SyncTalk: The Devil is in the Synchronization for Talking Head Synthesis

Ziqiao Peng[1]   Wentao Hu[2]   Yue Shi[3]   Xiangyu Zhu[4]   Xiaomei Zhang[4]   Hao Zhao[5]

Jun He[1]   Hongyan Liu[5*]   Zhaoxin Fan[1*]

[1]Renmin University of China   [2]Beijing University of Posts and Telecommunications

[3]Psyche AI Inc.   [4]Chinese Academy of Sciences   [5]Tsinghua University
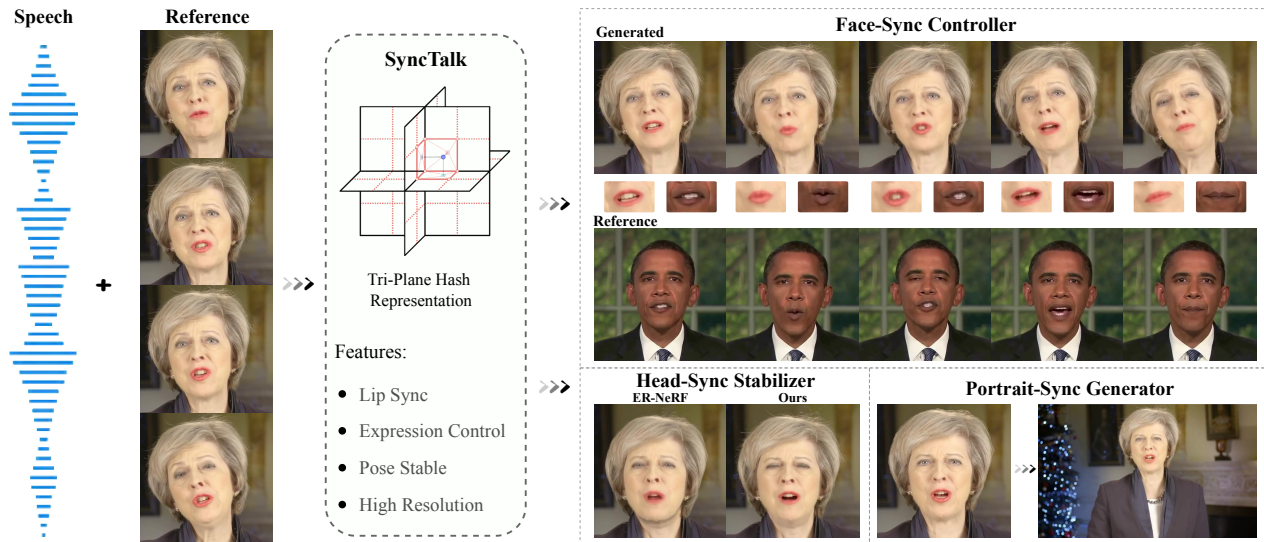
Figure 1. The proposed SyncTalk synthesizes synchronized talking head videos, employing tri-plane hash representations to maintain subject identity. It can generate synchronized lip movements, facial expressions, and stable head poses, and restores hair details to create high-resolution videos.

Abstract

Achieving high synchronization in the synthesis of realistic, speech-driven talking head videos presents a significant challenge. Traditional Generative Adversarial Networks (GAN) struggle to maintain consistent facial identity, while Neural Radiance Fields (NeRF) methods, although they can address this issue, often produce mismatched lip movements, inadequate facial expressions, and unstable head poses. A lifelike talking head requires synchronized coordination of subject identity, lip movements, facial expressions, and head poses. The absence of these synchronizations is a fundamental flaw, leading to unrealistic and artificial outcomes. To address the critical issue of synchronization, identified as the "devil" in creating realistic talking heads, we introduce SyncTalk. This NeRF-based method effectively maintains subject identity, enhancing synchronization and realism in talking head synthesis. SyncTalk employs a Face-Sync Controller to align lip movements with speech and innovatively uses a 3D facial blendshape model to capture accurate facial expressions. Our Head-Sync Stabilizer optimizes head poses, achieving more natural head movements. The Portrait-Sync Generator restores hair details and blends the generated head with the torso for a seamless visual experience. Extensive experiments and user studies demonstrate that SyncTalk outperforms state-of-the-art methods in synchronization and realism. We recommend watching the supplementary video: https://ziqiaopeng.github.io/synctalk

1. Introduction

The quest to generate dynamic and realistic speech-driven talking heads has intensified, driven by expanding applications in digital assistants [39], virtual reality [32], and film-making [18]. The ultimate goal is to enhance the realism of synthetic videos to align with human perceptual expectations. However, a fundamental challenge that persists across existing methods is the need for synchroniza-

*Corresponding authors.

tion. Traditional methods based on Generative Adversarial Networks (GAN) [13, 41, 50, 53], while adept at modeling the lip movements of speakers, often produce inconsistent identities across different frames, leading to issues such as different tooth sizes and fluctuating lip thickness. Similarly, emerging technologies based on Neural Radiance Fields (NeRF) [14, 21, 35, 43, 44] excel in maintaining identity consistency and preserving facial details. However, they struggle with mismatches in lip movements, challenging facial expression control, and unstable head poses, thereby diminishing the overall realism of the video.

In this paper, we find that the "devil" is in the synchronization. Existing methods need more synchronization in four key areas: subject identity, lip movements, facial expressions, and head poses. Firstly, in GAN-based methods, maintaining the subject's identity in the video is challenging due to the instability of features in consecutive frames and the use of only a few frames as references for facial reconstruction [34]. Secondly, lip movements are not synchronized with speech. In NeRF-based methods, audio features trained only on a 5-minute speech dataset struggle to generalize to different speech inputs [44]. Thirdly, there is a lack of facial expression control, with most methods only producing lip movements or controlling blinking, resulting in unnatural facial actions [12]. Fourthly, the head pose is not synchronized. Previous methods relied on sparse landmarks to compute projection error, but jitter and inaccuracy in these landmarks lead to unstable head poses [49], as shown in Fig. 1. These synchronization issues introduce artifacts and significantly reduce realism.

To address these synchronization challenges, we introduce SyncTalk, a NeRF-based method focused on highly synchronized, realistic, speech-driven talking head synthesis, employing tri-plane hash representations to maintain subject identity. Through the Face-Sync Controller and Head-Sync Stabilizer, SyncTalk significantly enhances the synchronization and visual quality of the synthesized videos. Visual quality is further improved by the Portrait-Sync Generator, which meticulously refines visual details. The entire rendering process can achieve 50 FPS and outputs high-resolution videos.

Within the Face-Sync Controller, we pre-train an audio-visual encoder on the 2D audio-visual dataset, resulting in a generalized representation that ensures synchronized lip movements across different speech samples. For controlling facial expressions, we employ a semantically enriched 3D facial blendshape model [33]. This model is distinguished by its ability to control specific facial expression regions through 52 parameters. Regarding the Head-Sync Stabilizer, we use a head motion tracker [14] to infer the head's rough rotation and translation parameters. Due to the instability of the rough parameters, inspired by Simultaneous Localization and Mapping (SLAM), we incorporate a head

point tracker to track dense keypoints and integrate a bundle adjustment method to optimize the head pose, thereby achieving stable and continuous head motion. To further enhance the visual fidelity of SyncTalk, we have designed a Portrait-Sync Generator. This module repairs artifacts in the NeRF modeling, particularly the intricate details of hair and background flaws, and outputs high-resolution video.

In summary, the main contributions of our work are as follows:

- We present a Face-Sync Controller that utilizes an Audio-Visual Encoder in conjunction with a Facial Animation Capturer, ensuring accurate lip synchronization and dynamic facial expression rendering.
- We introduce a Head-Sync Stabilizer that tracks head rotation and facial movement keypoints. Utilizing the bundle adjustment method, this Stabilizer guarantees smooth and synchronous head motion.
- We design a Portrait-Sync Generator that improves visual fidelity by repairing artifacts in NeRF modeling and refining intricate details like hair and background in high-resolution videos.

2. Related Work

2.1. GAN-based Talking Head Synthesis

Recently, GAN-based talking head synthesis [5, 6, 9, 20, 26, 36, 40, 54, 55] has emerged as an essential research area in computer vision, but they struggle to maintain the identity of the subject in videos consistently. A cluster of noteworthy techniques, including [13, 34, 37, 41, 50, 53], mainly focus on generating video streams for the lip region. These methods create new visuals for talking head portraits by changing the lip region. For instance, Wav2Lip [34] introduces a lip sync expert for supervising lip movements. However, due to the use of five frames from the reference frame to reconstruct the lip, it struggles to maintain the subject's identity. In contrast, methods like [6, 24, 42, 56] perform full-face synthesis but struggle to ensure sync between facial expressions and head poses.

Apart from video stream techniques, efforts have also been made to enable a single image to "speak" using speech, as demonstrated in [17, 45, 48]. For example, SadTalker [48] can generate videos of a person speaking from a single image. However, such methods fail to generate natural head poses and facial expressions and struggle to maintain the subject's identity, affecting the sync effect and leading to an unrealistic visual perception.

Compared to these methods, SyncTalk uses NeRF to perform three-dimensional modeling of the face. Its capability to represent continuous 3D scenes in canonical spaces translates to exceptional performance in maintaining subject identity consistency and detail preservation.
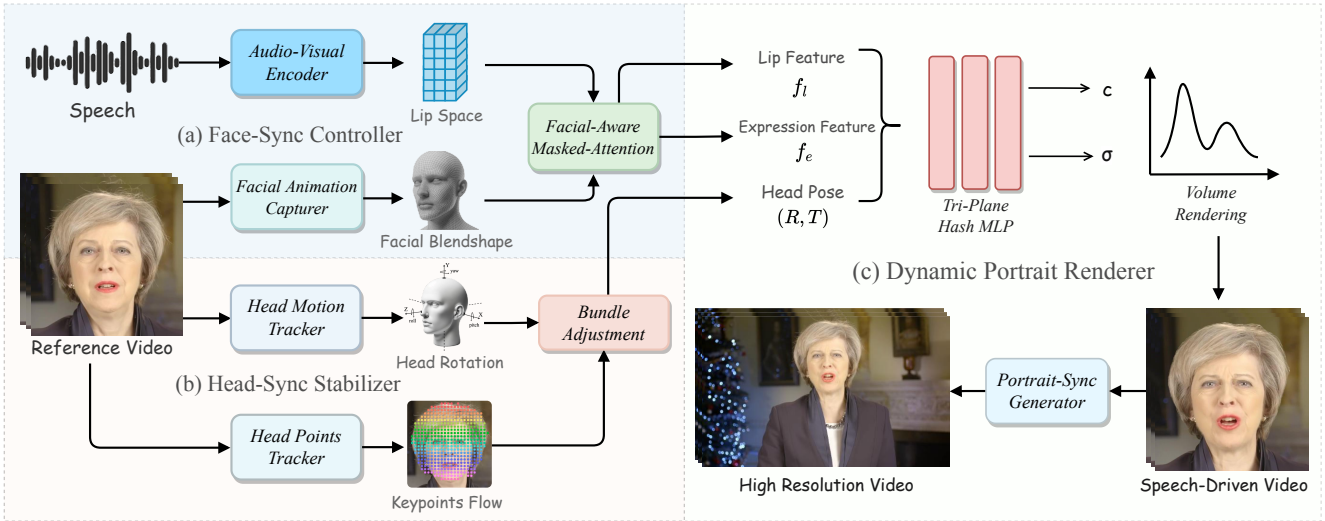
Figure 2. **Overview of SyncTalk.** Given a cropped reference video of a talking head and the corresponding speech, SyncTalk can extract the Lip Feature f_l, Expression Feature f_e, and Head Pose (R, T) through two synchronization modules (a) and (b). The Tri-Plane Hash Representation then models the head, outputting a rough speech-driven video. The Portrait-Sync Generator further restores details such as hair and background, ultimately producing a high-resolution talking head video.

2.2. NeRF-based Talking Head Synthesis

With the recent rise of NeRF, numerous fields have begun to utilize it to tackle related challenges [11, 25]. Previous work [14, 22, 43, 44] has integrated NeRF into the task of synthesizing talking heads and has used audio as the driving signal, but these methods are all based on the vanilla NeRF model. For instance, AD-NeRF [14] requires approximately 10 seconds to render a single image. RAD-NeRF [38] aims for real-time video generation and employs a NeRF based on Instant-NGP [30]. ER-NeRF [21] innovatively introduces triple-plane hash encoders to trim the empty spatial regions, advocating for a compact and accelerated rendering approach. GeneFace [44] attempts to reduce NeRF artifacts by translating speech features into facial landmarks, but this often results in inaccurate lip movements. Attempts to create character avatars with NeRF-based methods, such as [10, 51, 52, 57], cannot be directly driven by speech. These methods only use audio as a condition, without a clear concept of sync, and usually result in average lip movement. Additionally, previous methods lack control over facial expressions, being limited to controlling blinking only, and cannot model actions like raising eyebrows or frowning. Furthermore, these methods have a significant issue with unstable head poses, leading to the separation between the head and the torso.

In contrast, we use the Face-Sync Controller to model the relationship between audio and lip movements, thereby enhancing the synchronization of lip movements and expressions, and the Head-Sync Stabilizer to stabilize head poses. By addressing these synchronization "devils", our method improves visual quality.

3. Method

3.1. Overview

In this section, we introduce the proposed SyncTalk, as shown in Fig. 2. SyncTalk mainly consists of 3 parts: a) lip movements and facial expressions controlled by the Face-Sync Controller, b) stable head pose provided by the Head-Sync Stabilizer, and c) high-synchronization facial frames rendered by the Dynamic Portrait Renderer. We will describe the content of these three parts in detail in the following subsections.

3.2. Face-Sync Controller

Audio-Visual Encoder. Existing methods based on NeRF utilize pre-trained models such as DeepSpeech [2], Wav2Vec 2.0 [3], or HuBERT [16]. These are audio feature extraction methods designed for speech recognition tasks. Using an audio encoder designed for Automatic Speech Recognition (ASR) tasks does not truly reflect lip movements. This is because the pre-trained model is based on the distribution of features from audio to text, whereas we need the feature distribution from audio to lip movements.

Considering the above, we use an audio and visual synchronization audio encoder trained on the 2D audio-visual synchronization dataset LRS2 [1]. This ensures that the audio features extracted by our method and lip movements have the same feature distribution. The specific implementation method is as follows: We use a pre-trained lip synchronization discriminator [8]. It can give confidence for the lip synchronization effect of the video. The lip synchronization discriminator takes as input a continuous face window F and the corresponding audio frame A. They are

judged as positive samples (with label $y = 1$) if they overlap entirely. Otherwise, they are judged as negative samples (with label $y = 0$). The discriminator calculates the cosine similarity between these sequences as:

$$\text{sim}(F, A) = \frac{F \cdot A}{\|F\|_2 \|A\|_2},\tag{1}$$

and then uses binary cross-entropy loss:

$$L_{\text{sync}} = -\left(y \log(\text{sim}(F, A)) + (1 - y) \log(1 - \text{sim}(F, A))\right),\tag{2}$$

to minimize the distance for synchronized samples and maximize the distance for non-synchronized samples.

Under the supervision of the lip synchronization discriminator, we pre-train a highly synchronized audio-visual feature extractor related to lip movements. First, we use convolutional networks to obtain audio features $\text{Conv}(A)$ and encode facial features $\text{Conv}(F)$. These features are then concatenated. In the decoding phase, we use stacked convolutional layers to restore facial frames using the operation $\text{Dec}(\text{Conv}(A) \oplus \text{Conv}(F))$. The L_1 reconstruction loss during training is given by:

$$L_{\text{recon}} = \|F - \text{Dec}(\text{Conv}(A) \oplus \text{Conv}(F))\|_1.\tag{3}$$

Simultaneously, we sample synchronized and non-synchronized segments using lip movement discriminators, and employ a same sync loss as Eq. 2. By minimizing both losses, we train a facial generation network related to audio. After training, we use $\text{Conv}(A)$ as the lip space extracted from the audio. Ultimately, we obtain our highly synchronized audio-visual encoder related to lip movements.

Facial Animation Capturer. It is observed that previous methods based on NeRF [14, 21, 44] could only change blinking and could not model facial expressions accurately. This leads to issues such as rigid facial expressions and incorrect facial details if the trained character has significant facial movements, like squinting, raising eyebrows, or frowning. Considering the need for more synchronized and realistic facial expressions, we add an expression synchronization control module. Specifically, we introduce a 3D facial prior using 52 semantically facial blendshape coefficients [33] represented by B to model the face, as shown in Fig. 3. Because the 3D face model can retain the structure information of face motion, it can reflect the content of facial movements well without causing facial structural distortion. During the training, we first use a sophisticated facial blendshape capture module to capture facial expressions as $E(B)$, and select seven core facial expression control coefficients to control the eyebrow, forehead, and eye areas. They are highly correlated with expression and independent of lip movements. We can synchronize the speaker's expression during the inference process because the facial coefficients are semantically informed.
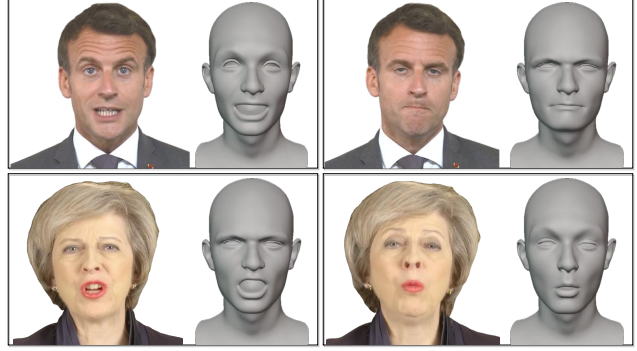

Figure 3. **Facial Animation Capturer.** We use 3D facial blendshape coefficients to control the expressions of characters.

Facial-Aware Masked-Attention. To reduce the mutual interference between lip features and expression features during training, we introduce the Facial-Aware Disentangle Attention module. Building on the region attention vector V [21], we add masks M_{lip} and M_{exp} to the attention areas for lips and expressions, respectively. Specifically, the new attention mechanisms are given by:

$$\begin{aligned} V_{\text{lip}} &= V \odot M_{\text{lip}}, \\ V_{\text{exp}} &= V \odot M_{\text{exp}}. \end{aligned}\tag{4}$$

These formulations allow the attention mechanisms to focus solely on their respective parts, thereby reducing entanglement between them. Before the disentanglement, lip movements might induce blinking tendencies and affect hair volume. By introducing the mask module, the attention mechanism can focus on either expressions or lips without affecting other areas, thereby reducing the artifact caused by coupling. Finally, we obtain the disentangled lip feature $f_l = f_{\text{lip}} \odot V_{\text{lip}}$ and expression feature $f_e = f_{\text{exp}} \odot V_{\text{exp}}$.

3.3. Head-Sync Stabilizer

Head Motion Tracker. The head pose, denoted as p, refers to the rotation angle of a person's head in 3D space and is defined by a rotation R and a translation T. An unstable head pose can lead to head jitter. To obtain a rough estimate of the head pose, initially, the best focal length is determined through i iterations within a predetermined range. For each focal length candidate, f_i, the system re-initializes the rotation and translation values. The objective is to minimize the error between the projected landmarks from the 3D Morphable Models (3DMM) [31] and the actual landmarks in the video frame. Formally, the optimal focal length f_{opt} is given by:

$$f_{\text{opt}} = \arg \min_{f_i} E_i(L_{2D}, L_{3D}(f_i, R_i, T_i)),\tag{5}$$

where E_i represents the Mean Squared Error (MSE) between these landmarks, $L_{3D}(f_i, R_i, T_i)$ represents the projected landmarks from the 3DMM for a given focal length

f_i, the corresponding rotation and translation parameters R_i and T_i, L_{2D} are the actual landmarks from the video frame. Subsequently, leveraging the optimal focal length f_{opt}, the system refines the rotation R and translation T parameters for all frames to better align the model's projected landmarks with the actual video landmarks. This refinement process can be mathematically represented as:

$$(R_{\text{opt}}, T_{\text{opt}}) = \arg\min_{R,T} E(L_{2D}, L_{3D}(f_{\text{opt}}, R, T)), \quad (6)$$

where E denotes the MSE metric, between the 3D model's projected landmarks L_{3D} for the optimal focal length f_{opt}, and the actual 2D landmarks L_{2D} in the video frame. The optimized rotation R_{opt} and translation T_{opt} are obtained by minimizing this error across all frames.

Head Points Tracker. Considering methods based on NeRF and their requirements for inputting head rotation R and translation T, previous methods utilize 3DMM-based techniques to extract head poses and generate an inaccurate result. To improve the precision of R and T, We use an optical flow estimation model from [43] to track facial keypoints K. Specifically, using a pre-trained optical flow estimation model, after obtaining the facial motion optical flow, we use the Laplacian filter to select the keypoints where the most significant flow changes are located and track the motion trajectories of these keypoints in the flow sequence. Through this module, our method ensures a more precise and consistent facial keypoint alignment across all frames, enhancing the accuracy of head pose parameters.

Bundle Adjustment. Given the keypoints and the rough head pose, we introduce a two-stage optimization framework from [14] to enhance the accuracy of keypoints and head pose estimations. In the first stage, we randomly initialize the 3D coordinates of j keypoints and optimize their positions to align with the tracked keypoints on the image plane. This process involve minimizing a loss function L_{init}, which captures the discrepancy between projected keypoints P and the tracked keypoints K, as given by:

$$L_{\text{init}} = \sum_j \|P_j - K_j\|_2. \quad (7)$$

Subsequently, in the second stage, we embark on a more comprehensive optimization to refine the 3D keypoints and the associated head jointly pose parameters. Through the Adam Optimization [19], the algorithm adjust the spatial coordinates, rotation angles R, and translations T to minimize the alignment error L_{sec}, expressed as:

$$L_{\text{sec}} = \sum_j \|P_j(R, T) - K_j\|_2. \quad (8)$$

After these optimizations, the resultant head pose and translation parameters are observed to be smooth and stable.

3.4. Dynamic Portrait Renderer

Tri-Plane Hash Representation. Using a collection of multi-perspective images along with corresponding camera poses, NeRF [27] harnesses these resources to manifest a 3D static scene. This representation utilizes an implicit function $\mathcal{F}$, delineated as $\mathcal{F} : (\text{x}, \text{d}) \to (\text{c}, \sigma)$, where the 3D spatial location is given by $\text{x} = (x, y, z)$, and the direction of viewing is characterized by $\text{d} = (\theta, \phi)$. The resultant values, $\text{c} = (r, g, b)$ and σ, represent the radiance and density, respectively. The predicted pixel color, denoted by $\hat{C}(\text{r})$, intertwined with ray $\text{r}(t) = \text{o} + t\text{d}$ originating from the camera's core position o, is derivable using:

$$\hat{C}(\text{r}) = \int_{t_n}^{t_f} \sigma(\text{r}(t)) \cdot \text{c}(\text{r}(t), \text{d}) \cdot T(t)dt, \quad (9)$$

where t_n and t_f are the near and far bounds, and $T(t)$ is the accumulated transmittance. Addressing the challenges of hash collisions and optimizing audio feature processing, we incorporate three uniquely oriented 2D hash grids [21]. A coordinate, given by $\text{x} = (x, y, z) \in \mathbb{R}^{\text{XYZ}}$, undergoes an encoding transformation for its projected values via three individual 2D-multiresolution hash encoders [30]:

$$\mathcal{H}^{\text{AB}} : (a, b) \to \text{f}_{ab}^{\text{AB}}, \quad (10)$$

where the output $\text{f}_{ab}^{\text{AB}} \in \mathbb{R}^{LD}$, with the number of levels L and feature dimensions per entry D, signifies the planar geometric feature corresponding to the projected coordinate (a, b) and $\mathcal{H}^{\text{AB}}$ denotes the multiresolution hash encoder for plane $\mathbb{R}^{\text{AB}}$. By fusing the outcomes, the conclusive geometric feature $\text{f}g \in \mathbb{R}^{3 \times LD}$ is derived as:

$$\text{f}_{\text{x}} = \mathcal{H}^{\text{XY}}(x, y) \oplus \mathcal{H}^{\text{YZ}}(y, z) \oplus \mathcal{H}^{\text{XZ}}(x, z), \quad (11)$$

where the concatenation of features is symbolized by $\oplus$, resulting in a $3 \times LD$-channel vector. Employing f_{x}, the perspective direction d, the lip feature f_l, and the expression feature f_e, the tri-plane hash's implicit function is defined as:

$$\mathcal{F}^{\mathcal{H}} : (\text{x}, \text{d}, f_l, f_e; \mathcal{H}^3) \to (\text{c}, \sigma), \quad (12)$$

where $\mathcal{H}^3$ amalgamates the triad of planar hash encoders, as illustrated in Eq. 10.

Our training employs a two-step coarse-to-fine strategy, initially using MSE loss to assess the difference between predicted $\hat{C}(\text{r})$ and actual image colors $C(\text{r})$. Recognizing MSE's limitations in detail capture, we advance to a refinement stage, incorporating LPIPS loss for enhanced detail, similar to ER-NeRF [21]. We extract random patches $\mathcal{P}$ from the image, combining LPIPS (weighted by λ) with MSE for improved detail representation, as shown:

$$\mathcal{L}_{\text{total}} = \sum_{\text{r}} \|C(\text{r}) - \hat{C}(\text{r})\|_2 + \lambda \times \mathcal{L}_{\text{LPIPS}}(\hat{\mathcal{P}}, \mathcal{P}). \quad (13)$$

Methods	PSNR ↑	LPIPS ↓	MS-SSIM ↑	FID ↓	NIQE ↓	BRISQUE ↓	LMD ↓	AUE ↓	LSE-C ↑
GAN Wav2Lip (ACM MM 20 [34])	33.4385	0.0697	0.9781	16.0228	14.5367	44.2659	4.9630	2.9029	**9.2387**
VideoReTalking (SIGGRAPH Asia 22 [7])	31.7923	0.0488	0.9680	9.2063	14.2410	43.0465	5.8575	3.3308	7.9683
DINet (AAAI 23 [50])	31.6475	0.0443	0.9640	9.4300	14.6850	40.3650	4.3725	3.6875	6.5653
TalkLip (CVPR 23 [41])	32.5154	0.0782	0.9697	18.4997	14.6385	46.6717	5.8605	2.9579	5.9472
IP-LAP (CVPR 23 [53])	35.1525	0.0443	<u>0.9803</u>	8.2125	14.6400	42.0750	3.3350	<u>2.8400</u>	4.9541
NeRF AD-NeRF (ICCV 21 [14])	26.7291	0.1536	0.9111	28.9862	14.9091	55.4667	2.9995	5.5481	4.4996
RAD-NeRF (arXiv 22 [38])	31.7754	0.0778	0.9452	8.6570	<u>13.4433</u>	44.6892	2.9115	5.0958	5.5219
GeneFace (ICLR 23 [44])	24.8165	0.1178	0.8753	21.7084	13.3353	46.5061	4.2859	5.4527	5.1950
ER-NeRF (ICCV 23 [21])	32.5216	0.0334	0.9501	5.2936	13.7048	<u>34.7361</u>	2.8137	4.1873	5.7749
SyncTalk (w/o Portrait)	<u>35.3542</u>	<u>0.0235</u>	0.9769	<u>3.9247</u>	**13.1333**	**33.2954**	<u>2.5714</u>	**2.5796**	<u>8.1331</u>
SyncTalk (Portrait)	**37.4017**	**0.0113**	**0.9841**	**2.7070**	14.2165	37.3042	**2.5043**	3.2074	8.0263

Table 1. **The quantitative results of the head reconstruction.** "Portrait" refers to the use of the Portrait-Sync Generator. We achieve state-of-the-art performance on most metrics. We highlight **best** and <u>second-best</u> results.

Portrait-Sync Generator. During the training process, to address NeRF's limitations in capturing fine details like hair strands and dynamic backgrounds, we introduce a Portrait-Sync Generator with two key sections. First, NeRF renders the face area (F_r), creates $G(F_r)$ through Gaussian blur, and then uses our synchronized head pose to be able to merge with the original image (F_o) to enhance hair detail fidelity. Second, when the head and torso are combined, if the character in the source video speaks while the generated face is silent, a dark gap area might appear, as shown in Fig. 5 (b). We fill these areas with the average neck color (C_n). This approach results in more realistic details and improved visual quality through the Portrait-Sync Generator.

4. Experiments

4.1. Experimental Settings

Dataset. To ensure a fair comparison, we use the same well-edited video sequences from [14, 21, 44], including English and French. The average length of these videos is approximately 8,843 frames, and each video is recorded at 25 FPS. Except for the video from AD-NeRF [14], which has a resolution of 450×450, all other videos have a resolution of 512×512, with the character-centered.

Comparison Baselines. We compare our method with five GAN-based methods, including Wav2Lip [34], Video-ReTalking [7], DINet [50], TalkLip [41], and IP-LAP [53], and NeRF-based methods such as AD-NeRF [14], RAD-NeRF [38], GeneFace [44], and ER-NeRF [21].

Implementation Details. In the coarse stage, the portrait head is trained for 100,000 iterations and 25,000 in the fine stage, sampling 256^2 rays per iteration using a 2D hash encoder (L=14, F=1). We employ the AdamW optimizer [23], with learning rates of 0.01 for the hash encoder and 0.001 for other modules. Total training time is approximately 2 hours on an NVIDIA RTX 3090 GPU.

Methods	Audio A		Audio B	
	LSE-D ↓	LSE-C ↑	LSE-D ↓	LSE-C ↑
DINet (AAAI 23 [50])	<u>8.5031</u>	5.6956	<u>8.2038</u>	5.1134
TalkLip (CVPR 23 [41])	8.7615	<u>5.7449</u>	8.7019	<u>5.5359</u>
IP-LAP (CVPR 23 [53])	9.8037	3.8578	9.1102	4.389
GeneFace (ICLR 23 [44])	9.5451	4.2933	9.6675	3.7342
ER-NeRF (ICCV 23 [21])	11.813	2.4076	10.7338	3.0242
SyncTalk (Ours)	**7.7211**	**6.6659**	**8.0248**	**6.2596**

Table 2. **The quantitative results of the lip synchronization.** We use two different audio samples to drive the same subject, then highlight **best** and <u>second-best</u> results.

4.2. Quantitative Evaluation

Full Reference Quality Assessment. In terms of image quality, we use full reference metrics such as Peak Signal-to-Noise Ratio (PSNR), Learned Perceptual Image Patch Similarity (LPIPS) [46], Multi-Scale Structure Similarity (MS-SSIM), and Frechet Inception Distance (FID) [15] as evaluation metrics.

No Reference Quality Assessment. In high PSNR images, texture details may not align with human visual perception [47]. For more precise output definition and comparison, we use two No Reference methods: the Natural Image Quality Evaluator (NIQE)[29] and the Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE)[28].

Synchronization Assessment. For synchronization, we use landmark distance (LMD) to measure the synchronicity of facial movements, action units error (AUE) [4] to assess the accuracy of facial movements, and introduce Lip Sync Error Confidence (LSE-C), consistent with Wav2Lip [34], to evaluate the synchronization between lip movements and audio.

Evaluation Results. The evaluation results of the head reconstruction are shown in Tab. 1. We compare recent methods based on GAN and NeRF. It can be observed that our image quality is superior to other methods in all aspects. In terms of synchronization, our results surpass most methods.
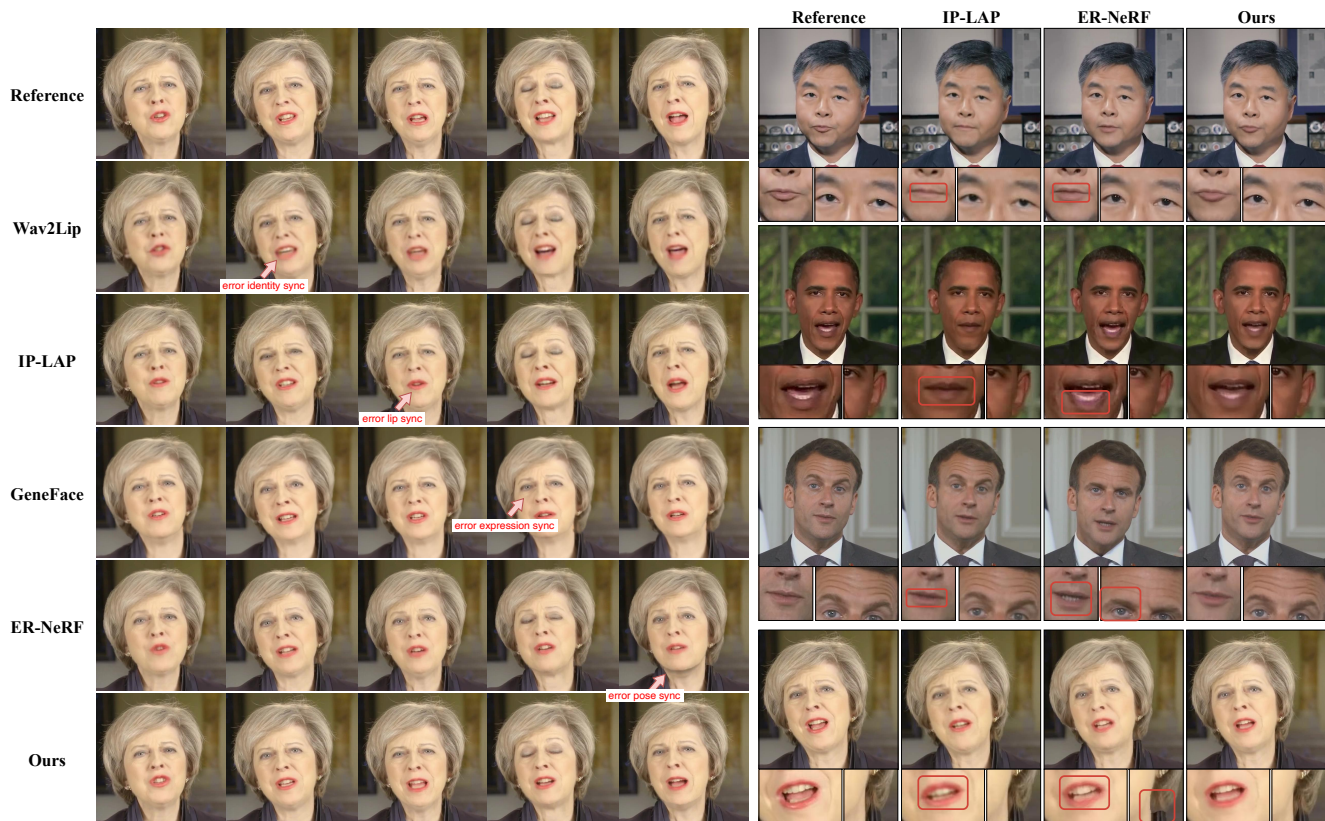
Figure 4. **Qualitative comparison of facial synthesis by different methods.** Our method has the best visual effect on lip movements and facial expressions without the problem of separation of head and torso. Please zoom in for better visualization.

We compare the two output modes of SyncTalk, one processed through the Portrait-Sync Generator and one without it. After processing through the Portrait-Sync Generator, hair details are restored, and the image quality is improved. Since we can maintain the subject's identity well, we surpass GAN-based methods in image quality. Thanks to the sync of lip, expression, and pose, we also outperform NeRF-based methods in image quality. Especially in terms of the LPIPS metric, our method has three times improvement compared to the previous state-of-the-art method ER-NeRF [21]. We compare the latest SOTA method drivers using out-of-distribution (OOD) audio, and the results are shown in Tab. 2. We introduce Lip Synchronization Error Distance (LSE-D) and Confidence (LSE-C) for lip-audio sync evaluation, aligning with [34]. Our method shows state-of-the-art lip synchronization, overcoming small-sample NeRF limitations by incorporating a pre-trained audio-visual encoder for lip modeling.

We also test the rendering speed. On an NVIDIA RTX 3090 GPU, and having preloaded the data onto the GPU, the head is outputted at 52 FPS with a resolution of 512×512. In Portrait mode, using the Portrait-Sync Generator, we achieve rendering speeds of 50 FPS with our CUDA acceleration. This far exceeds the video input speed of 25 FPS, allowing for real-time generation of video streams.

4.3. Qualitative Evaluation

Evaluation Results. To more intuitively evaluate image quality, we display a comparison between our method and other methods in Fig. 4. In this figure, it can be observed that SyncTalk demonstrates more precise and more accurate facial details. Compared to Wav2Lip [34], our method better preserves the subject's identity while offering higher fidelity and resolution. Against IP-LAP [53], our method excels in lip shape synchronization, primarily due to the audio-visual consistency brought by the audio-visual encoder. Compared to GeneFace [44], our method can accurately reproduce actions such as blinking and eyebrow-raising through expression sync. In contrast to ER-NeRF [21], our method avoids the separation between the head and body through the Pose-Sync Stabilizer and generates more accurate lip shapes. Our method achieves the best overall visual effect; we recommend watching the supplementary video for comparison.

User Study. To provide a more comprehensive evaluation of the proposed model, we design an exhaustive user study questionnaire. We extract 24 video clips, each lasting more than 10 seconds, which include head poses, facial expressions, and lip movements. Each method is represented by three of these clips. We invite 35 participants

	Wav2Lip [34]	DINet [50]	TalkLip [41]	IP-LAP [53]	AD-NeRF [14]	GeneFace [44]	ER-NeRF [21]	SyncTalk
Lip-sync Accuracy	3.839	3.696	2.893	3.161	2.696	2.982	3.189	**4.304**
Exp-sync Accuracy	3.536	3.482	2.607	3.411	2.250	3.036	2.946	**4.036**
Pose-sync Accuracy	3.571	3.571	2.875	3.696	2.232	2.929	2.607	**3.980**
Image Quality	2.500	2.696	2.054	3.571	2.464	3.482	3.036	**4.054**
Video Realness	2.929	2.429	2.429	3.161	2.036	2.732	2.518	**4.018**

Table 3. **User Study.** Rating is on a scale of 1-5; the higher the better. The term "Exp-sync Accuracy" is an abbreviation for "Expression-sync Accuracy". We highlight **best** and <u>second-best</u> results.

	PSNR ↑	LPIPS ↓	LMD ↓
Ours	**37.311**	**0.0121**	**2.8032**
replace Audio-Visual Encoder with Hubert [16]	33.516	0.0276	3.3961
replace Facial Animation Capture with ER-NeRF [21]'s	30.273	0.0415	3.0516
w/o Facial-Aware Masked-Attention	36.536	0.0165	2.9139
w/o Head-Sync Stabilizer	28.984	0.0634	3.5373
w/o Portrait-Sync Generator	32.239	0.0395	2.8154

Table 4. **Ablation study for our components.** We show the PSNR, LPIPS, and LMD in different cases.

to provide scores. The questionnaire is designed using the Mean Opinion Score (MOS) scoring protocol, asking participants to rate the generated videos from five perspectives: (1) Lip-sync Accuracy, (2) Expression-sync Accuracy, (3) Pose-sync Accuracy, (4) Image Quality, and (5) Video Realness. On average, participants take 19 minutes to complete the questionnaire, with a standardized Cronbach α coefficient of 0.96. The results of the User Study are displayed in Tab. 3. SyncTalk surpasses previous methods in all evaluations. Furthermore, we achieve the highest score in video authenticity, surpassing the second-place method, IP-LAP [53], by a margin of 20%. User studies indicate that our method can generate visually excellent quality as perceived by humans, achieving high realism.

4.4. Ablation Study

We conduct an ablation study to examine the contributions of different components in our model. We select three core metrics for evaluation: PSNR, LPIPS, and LMD. We select a subject named "May" for testing, and the results are presented in Tab. 4.

The Audio-Visual Encoder provides the primary lip sync information. When this module is replaced, all three metrics become worse, with the LMD error increasing by 21.15% in particular, indicating decreased lip motion synchronization, as shown in Fig. 5 (a), and showing that our audio-visual encoders can extract accurate lip features. Replacing Facial Animation Capture with ER-NeRF [21]'s blink module

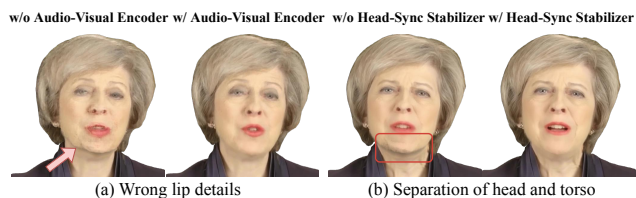(a) Wrong lip details   (b) Separation of head and torso

Figure 5. **Ablation Study** on Audio-Visual Encoder and Head-Sync Stabilizer. Removing them will lead to (a) and (b).

affects eyebrow movements and image quality.

The Facial-Aware Masked-Attention mainly alleviates motion entanglement between the lips and the rest of the face, slightly affecting the image quality after removal. Without the Head-Sync Stabilizer, all metrics significantly declined, notably LPIPS, leading to head pose jitter and head and torso separation, as shown in Fig. 5 (b). The Portrait-Sync Generator restores details like hair. Removing this module impacts the restoration of details like hair, leading to noticeable segmentation boundaries.

5. Conclusion

In this paper, we introduce SyncTalk, a highly synchronized NeRF-based method for realistic speech-driven talking head synthesis. Our framework includes a Facial Sync Controller, Head Sync Stabilizer, and Portrait Sync Generator, which maintain subject identity and generate synchronized lip movements, facial expressions, and stable head poses. Through extensive evaluation, SyncTalk demonstrates superior performance in creating realistic and synchronized talking head videos compared to existing methods. We expect that SyncTalk will not only enhance various applications but also inspire further innovation in the field of talking head synthesis.

Acknowledgments

References

[1] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence*, 44(12):8717–8727, 2018. 3

[2] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pages 173–182. PMLR, 2016. 3

[3] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020. 3

[4] Tadas Baltrušaitis, Marwa Mahmoud, and Peter Robinson. Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–6. IEEE, 2015. 6

[5] Lele Chen, Zhiheng Li, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Lip movements generation at a glance. In *Proceedings of the European conference on computer vision (ECCV)*, pages 520–535, 2018. 2

[6] Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7832–7841, 2019. 2

[7] Kun Cheng, Xiaodong Cun, Yong Zhang, Menghan Xia, Fei Yin, Mingrui Zhu, Xuan Wang, Jue Wang, and Nannan Wang. Videoretalking: Audio-based lip synchronization for talking head video editing in the wild. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 6

[8] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13*, pages 251–263. Springer, 2017. 3

[9] Dipanjan Das, Sandika Biswas, Sanjana Sinha, and Brojeshwar Bhowmick. Speech-driven facial animation using cascaded gans for learning of motion and texture. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 408–424. Springer, 2020. 2

[10] Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8649–8658, 2021. 3

[11] Kyle Gao, Yina Gao, Hongjie He, Dening Lu, Linlin Xu, and Jonathan Li. Nerf: Neural radiance field in 3d vision, a comprehensive review. *arXiv preprint arXiv:2210.00379*, 2022. 3

[12] Shreyank N Gowda, Dheeraj Pandey, and Shashank Narayana Gowda. From pixels to portraits: A comprehensive survey of talking head generation techniques and applications. *arXiv preprint arXiv:2308.16041*, 2023. 2

[13] Jiazhi Guan, Zhanwang Zhang, Hang Zhou, Tianshu Hu, Kaisiyuan Wang, Dongliang He, Haocheng Feng, Jingtuo Liu, Errui Ding, Ziwei Liu, et al. Stylesync: High-fidelity generalized and personalized lip sync in style-based generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1505–1515, 2023. 2

[14] Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5784–5794, 2021. 2, 3, 4, 5, 6, 8

[15] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[16] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021. 3, 8

[17] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Qianyi Wu, Wayne Wu, Feng Xu, and Xun Cao. Eamm: One-shot emotional talking face via audio-based emotion-aware motion model. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 2

[18] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. Deep video portraits. *ACM transactions on graphics (TOG)*, 37(4): 1–14, 2018. 1

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[20] Prajwal KR, Rudrabha Mukhopadhyay, Jerin Philip, Abhishek Jha, Vinay Namboodiri, and CV Jawahar. Towards automatic face-to-face translation. In *Proceedings of the 27th ACM international conference on multimedia*, pages 1428–1436, 2019. 2

[21] Jiahe Li, Jiawei Zhang, Xiao Bai, Jun Zhou, and Lin Gu. Efficient region-aware neural radiance fields for high-fidelity talking portrait synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7568–7578, 2023. 2, 3, 4, 5, 6, 7, 8

[22] Xian Liu, Yinghao Xu, Qianyi Wu, Hang Zhou, Wayne Wu, and Bolei Zhou. Semantic-aware implicit neural audio-driven video portrait generation. In *European Conference on Computer Vision*, pages 106–125. Springer, 2022. 3

[23] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[24] Yuanxun Lu, Jinxiang Chai, and Xun Cao. Live speech portraits: real-time photorealistic talking-head animation. *ACM Transactions on Graphics (TOG)*, 40(6):1–17, 2021. 2

[25] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7210–7219, 2021. 3

[26] Moustafa Meshry, Saksham Suri, Larry S Davis, and Abhinav Shrivastava. Learned spatial representations for few-shot talking-head synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13829–13838, 2021. 2

[27] 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. *Communications of the ACM*, 65(1):99–106, 2021. 5

[28] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012. 6

[29] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 6

[30] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022. 3, 5

[31] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *2009 sixth IEEE international conference on advanced video and signal based surveillance*, pages 296–301. Ieee, 2009. 4

[32] Ziqiao Peng, Yihao Luo, Yue Shi, Hao Xu, Xiangyu Zhu, Hongyan Liu, Jun He, and Zhaoxin Fan. Selftalk: A self-supervised commutative training diagram to comprehend 3d talking faces. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5292–5301, 2023. 1

[33] Ziqiao Peng, Haoyu Wu, Zhenbo Song, Hao Xu, Xiangyu Zhu, Jun He, Hongyan Liu, and Zhaoxin Fan. Emotalk: Speech-driven emotional disentanglement for 3d face animation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20687–20697, 2023. 2, 4

[34] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. 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*, pages 484–492, 2020. 2, 6, 7, 8

[35] Shuai Shen, Wanhua Li, Zheng Zhu, Yueqi Duan, Jie Zhou, and Jiwen Lu. Learning dynamic facial radiance fields for few-shot talking head synthesis. In *European Conference on Computer Vision*, pages 666–682. Springer, 2022. 2

[36] Linsen Song, Wayne Wu, Chen Qian, Ran He, and Chen Change Loy. Everybody's talkin': Let me talk as you want. *IEEE Transactions on Information Forensics and Security*, 17:585–598, 2022. 2

[37] Yasheng Sun, Hang Zhou, Kaisiyuan Wang, Qianyi Wu, Zhibin Hong, Jingtuo Liu, Errui Ding, Jingdong Wang, Ziwei Liu, and Koike Hideki. Masked lip-sync prediction by audio-visual contextual exploitation in transformers. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 2

[38] Jiaxiang Tang, Kaisiyuan Wang, Hang Zhou, Xiaokang Chen, Dongliang He, Tianshu Hu, Jingtuo Liu, Gang Zeng, and Jingdong Wang. Real-time neural radiance talking portrait synthesis via audio-spatial decomposition. *arXiv preprint arXiv:2211.12368*, 2022. 3, 6

[39] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural voice puppetry: Audio-driven facial reenactment. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pages 716–731. Springer, 2020. 1

[40] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Realistic speech-driven facial animation with gans. *International Journal of Computer Vision*, 128:1398–1413, 2020. 2

[41] Jiadong Wang, Xinyuan Qian, Malu Zhang, Robby T Tan, and Haizhou Li. Seeing what you said: Talking face generation guided by a lip reading expert. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14653–14662, 2023. 2, 6, 8

[42] Suzhen Wang, Lincheng Li, Yu Ding, Changjie Fan, and Xin Yu. Audio2head: Audio-driven one-shot talking-head generation with natural head motion. *arXiv preprint arXiv:2107.09293*, 2021. 2

[43] Shunyu Yao, RuiZhe Zhong, Yichao Yan, Guangtao Zhai, and Xiaokang Yang. Dfa-nerf: Personalized talking head generation via disentangled face attributes neural rendering. *arXiv preprint arXiv:2201.00791*, 2022. 2, 3, 5

[44] Zhenhui Ye, Ziyue Jiang, Yi Ren, Jinglin Liu, JinZheng He, and Zhou Zhao. Geneface: Generalized and high-fidelity audio-driven 3d talking face synthesis. *arXiv preprint arXiv:2301.13430*, 2023. 2, 3, 4, 6, 7, 8

[45] Chenxu Zhang, Yifan Zhao, Yifei Huang, Ming Zeng, Saifeng Ni, Madhukar Budagavi, and Xiaohu Guo. Facial: Synthesizing dynamic talking face with implicit attribute learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3867–3876, 2021. 2

[46] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[47] Wenlong Zhang, Yihao Liu, Chao Dong, and Yu Qiao. Ranksrgan: Generative adversarial networks with ranker for image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3096–3105, 2019. 6

[48] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8652–8661, 2023. 2

[49] 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. 2

[50] Zhimeng Zhang, Zhipeng Hu, Wenjin Deng, Changjie Fan, Tangjie Lv, and Yu Ding. Dinet: Deformation inpainting network for realistic face visually dubbing on high resolution video. *arXiv preprint arXiv:2303.03988*, 2023. 2, 6, 8

[51] Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C Bühler, Xu Chen, Michael J Black, and Otmar Hilliges. Im avatar: Implicit morphable head avatars from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13545–13555, 2022. 3

[52] Yufeng Zheng, Wang Yifan, Gordon Wetzstein, Michael J Black, and Otmar Hilliges. Pointavatar: Deformable point-based head avatars from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21057–21067, 2023. 3

[53] Weizhi Zhong, Chaowei Fang, Yinqi Cai, Pengxu Wei, Gangming Zhao, Liang Lin, and Guanbin Li. Identity-preserving talking face generation with landmark and appearance priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2023. 2, 6, 7, 8

[54] 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*, pages 9299–9306, 2019. 2

[55] 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. 2

[56] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makelttalk: speaker-aware talking-head animation. *ACM Transactions On Graphics (TOG)*, 39(6):1–15, 2020. 2

[57] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Instant volumetric head avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4574–4584, 2023. 3