of the original video for motion smoothness, and the prior knowledge of a large video dataset for lip accuracy at the same time. We then propose our neural rendering module. This module adopts a pre-trained NeRF-based head topology and uses only few frames for efficient fine-tuning to obtain a person-specific rendering model for generating photo-realistic and full-head images. Moreover, the disentangled animation features (i.e., the features of the lip, head pose, and expression have been disentangled) makes it reduce the data requirements for aligning the speech and non-verbal visual content. And the explicit motion control makes it become a practical tool for smooth video clip insertion.

Our main contributions are summarized as follows:

- We propose a novel video editing framework to realize word-level talking-head video editing with prominently superior lip accuracy and motion smoothness.
- We design two effective components, an animation prediction module, and a neural rendering module, to fully utilize context features around the edited clips, facilitating the model performance on seconds of source video data.
- Extensive experiments demonstrate that our method can achieve smoother editing with higher quality and less training data than previous methods.

*Project page: https://songlin1998.github.io/THEdit/

2 RELATED WORK

2.1 Talking-Head Video Synthesis and Editing

Video Synthesis. Talking-head video synthesis aims to generate avatar videos speaking the conditioned audio. The intuitive way is directly using the audio features as a condition to generate the frame image in the 2D setting [20, 32, 38, 47, 48, 53, 55, 58], which usually adopt GANs or image-to-image translation as the core technologies. For better-structured control-ability, some works [7, 12, 39] use 2D landmarks as the intermediate representation to guide the audio-driven generation. Other works [11, 21, 31, 43] adopt 3D prior for explicit and more fine-grained control of the faces. Taking advantage of 3D structure modeling, these approaches can achieve more natural talking style than 2D methods. However, since their networks are optimized on a specific identity for idiosyncrasies learning, per-identity training on a large dataset is needed. Recently, the Neural Radiance Field [29] based on volume rendering [22] inspires more works [17, 25, 36, 42] to make improvements in photo-realistic rendering, full-frame synthesis, and 3D consistency. To sum up, the development of the rendering model prompts the flexible control of the audio-driven face animation, and more fine-grained control of the faces is the goal that the researchers pursue constantly.

Video Editing. The word-level video editing is based on video synthesis but has higher requirements in the training data and motion transition smoothness between the edited clip and the original sequence. Fried et al. [14] introduced text-based editing for the talking-head videos based on a viseme-phoneme dictionary. However, their work requires one hour of the target video and takes hours to produce a result. Yao et al.. [50] used neural retargeting to handle the data problem and proposed an iterative talking-head video editing tool. However, this method still requires 2∼3 minutes of the source video to fit the 3D parametric model. These two methods are limited to the editing of the lower face region, which fails to edit the full head pose and expression. Other methods like VideoReTalking [8] also try to edit the expressions of the video.

2.2 Text-to-Speech Editing and Speech-to-Text

Recent text-to-speech (TTS) editing has been developed rapidly. Tang et al. [41] proposed a text-based speech editing method that can generate seamlessly inserted speech. RetrieverTTS [51] is a paradigm for text-based speech insertion that enables any-length insertion with naturalness and speaker similarity. Speech recognition (or called Speech-to-Text, STT) also has been studied for decades, and deep learning has greatly improved its performance. Previous works [1, 19] have been able to handle a diverse variety of speech including noisy environments, accents, and languages. Speech editing is not the focus of our work. We adopt TTS editing to synthesize the edited speech and the STT tool to estimate the starting point and duration of every word for audio-text alignment.

2.3 Neural Radiance Field

Neural Radiance Field [29] (NeRF) provides an implicit information storage mode for 3D space, which uses 3D position and view direction to query a fully-connected network to obtain the needed color and density for volume rendering to render a pixel. NeRF can not only render photo-realistic images of static scenes, but also be extended to the objects which have dynamic and non-rigid deformations, especially human faces [28]. NeRFace [15] intuitively encode the expression parameters into the NeRF for dynamic faces rendering, while these works [30, 33, 45] disentangle the face representation as deformed space and canonical space. However, the core of NeRF is adopting 3D-aware modeling to overfit a scene, and a large dataset is needed to make the model implicitly learn the 3D modeling process. To tackle this data requirement for a few-shot application, DFRF [36] conditions the face radiance field on 2D appearance images to learn the face prior and obtain a generalized talking head synthesis method with few training data. This can render photo-realistic results without the burden of efficiency.

3 METHOD

Our goal is to learn a talking-head video editing model for an arbitrary person in few-shot settings because a large corpus of source videos is often impractical for most applications. Under such settings, several challenges need to be addressed: (1) lack of enough paired audio-motion data from the target identity for accurate animation predictions, (2) the identity cannot be well-preserved between edited frames and adjacent original video frames, and (3) lip, pose and expression-related motions among adjacent frames and edited frames of editing point are not smooth.

To mitigate these issues for talking-head video editing, we propose a novel framework. As illustrated in Fig 2, our framework consists of a ***Pre-Alignment*** module, an ***Animation Prediction*** module, and a ***Neural Rendering*** module. Taking video insertion as an example, the input of our framework includes sequential video frames and novel edited speech segments "We all don't like apples", where we insert word "don't" into original scripts "We all like apples". The pre-alignment module (Sec. 3.1) is first introduced to extract the visual (and acoustic) features from the video frames (and speech segments) surrounding the editing point. The animation prediction module (Sec. 3.2) estimates accurate and smooth animation feature sequences. Taking input of sequential animation features and context video frames, the neural rendering module (Sec. 3.3) is capable of generating new edited frames which are both photo-realistic and motion-smoothing with adjacent original video frames. Finally, we detail the implementation guidance of insertion, deletion, and substitution of real applications in Sec. 3.4.

3.1 Pre-Alignment

To edit a video with a given transcript, the first step is to determine the editing point. And then prepare a new speech segment and extract the context video frames for the animation prediction module and neural rendering module, respectively. The new speech segment can be generated by the text-to-speech tools [34, 40, 51] while the adjacent video frames are extracted from the target video in 25 fps (Frames Per Second).

To ensure lip accuracy and motion smoothness of the edited frames with the adjacent video frames, it is non-trivial to generate talking videos while aligning the pose/expressions of generated frames with smooth motions surrounding the editing point. To alleviate this problem, we represent the motions of talking-head videos with verbal (lip-related) and non-verbal (head pose and expression) features. The former consists of two parts, one is extracted from the speech to be edited via DeepSpeech [1] as speech features, and the other is the lip-related features extracted from adjacent video frames via lip reading model LipNet [2] for context information embedding. The latter is extracted from adjacent video frames

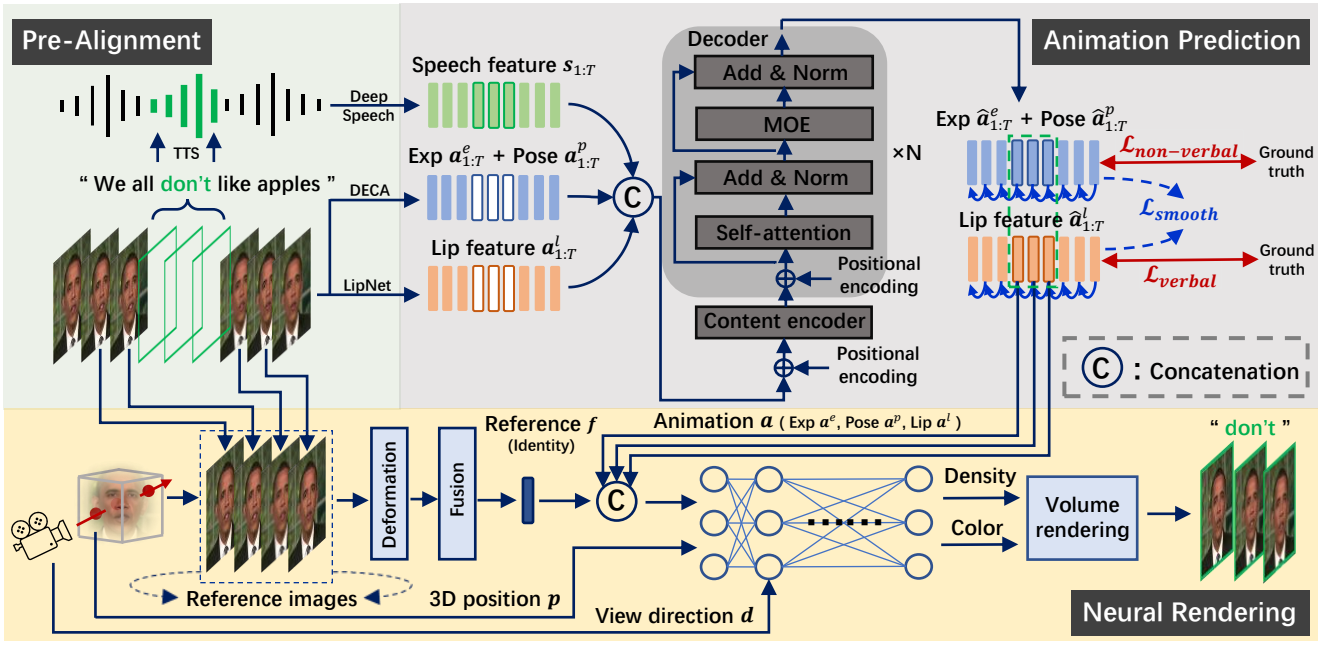

Figure 2: Method overview.

with DECA [13] to constrain the head pose and expression of the predicted faces.

3.2 Animation Prediction

With the disentangled representations of expressions (a^e_t), poses (a^p_t), and lip features (a^l_t) from video frames (those features are further concatenated as feature sequence, denoted as $(a_1, ..., a_T)$, and T is the sequence length), we can synthesize talking-head videos with neural rendering module (Sec. 3.3). However, since the edited speech is changeable according to different editing operations by the user, the edited lip features cannot be well associated with the adjacent video clips in poses, expressions, and lips for all edited speech. How to grasp the subtle variations in different speech and ensure smooth translation at the beginning and end of edited clips remains unsolved. To combat these, we propose to consider contextual information in our animation prediction module from Context-Aware Input Format and Sequential Modeling:

Context-Aware Input Format. Our first solution is building a feature sequence that consists of not only the edited speech and contextual speech, but also the pose, expression, and lip features from adjacent frames. Such input format has two benefits. First, the speech-to-lip mapping is personalized and associated with one's 'talking style'. Aligning the speech features with original lip features provides our animation prediction model with more speech-to-lip feature pairs and helps to model such dynamics for edited speech. Second, involving the contextual pose and expression features from visual frames guides the animation prediction model to synthesize smooth motions around the beginning and end point of the edited speech, thus ensuring better smoothness for the follow-up models.

Sequential Modeling. Modeling short- and long-range correlations for better accuracy and smoothness is very common in sequence-to-sequence learning. Previously, some representative methods adopt the autoregressive generation strategy [23, 37] (i.e., generating the current frame by looking at the previous frames), or introduce a temporal discriminator [7, 47]. However, these methods are less optimal in that (1) the less-satisfying model performance in sequence-to-sequence learning and (2) they are non-parallel and not efficient in fast training and inference. To efficiently model the sequential lip features along with speech features, pose, and expression, we adopt the non-autoregressive strategy to model the context information of the animation feature sequence. Specifically, we utilize a transformer-based network [46] to capture the short- and long-range correlations within the feature sequence.

We optimize our model in a self-supervised manner. To generate data pairs for training, we randomly mask some words and preserve speech features while replacing the corresponding pose, expression, and lip features, with a mean value of its contextual frames. The animation prediction module is expected to predict correct pose, expression, and lip features which not only match the speech content but also ensure the smooth translation at the beginning and end of editing subclips. Without loss of generality, we define the input speech feature as $(s_1, s_2, ..., s_T)$, each of which is acquired by DeepSpeech [1] and an MLP projector. The feature of edited speech can be written as $(s_1, ..., \bar{s}_j, ..., \bar{s}_k, ..., s_T)$, $1 \leqslant j \leqslant k \leqslant T$ if the video frames from j-th to k-th are expected to be edited. We denote the animation prediction module as $\mathcal{P}$, and the animation prediction can be formulated as follows:

$$\mathcal{P}(S, A) = (\hat{a}_1, \hat{a}_2, ..., \hat{a}_T), \tag{1}$$

where we assign speech feature sequence as $S = (s_1, ..., \bar{s}_j, ..., \bar{s}_k, ..., s_T)$ and animation feature sequence as $A = (a_1, ..., \bar{a}_j, ..., \bar{a}_k, ..., a_T)$.

We adopt the average vector of the adjacent frames as the vectors in $\{\bar{s}_t\}$ and $\{\bar{a}_t\}$ where $j \leqslant t \leqslant k$. In the inference stage, we predict

the $\{\hat{\boldsymbol{a}}_{\boldsymbol{t}}\}$ where $1 \leqslant t \leqslant T$. We then extract $\{\hat{\boldsymbol{a}}_{\boldsymbol{t}}\}$ where $j \leqslant t \leqslant k$ as the animation feature sequence of the edited words.

We use the L_2 loss for the animation features to optimize $\mathcal{P}$. To enforce motion smoothness, we add L_2 regularization on the gradient of the predicted sequence. These training objectives can be formulated as follows:

$$\mathcal{L}_{verbal} = \sum_{1 \leqslant t \leqslant T} \|\boldsymbol{a}_t^l - \hat{\boldsymbol{a}}_t^l\|_2, \tag{2}$$

$$\mathcal{L}_{non\text{-}verbal} = \sum_{1 \leqslant t \leqslant T} (\|\boldsymbol{a}_t^e - \hat{\boldsymbol{a}}_t^e\|_2 + \|\boldsymbol{a}_t^p - \hat{\boldsymbol{a}}_t^p\|_2), \tag{3}$$

$$\mathcal{L}_{smooth} = \sum_{1 < t \leqslant T} \|\hat{\boldsymbol{a}}_{t+1} - \hat{\boldsymbol{a}}_t\|_2. \tag{4}$$

The total loss for animation prediction is written as:

$$\mathcal{L}_{animation} = \mathcal{L}_{verbal} + \lambda_1 \mathcal{L}_{non\text{-}verbal} + \lambda_2 \mathcal{L}_{smooth}. \tag{5}$$

Finally, our animation prediction module has three capabilities: the copy ability of the available context frame, the animation prediction of the edited frame, and the smoothness of this sequence.

3.3 Neural Rendering

Compared with the 3D face parametric models [4, 14, 24, 50], the neural rendering models such as NeRF [17, 25, 29, 36] provide a more flexible way to generate photo-realistic full-head frames. However, previous speech-driven NeRF-based methods have two problems:

Many-to-Many Mapping. One audio can be said with different expressions, while one expression can also say different audios. This results in requiring more data to learn this many-to-many mapping from speech signal to pose and expression. Moreover, using speech features as a condition, which entangles lots of content-unrelated audio information such as loudness, further increases the difficulty.

Context-Aware Insertion. The edited frames can be smoothly inserted into the original video only when their motions are consistent with the adjacent frames. The existing NeRF-based talking-head synthesis methods [17, 25, 36] often generate the visual frames conditioned on speech features, where the inserted speech fails to control the non-verbal motion information for a smooth transition.

We propose our head rendering model to tackle the above problems, which consists of a facial neural radiance field and a deformation module [36]. The facial neural radiance field $\mathcal{F}$ is based on the volume rendering, which takes in a 3D point query $\boldsymbol{p} = (x, y, z) \in \mathbb{R}^3$, a 2D view direction $\boldsymbol{d} = (\theta, \phi) \in \mathbb{R}^2$, and condition features including animation $\boldsymbol{a}$ and reference $\boldsymbol{f}$. Then, the $\mathcal{F}$ uses an MLP to predict the color $\boldsymbol{c} \in \mathbb{R}^3$ and the density $\sigma \in \mathbb{R}$ to render a pixel of the target image, which can be formulated as:

$$(\boldsymbol{c}, \sigma) = \mathcal{F}(\boldsymbol{p}, \boldsymbol{d}, \boldsymbol{a}, \boldsymbol{f}). \tag{6}$$

The context-aware animation feature $\boldsymbol{a}$ is adopted from the animation prediction module (Sec. 3.2), which concatenates expression $\boldsymbol{a}^e \in \mathbb{R}^{50}$, head pose $\boldsymbol{a}^p \in \mathbb{R}^6$, and lip $\boldsymbol{a}^l \in \mathbb{R}^{72}$. As for the reference feature $\boldsymbol{f} \in \mathbb{R}^{128}$, we first sample M frames adjacent to the editing frames to transform sampled frames to the image feature map $\boldsymbol{I} \in \mathbb{R}^{M \times H \times W \times 128}$, where H and W are the height and width of the reference frames. We denote the implicit feature of every pixel in $\boldsymbol{I}$ as $\boldsymbol{f}_i \in \mathbb{R}^{128}$. Then, we take a three-layer MLP as deformation module $\mathcal{D}$ to predict the offset $\boldsymbol{o}_i \in \mathbb{R}^2$ between the target pixel and every reference pixel conditioned on the animation feature of the target frame as follows:

$$\boldsymbol{o}_i = \mathcal{D}(\boldsymbol{p}, \boldsymbol{a}^e, \boldsymbol{a}^p, \boldsymbol{a}^l, \boldsymbol{f}_i), \tag{7}$$

We adopt the differentiable trick [36] and attention-based feature fusion module [26] to obtain the final reference feature $\boldsymbol{f}$.

Finally, we use volume rendering to integrate these pixels into portrait images. The background, torso, and neck parts are regarded as the background and are restored frame by frame. The foreground head part follows the volume rendering. The accumulated color C of a camera ray $\boldsymbol{r} \in \mathcal{R}$ under the condition of animation and reference feature can be formulated as follows:

$$C(\boldsymbol{r}) = \int_{v_{near}}^{v_{far}} \sigma(v, \boldsymbol{a}, \boldsymbol{f}) \boldsymbol{c}(v, \boldsymbol{d}, \boldsymbol{a}, \boldsymbol{f}) T(v) dv, \tag{8}$$

where $\boldsymbol{a}$ is concatenated by $\boldsymbol{a}^e$, $\boldsymbol{a}^p$, and $\boldsymbol{a}^l$. The

$$T(v) = exp(-\int_{v_{near}}^{v} \sigma(\boldsymbol{r}(v)) dv), \tag{9}$$

is the integral transmittance along the camera ray $\boldsymbol{r} \in \mathcal{R}$, where v_{near} and v_{far} are the near and far bound of the camera ray. We use the L_2 loss as the optimization objective to minimize the distance between the predicted color $\hat{C}$ and its ground truth C:

$$\mathcal{L}_{rendering} = \sum_{\boldsymbol{r} \in \mathcal{R}} \|C(\boldsymbol{r}) - \hat{C}(\boldsymbol{r})\|_2. \tag{10}$$

3.4 Implementations of Video Editing

For better technical reproducibility, we detail the editing operations of insertion, deletion, and substitution as follows:

Insertion. The starting frame will be located, and the novel speech with its duration of this inserted word will be obtained from the TTS tool. Use the animation prediction module to obtain its corresponding animation feature sequence, and then use the neural rendering module to obtain the edited visual frame sequence.

Deletion. The frames corresponding to the target word will first be deleted. Then, the disconnected visual frames will be re-generated to obtain the smoothed visual frames.

Substitution. The target word to be altered will first be located with its starting point and duration. Use the same pipeline of the insertion operation to obtain the frames of the novel word to insert into the original frame sequence.

4 EXPERIMENTS

4.1 Experimental Settings

Network Architectures. The animation prediction module consists of a content encoder and several decoders. The sequence length T is set as 50. The content encoder is stacked with the blocks in the Transformer encoder [46], which is composed of a stack of $N = 6$ identical layers and each has a multi-head self-attention layer and a position-wise feed-forward network. The decoder is constructed by several Transformer blocks based on Mixture-of-Experts (MOE) layer [35]. This MOE layer consists of multiple experts and a gating network. The experts share the same architecture but have different parameters, which each is a convolution layer with a fully connected layer. The output of the gating network is a multi-dimensional vector, and then the experts with top k values will be

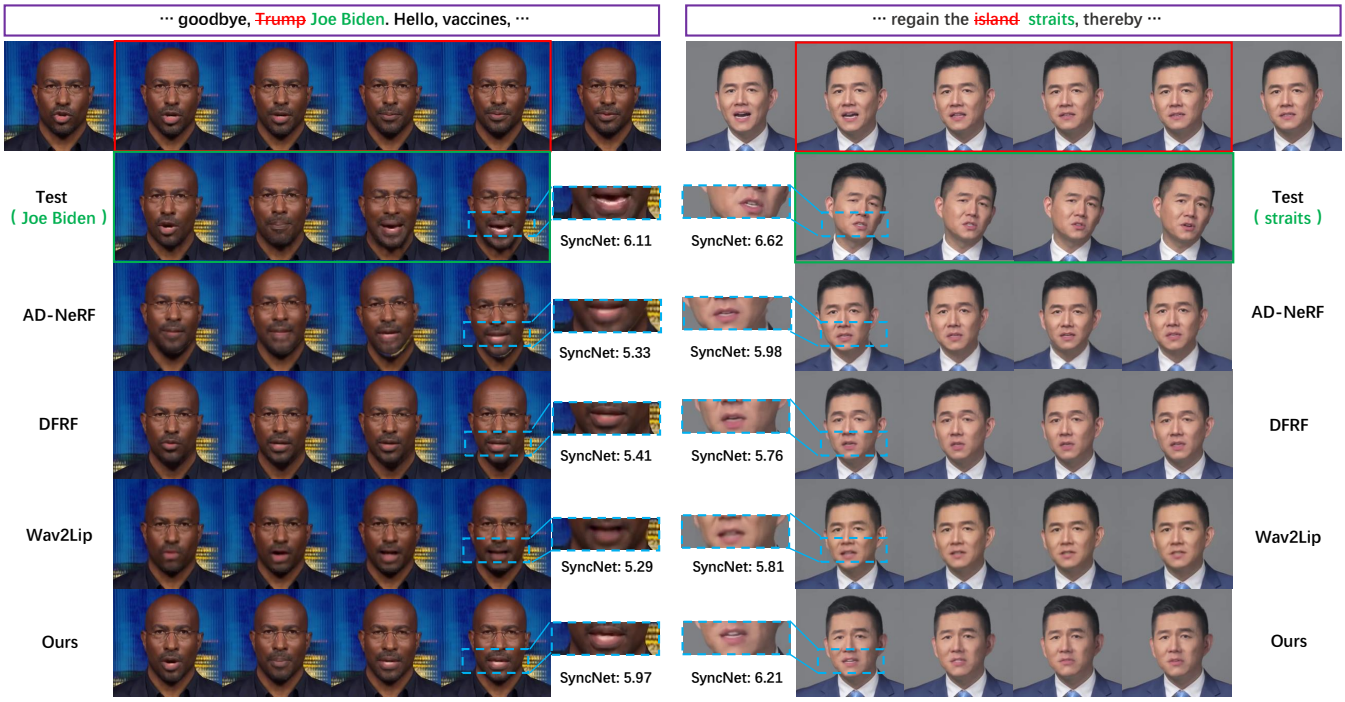

Figure 3: The results of substituting the word of the talking-head videos. We further evaluate the SyncNet [9] scores of the edited video clips.

	LMD [6]↓	SyncNet [9]↑	PSNR↑	SSIM↑	LPIPS↓	LF [18]↓	LDS↓	User Study	
								Mean↑	Real↑
Ground truth	0.0	6.24	N/A	1.0	0.0	0.12	2.23	4.160	81.0
AD-NeRF [17]	5.63	5.24	29.75	0.842	0.045	0.89	2.77	2.249	19.6
DFRF (5s/10k) [36]	7.16	5.33	28.21	0.814	0.092	0.91	2.84	2.271	14.4
DFRF (15s/40k) [36]	5.03	6.02	30.98	0.871	0.041	0.80	2.45	2.406	22.1
Wav2Lip [32]	5.97	**6.16**	26.12	0.791	0.087	0.85	2.66	2.333	20.1
Ours (5s/10k)	6.46	5.52	30.23	0.838	0.089	0.88	2.73	2.423	17.2
Ours (15s/40k)	**4.97**	6.11	**33.79**	**0.879**	**0.039**	**0.79**	**2.36**	**2.843**	**26.7**

Table 2: The quantitative evaluation of the accuracy, quality, and smoothness of the edited videos.

kept. The number of experts is set as 48, which is roughly equal to the number of distinct phonemes in Mandarin or English. We set the k as 16, following the defined basic visemes [49].

Dataset and Training. The animation prediction module is trained on the GRID [10] dataset (30×40 minutes) and takes 3 days. As for the neural rendering module, we train the basic NeRF model using the same 3-minute multi-identity video provided by the DFRF [36]. For an arbitrary unseen identity, we only fetch the 15-second video clip to be edited and fine-tune the basic NeRF model using this clip to obtain the corresponding personalized NeRF model. Note that the basic NeRF model is trained for 300k iterations and every personalized NeRF model is trained for 10k iterations. We set $\lambda_1 = 0.5$ and $\lambda_2 = 5$. We select thirteen 30-second videos of different identities (6 females, 7 males; 7 Caucasians, 3 Asians, 2 Africans, 1 Italian) for evaluation.

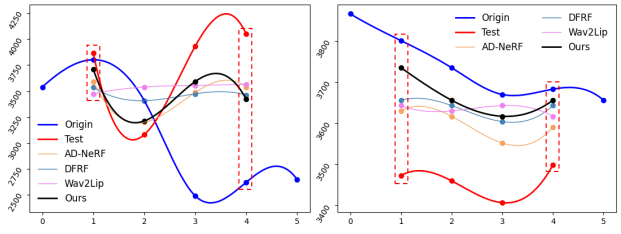

Figure 4: The mouth area of the frames shown in Fig. 3 (Left: the African male, Right: the Asian male). Our curve (black) simultaneously maintains the closeness to the test curve (red), and the smooth transition with the original video (please see the transition points framed by dotted lines, and the closer it gets, the smoother it becomes) .

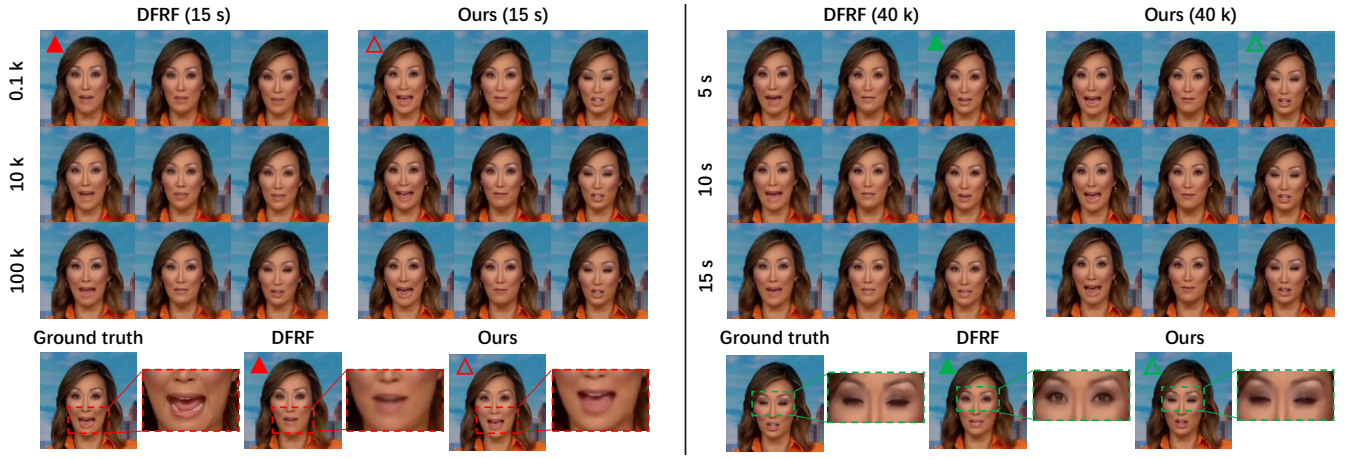

Figure 5: The qualitative results using DFRF [36] and ours with different amounts of training data and fine-tuning epochs.

		LMD [6]↓		SyncNet [9]↑		PSNR↑		SSIM↑		LPIPS↓	
		DFRF	Ours	DFRF	Ours	DFRF	Ours	DFRF	Ours	DFRF	Ours
5 s	0.1 k	8.34	**6.99**	4.88	**5.40**	27.33	**28.01**	0.725	**0.802**	0.168	**0.160**
	10 k	7.02	**6.40**	5.39	**5.65**	28.21	**30.23**	0.811	**0.837**	0.092	**0.089**
	40 k	6.44	**5.72**	5.64	**5.68**	28.92	**31.00**	0.829	**0.848**	0.071	**0.066**
10 s	0.1 k	8.01	**6.81**	4.97	**5.45**	27.89	**28.62**	0.734	**0.811**	0.152	**0.149**
	10 k	6.85	**6.08**	5.48	**5.67**	28.67	**30.68**	0.845	**0.849**	0.089	**0.070**
	40 k	5.42	**5.08**	6.11	**6.12**	30.43	**32.85**	0.852	**0.870**	0.045	**0.041**
15 s	0.1 k	7.87	**6.65**	5.11	**5.61**	28.02	**29.75**	0.750	**0.825**	0.141	**0.138**
	10 k	6.04	**5.39**	5.88	**6.06**	29.75	**31.11**	0.849	**0.891**	0.065	**0.059**
	40 k	5.03	**4.94**	6.19	**6.17**	30.98	**33.72**	0.881	**0.897**	0.041	**0.039**

Table 3: The quantitative evaluation of the DFRF [36] and ours under different amounts of training data and fine-tuning epochs.

Baseline Methods. We select Wav2Lip [32], AD-NeRF [17], and DFRF [36] to validate our rendering quality and superior smoothness among non-NeRF and NeRF methods. Wav2Lip has to use the original video as input to obtain the edited video which is conditioned on the novel speech audio, while AD-NeRF and DFRF estimate the head pose based on Face2Face [44] and manually select the pose for the edited sequence. Moreover, Our editing method has three similar NeRF-based works: AD-NeRF [17] needs 36 hours to train the model with a source video of 5 minutes; SSP-NeRF [25] has the same setting of AD-NeRF; DFRF [36] takes 15-second videos for training. So we select DFRF to compare the few-shot performance under different amounts of training data and epochs.

Metrics. We evaluate the results in the following three aspects: (1) **Accuracy:** The Landmark Distance (**LMD**) [6] and the lip synchronization confidence calculated by the SyncNet [9] are adopted to evaluate the accuracy of generated lip movements. (2) **Quality:** We adopt PSNR, SSIM, and LPIPS to evaluate the image quality of the generated frames. We also adopt LipForensics [18] (**LF**) to calculate the probability of the generated clips predicted as fake. (3) **Smoothness:** We propose a new metric called Landmark Distance Smoothness (**LDS**), to evaluate whether the connection between the inserted frames and the original frames is smooth, which calculates the mean squared distance of the 2D landmarks [5] between the starting edited frame and its previous frame, as well as the end edited frame and its next frame. Moreover, we conduct a user study of the 5-Point Likert scale proposed by previous works [14, 50], where the 'Mean' is the weighted average of the scores and the 'Real' is the total percentage of the scores of 4 and 5.

4.2 Quality, Accuracy, and Smoothness

The substitution operation includes novel content creation similar to insertion and also disconnected part smoothing similar to deletion, so we compare our method with different NeRF and non-NeRF methods in the most challenging substitution setting. As shown in Fig. 3, we show two different representative cases: the lip motion corresponding to test speech is much different from the original frames (Fig. 3 left); the head poses of the test video are much different from the original frames (Fig. 3 right). The former case shows that we can achieve a good trade-off between the smooth transition with the adjacent frames and lip accuracy with the driven speech audio. The latter case shows our disentangled control of verbal and non-verbal performance. As shown in Fig. 4, we further use face parsing [52] to calculate the number of pixels belonging to the mouth region of the frames in Fig. 3, where the left table is

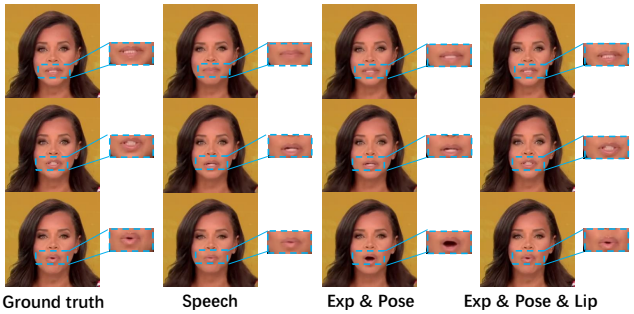

Figure 6: The ablation study of using different animation conditions under the same amount of training data and fine-tuning epochs.

Figure 7: The ablation study of animation and rendering context information for improving the motion smoothness and lighting consistency respectively. The generated frames are marked as green.

the Fig. 3 left, and the right is the Fig. 3 right. Our inserted frames achieve closer mouth areas with the original adjacent frames, which shows a smoother transition. For quantitative evaluation, we compare the AD-NeRF [17], DFRF [36] and Wav2Lip [32] on our test dataset, as shown in Tab. 2. We achieve better accuracy, quality, and smoothness compared with different NeRF and non-NeRF methods.

4.3 Efficiency-Quality Trade-Off

As shown in Fig. 5, we compare the rendering quality of the DFRF [36] and ours under different training data amounts and training epochs. Compared with DFRF, we adopt the head pose, expression, and lip feature embedding as an intermediate representation between vision and speech modality, to disentangle the verbal and non-verbal performance, instead of using the speech as a condition. This can reduce the data requirement of aligning the speech with non-verbal vision content such as head pose and expression. With fewer training data and iterations, our method can make the neural radiance field focus on the small but important parts (e.g., eyes and mouth) of talking-head videos. This can result in better control of lip synchronization and fine-grained animation. For quantitative comparison, we select thirteen different identities to evaluate, as shown in Tab. 3.

4.4 Ablation Study

We conduct the ablation study with different conditions under the same experimental setting. As shown in Fig. 6, our method can achieve better lip synchronization using the disentangled control of verbal and non-verbal information. Moreover, we evaluate the animation context (**AC**) by using only the driven speech feature embedding, and the rendering context (**RC**) by adopting the arbitrary frames as reference. As shown in the first and third rows of Fig. 7, our method maintains a smooth head pose transition of the generated frames and adjacent frames of the original video, while flickering motion inconsistency appears without animation context. As shown in the second and third rows of Fig. 7, we further evaluate the effect of reference frames, and not adopting the frames around the editing point may lead to lighting inconsistency when inserting the novel frames into the original video, where the mouth part is especially obvious. More quantitative results of different identities can be seen in Tab. 4.

	LMD [6]↓	SyncNet [9]↑	PSNR↑	LDS↓
Speech	5.11	6.02	30.25	2.45
Exp&Pose	5.08	6.05	30.33	2.53
w/o AC	5.02	5.94	32.98	2.82
w/o RC	5.23	6.08	29.87	2.38
Ours	**4.97**	**6.11**	**33.79**	**2.36**

Table 4: The quantitative evaluation of ablation study.

5 CONCLUSIONS

As for data requirement, efficiency, lip synchronization, and motion smoothness in the talking-head video editing setting, we propose a novel editing framework with an animation prediction module and a neural rendering module for the word-level editing of talking-head videos. We fully utilize the contextual information to realize the motion smoothness of animation sequence prediction and solve the ill-posed problem of obtaining a person-specific rendering model with only few-second source video. Extensive experiments demonstrate that our method achieves smoother editing with higher quality and less data than previous methods.

Limitations. The rendering requirement for editing excludes multi-view synthesis, because word-level editing only requires few-second video clip generation without large pose variation. We are not able to tackle the multi-view generation with only frontal-view faces as input data in this editing framework. Moreover, the lighting inconsistency in some hard cases (e.g., hair region) can be solved by adopting more computation resource. The performance of our method has exceeded the state-of-the-art methods, and more challenging requirements (e.g., large-pose editing, hair rendering, and sentence-level editing) will be studied in our future work.

6 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 61972395, 62272460, U19B2038, Beijing Natural Science Foundation under Grant No. 4232037, and a grant from Young Elite Scientists Sponsorship Program by CAST (YESS).

REFERENCES

[1] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. 2016. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*. PMLR, 173–182.

[2] Yannis M Assael, Brendan Shillingford, Shimon Whiteson, and Nando De Freitas. 2016. Lipnet: End-to-end sentence-level lipreading. *GPU Technology Conference* (2016).

[3] Floraine Berthouzoz, Wilmot Li, and Maneesh Agrawala. 2012. Tools for placing cuts and transitions in interview video. *ACM Transactions on Graphics (TOG)* 31, 4 (2012), 1–8.

[4] Volker Blanz and Thomas Vetter. 1999. A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*. 187–194.

[5] Adrian Bulat and Georgios Tzimiropoulos. 2017. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE International Conference on Computer Vision*. 1021–1030.

[6] Lele Chen, Zhiheng Li, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. 2018. Lip movements generation at a glance. In *European conference on computer vision*. 520–535.

[7] Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. 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.

[8] Kun Cheng, Xiaodong Cun, Yong Zhang, Menghan Xia, Fei Yin, Mingrui Zhu, Xuan Wang, Jue Wang, and Nannan Wang. 2022. VideoReTalking: Audio-based Lip Synchronization for Talking Head Video Editing In the Wild. In *SIGGRAPH Asia 2022 Conference Papers*. 1–9.

[9] Joon Son Chung and Andrew Zisserman. 2016. Out of time: automated lip sync in the wild. In *ACCV*. Springer, 251–263.

[10] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. 2006. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America* 120, 5 (2006), 2421–2424.

[11] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J Black. 2019. Capture, learning, and synthesis of 3D speaking styles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10101–10111.

[12] Dipanjan Das, Sandika Biswas, Sanjana Sinha, and Brojeshwar Bhowmick. 2020. Speech-driven facial animation using cascaded gans for learning of motion and texture. In *European conference on computer vision*. Springer, 408–424.

[13] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. 2021. Learning an animatable detailed 3D face model from in-the-wild images. *ACM Transactions on Graphics (ToG)* 40, 4 (2021), 1–13.

[14] Ohad Fried, Ayush Tewari, Michael Zollhöfer, Adam Finkelstein, Eli Shechtman, Dan B Goldman, Kyle Genova, Zeyu Jin, Christian Theobalt, and Maneesh Agrawala. 2019. Text-based editing of talking-head video. *ACM Transactions on Graphics (TOG)* 38, 4 (2019), 1–14.

[15] Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias Nießner. 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.

[16] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. 2022. Neural head avatars from monocular RGB videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18653–18664.

[17] Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. 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.

[18] Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. 2021. Lips don't lie: A generalisable and robust approach to face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5039–5049.

[19] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. 2014. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567* (2014).

[20] Amir Jamaludin, Joon Son Chung, and Andrew Zisserman. 2019. You said that?: Synthesising talking faces from audio. *International Journal of Computer Vision* 127, 11 (2019), 1767–1779.

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

[22] James T Kajiya and Brian P Von Herzen. 1984. Ray tracing volume densities. *ACM SIGGRAPH computer graphics* 18, 3 (1984), 165–174.

[23] Avisek Lahiri, Vivek Kwatra, Christian Frueh, John Lewis, and Chris Bregler. 2021. Lipsync3d: Data-efficient learning of personalized 3d talking faces from video using pose and lighting normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2755–2764.

[24] Jun Ling, Xu Tan, Liyang Chen, Runnan Li, Yuchao Zhang, Sheng Zhao, and Li Song. 2022. StableFace: Analyzing and Improving Motion Stability for Talking Face Generation. *arXiv preprint arXiv:2208.13717* (2022).

[25] Xian Liu, Yinghao Xu, Qianyi Wu, Hang Zhou, Wayne Wu, and Bolei Zhou. 2022. Semantic-aware implicit neural audio-driven video portrait generation. *European conference on computer vision* (2022).

[26] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. 2020. Object-centric learning with slot attention. *Advances in Neural Information Processing Systems* 33 (2020), 11525–11538.

[27] Yueming Lyu, Jing Dong, Bo Peng, Wei Wang, and Tieniu Tan. 2021. SOGAN: 3D-aware shadow and occlusion robust GAN for makeup transfer. In *Proceedings of the 29th ACM International conference on multimedia*. 3601–3609.

[28] Tianxiang Ma, Bingchuan Li, Qian He, Jing Dong, and Tieniu Tan. 2023. Semantic 3D-aware Portrait Synthesis and Manipulation Based on Compositional Neural Radiance Field. *arXiv preprint arXiv:2302.01579* (2023).

[29] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* 65, 1 (2021), 99–106.

[30] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. 2021. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5865–5874.

[31] Hai X Pham, Samuel Cheung, and Vladimir Pavlovic. 2017. Speech-driven 3D facial animation with implicit emotional awareness: a deep learning approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 80–88.

[32] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. 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.

[33] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. 2021. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10318–10327.

[34] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. Fastspeech: Fast, robust and controllable text to speech. *Advances in Neural Information Processing Systems* 32 (2019).

[35] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. 2017. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. *CoRR* abs/1701.06538 (2017). arXiv:1701.06538

[36] Shuai Shen, Wanhua Li, Zheng Zhu, Yueqi Duan, Jie Zhou, and Jiwen Lu. 2022. Learning Dynamic Facial Radiance Fields for Few-Shot Talking Head Synthesis. In *European conference on computer vision*.

[37] Luchuan Song, Bin Liu, Guojun Yin, Xiaoyi Dong, Yufei Zhang, and Jia-Xuan Bai. 2021. TACR-Net: Editing on Deep Video and Voice Portraits. In *Proceedings of the 29th ACM International Conference on Multimedia*. 478–486.

[38] Yang Song, Jingwen Zhu, Dawei Li, Xiaolong Wang, and Hairong Qi. 2019. Talking face generation by conditional recurrent adversarial network. *International Joint Conference on Artificial Intelligence* (2019).

[39] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. 2017. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (ToG)* 36, 4 (2017), 1–13.

[40] Xu Tan, Jiawei Chen, Haohe Liu, Jian Cong, Chen Zhang, Yanqing Liu, Xi Wang, Yichong Leng, Yuanhao Yi, Lei He, et al. 2022. NaturalSpeech: End-to-End Text to Speech Synthesis with Human-Level Quality. *arXiv preprint arXiv:2205.04421* (2022).

[41] Chuanxin Tang, Chong Luo, Zhiyuan Zhao, Dacheng Yin, Yucheng Zhao, and Wenjun Zeng. 2021. Zero-Shot Text-to-Speech for Text-Based Insertion in Audio Narration. *Interspeech* (2021).

[42] Jiaxiang Tang, Kaisiyuan Wang, Hang Zhou, Xiaokang Chen, Dongliang He, Tianshu Hu, Jingtuo Liu, Gang Zeng, and Jingdong Wang. 2022. Real-time neural radiance talking portrait synthesis via audio-spatial decomposition. *arXiv preprint arXiv:2211.12368* (2022).

[43] Sarah Taylor, Taehwan Kim, Yisong Yue, Moshe Mahler, James Krahe, Anastasio Garcia Rodriguez, Jessica Hodgins, and Iain Matthews. 2017. A deep learning approach for generalized speech animation. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 1–11.

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

[45] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. 2021. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 12959–12970.

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[47] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. 2020. Realistic speech-driven facial animation with gans. *International Journal of Computer Vision* 128, 5 (2020), 1398–1413.

[48] Suzhen Wang, Lincheng Li, Yu Ding, Changjie Fan, and Xin Yu. 2021. Audio2head: Audio-driven one-shot talking-head generation with natural head motion. *IJCAI* (2021).

[49] Zhiyong Wu, Shen Zhang, Lianhong Cai, and Helen M Meng. 2006. Real-time synthesis of Chinese visual speech and facial expressions using MPEG-4 FAP features in a three-dimensional avatar. *Interspeech* 4 (2006), 1802–1805.

[50] Xinwei Yao, Ohad Fried, Kayvon Fatahalian, and Maneesh Agrawala. 2021. Iterative text-based editing of talking-heads using neural retargeting. *ACM Transactions on Graphics (TOG)* 40, 3 (2021), 1–14.

[51] Dacheng Yin, Chuanxin Tang, Yanqing Liu, Xiaoqiang Wang, Zhiyuan Zhao, Yucheng Zhao, Zhiwei Xiong, Sheng Zhao, and Chong Luo. 2022. RetrieverTTS: Modeling Decomposed Factors for Text-Based Speech Insertion. *Interspeech* (2022).

[52] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. 2018. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *European conference on computer vision*. 325–341.

[53] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. 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.

[54] Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C Bühler, Xu Chen, Michael J Black, and Otmar Hilliges. 2022. Im avatar: Implicit morphable head avatars from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13545–13555.

[55] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. 2019. Talking face generation by adversarially disentangled audio-visual representation. In *AAAI*, Vol. 33. 9299–9306.

[56] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. 2021. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4176–4186.

[57] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. 2020. Makelttalk: speaker-aware talking-head animation. *ACM Transactions on Graphics (TOG)* 39, 6 (2020), 1–15.

[58] Hao Zhu, Huaibo Huang, Yi Li, Aihua Zheng, and Ran He. 2020. Arbitrary talking face generation via attentional audio-visual coherence learning. *IJCAI* (2020).