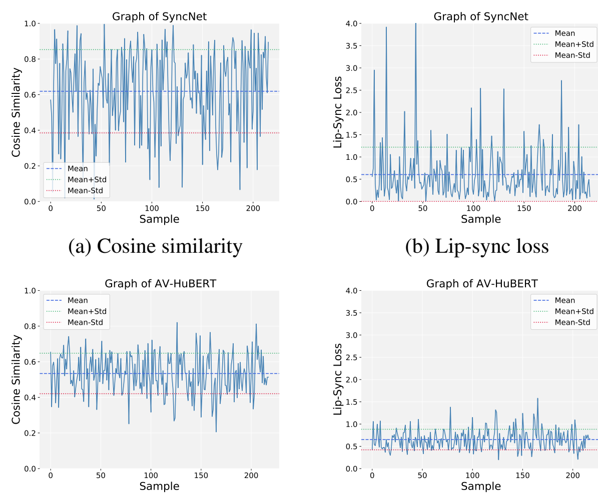
Audio-Visual Speech Representation Expert for Enhanced Talking Face Video Generation and Evaluation

Dogucan Yaman[1] Fevziye Irem Eyiokur[1] Leonard Bärmann[1] Seymanur Aktı[1]
Hazım Kemal Ekenel[2] Alexander Waibel[1,3]
[1]Karlsruhe Institute of Technology, [2]Istanbul Technical University, [3]Carnegie Mellon University
dogucan.yaman@kit.edu

Abstract

In the task of talking face generation, the objective is to generate a face video with lips synchronized to the corresponding audio while preserving visual details and identity information. Current methods face the challenge of learning accurate lip synchronization while avoiding detrimental effects on visual quality, as well as robustly evaluating such synchronization. To tackle these problems, we propose utilizing an audio-visual speech representation expert (AV-HuBERT) for calculating lip synchronization loss during training. Moreover, leveraging AV-HuBERT's features, we introduce three novel lip synchronization evaluation metrics, aiming to provide a comprehensive assessment of lip synchronization performance. Experimental results, along with a detailed ablation study, demonstrate the effectiveness of our approach and the utility of the proposed evaluation metrics.



(a) Cosine similarity (b) Lip-sync loss
(c) Cosine similarity (d) Lip-sync loss

Figure 1. Cosine similarity and lip-sync loss between GT audio-lip pairs on random LRS2 test samples, showcasing the instability of SyncNet [28] **(a, b)** and more robust performance of AV-HuBERT **(c, d)**.

1. Introduction

The goal of talking face generation is to create a video based on provided face and audio sequences, seeking synchronized lip movements that match the given audio while maintaining the identity and visual details. This task has gained considerable interest recently for its diverse applications, such as face dubbing and enhancement in video conferencing tools, film dubbing, and content creation [50, 54].

In the talking face generation task, visual quality of the generated faces and audio-lip synchronization (lip sync) are essential but also challenging aspects to have a natural video. Since visual artifacts and out-of-sync lip movements are easily recognizable by the audience, they significantly diminish the naturalness of the dubbed video. While visual quality is addressed with approaches across various domains, lip sync takes precedence as it is particular and crucial for talking face generation in maintaining the naturalness of dubbed videos. Up to now, several different works tackled the challenges of learning lip sync. The widely adopted method involves extracting features from both the audio and face sequences by a model that has been contrastively trained with audio-face pairs to learn lip sync. These features are then compared to measure the synchronization between the two modalities (e.g., with cosine similarity). Therefore, it is essential to extract meaningful as well as robust audio and visual features, as the speech representations from both modalities significantly influence the synchronization measurement. For this, the most common method is using a *lip-expert*, which is a slightly modified and retrained version of SyncNet [11][1]. Specifically, the features from the lips and audio sequence are extracted by the SyncNet [28] and the cosine similarity along with the

[1]Therefore, SyncNet name is being used for the original SyncNet [11] and *lip-expert* [28] interchangeably, as we also do it hereinafter.

cross-entropy loss is computed. This loss strategy is called *lip-sync loss*. Significant advancements have been made since the introduction of using this SyncNet [28] and most of the methods have benefited from this approach to learn lip sync in the literature. Recently, TalkLip [41] proposes using a lip-reading expert, AV-HuBERT [31, 32], for audio and visual feature extraction. Then, contrastive learning is employed to learn lip sync, yielding superior performance.

In Figs. 1a and 1b, we share our analysis about the performance of SyncNet [28] on LRS2 [1] ground-truth (GT) audio-lip pairs. The cosine similarity and lip-sync loss show fluctuations even on GT samples, contrary to the anticipated stable and high performance. This outcome implies that SyncNet [28] has significant stability and reliability issues that lead to poor lip sync performance. Moreover, we empirically find that using SyncNet [28] causes severe visual quality issues and unstable training, despite enhanced lip sync performance. To address these problems, inspired by TalkLip [41], we employ a pretrained audio-visual speech representation learning model, AV-HuBERT [31, 32], which was finetuned for the lip reading task, to extract audio and lip features. In contrast to TalkLip, we utilize cross-entropy-based lip-sync loss [28] to guide our model during training, providing a stabilized training signal (Figs. 1c and 1d). We refer to this as approach as *unsupervised*. In addition to this, we investigate two further methods as loss functions. Specifically, we employ AV-HuBERT features and compute the lip-sync loss by using the visual features of the generated faces and GT faces. We call this approach *visual-visual*, as the audio is not involved. We also obtain features from generated face-audio pairs and GT face-audio pairs with AV-HuBERT model for lip-sync loss and term this *multimodal*, since we acquire the features from the multimodal representation (face-audio pairs). We conduct an ablation study about lip sync learning by comparing these introduced approaches and present the results in Sec. 4.5.

Besides training a model with high-quality audio-lip synchronization, proper and robust evaluation of such capabilities is another key aspect of talking face generation, essential for analyzing and comparing different methods. One of the first lip sync evaluation metrics is Mouth Landmark Distance (LMD) [6], focusing on computing the distance between the landmarks in the mouth region of the generated faces and GT faces. However, this metric does not disentangle the lip sync from visual factors, as it is sensitive against shifting in the spatial domain. Additionally, lips with different articulatory parameters (e.g., aperture and spreading) yield poor LMD score, although they are still synchronized. More recent metrics, LSE-C & -D, are based on SyncNet [11] features and measure the confidence and distance scores to represent lip sync. The advantage of these metrics is that they do not require GT data, directly measuring the alignment between audio and faces. However, unstable SyncNet [11] performance makes these two metrics unreliable and also vulnerable to affine transformation. One of the main reasons of this are the poor shift-invariant characteristics of SyncNet [11]. To tackle these problems in lip sync evaluation, we propose three novel complementary metrics: Unsupervised Audio-Visual Synchronization (AVS_u), Multimodal Audio-Visual Synchronization (AVS_m), and Visual-only Lip Synchronization (AVS_v). We leverage the pretrained AV-HuBERT lip-reading expert for feature extraction and utilize cosine similarity for the score calculation. The differences and details of these metrics are presented in Sec. 4.1. Our contributions are summarized as follows:

- We propose to use a pretrained audio-visual speech representation learning model (AV-HuBERT), finetuned for lip reading task, for feature extraction from the audio and face sequences for lip-sync loss in training.
- We introduce three novel evaluation metrics by employing AV-HuBERT for feature extraction, yielding less vulnerable and more consistent assessments of performance.
- We conduct extensive experiments and ablation studies to demonstrate the effectiveness of our contributions.

2. Related Work

Talking face generation Traditional methods focus on achieving time-aligned videos by choosing the most fitting image-audio pairs [4, 35, 48]. Later on, the facial landmark representation for face generation is employed to obtain a synchronized lip [7, 12, 55, 58]. However, these methods suffer from poor lip sync, despite controllable face generation. Wav2Lip [28] proposes a *lip-expert*, modified SyncNet [11], and also a lip-sync loss to guide the model for lip sync learning. It shows superior lip sync performance. PC-AVS [57] introduces a pose-controllable 2D talking face generation without using any intermediate representation (e.g., facial landmarks, 3D head representation). GC-AVT [21] employs a similar approach as PC-AVS but involves emotion-controllable face generation. SyncTalkFace [27] benefits from an audio-lip memory mechanism to store and retrieve lip motion representation to achieve enhanced lip sync. On the other hand, VideoReTalking [8] reveals fundamental problems in talking face generation, namely the effect of lip motion of the identity reference over the talking face generation. To solve this problem, they transform the identity reference to have canonical expression with stable and flat lips, which yields improved training stability and lip sync. DINet [53] employs a deformation module to improve the pose alignment and lip sync. Recently, LipFormer [42] introduced a pre-learned facial codebook-based method to learn HR video generation by overcoming existing challenges. TalkLip [41] employs a global audio encoder to capture the content in the speech and also introduces AV-HuBERT-based audio-visual feature extraction for lip sync learning along with contrastive learning. However, TalkLip has severe visual artifacts, despite superior lip sync. This method is the closest approach to our talking face generation. However, we use lip-sync loss instead of contrastive learning and investigate two further methods for lip sync learning. Moreover, we achieve better visual quality without artifacts. SIDGAN [24] presents crucial analyses of learning synchronization. They also provide a shift-invariant model, similar to the lip-expert & SyncNet, for feature extraction to guide the model for lip sync learning. On the other hand, one of the first attempts to generate synchronized lips as a part of a system was done in [29] and the follow-up paper [39] also contains an entire system that involves speech-to-speech translation and face dubbing. These systems are also well-suited for utilization in meeting rooms, facilitating seamless communication among individuals conversing in different languages [38]. In addition to the above 2D-based approaches, Neural Radiance Fields-based (NeRFs) and 3D-based methods aim at synthesizing the entire head by representing the head in the 3D space [2, 3, 14, 22, 26, 30, 34, 36, 37, 40, 45–47, 49, 51, 52, 56, 58]. Though they are capable of controlling the pose and emotion much better as well as achieving enhanced visual quality, they have severe lip sync performance, yielding unrealistic videos.

Lip sync evaluation The very first proposed metric is Lip Landmark Distance (LMD) [6]. However, it has several issues, as we mentioned in Sec. 1 and Sec. 4.1. After the SyncNet [11] and Sync scores [11] were proposed, they demonstrated a more convenient performance than LMD. Later, Wav2Lip [28] proposed LSE-C and LSE-D metrics (confidence and distance) using SyncNet [11] audio and visual features, which became the gold standard in the literature. Despite the advantage of not requiring GT data, the unreliable performance of SyncNet makes these two metrics vulnerable. The Word Error Rate (WER) has recently been proposed as an evaluation metric for reading intelligibility [41]. In this work, we propose three novel lip sync evaluation metrics by utilizing a robust pretrained audio-visual speech representation learning model, AV-HuBERT.

3. Talking Face Generation

We propose a talking face generation approach to enhance lip sync and visual quality by leveraging an audio-visual lip-reading expert named AVHubert [31, 32]. Fig. 2 provides an overview of our model, which takes three inputs: an identity reference, a bottom-half masked pose reference, and an audio snippet. The face generator is responsible for synthesizing a set of images to retain synchronized lips with respect to the given audio while preserving the identity and visual quality. We extract features from the generated samples with the AV-HuBERT and then calculate lip-sync loss in the training along with the other losses.

3.1. Face Encoder

In 2D talking face generation, the main approach is to provide an identity reference and a pose reference to the model, which aims to synthesize a modified version of the pose reference with lip movements matching the given audio. Since the target image and the pose reference are the same, the mouth region of the pose reference has to be masked. On the other hand, the identity reference is utilized to preserve the subject's identity and is randomly sampled from a different part of the input video than the pose reference. To encode the pose and identity references, we employ two individual encoders. This approach demonstrates superior performance than using a single encoder for both modalities as in the traditional approaches, since each encoder follows its own objective more effectively, yielding better visual feature representation for both inputs [24]. Our identity and pose encoders share the same architecture. They have consecutive convolutional blocks and each block involves one strided-convolution layer followed by two non-strided convolution layers. After each layer, we employ a ReLU activation function [20, 25] and a batch normalization layer [17].

3.2. Audio Encoder

Our audio encoder embeds the mel-spectrogram representation of the audio snippet, acting as a condition for the face generator to drive the model to generate accurate lip movements, $E_A(A) = F^A \in \mathbb{R}^{1\times 1\times 512}$. We employ the audio encoder of the SyncNet [28] without finetuning, since it was trained in conjunction with a face encoder to learn lip synchronization. In this way, the audio encoder provides a feature representation more suitable for the purpose of generating synchronized lips.

3.3. Video Generation

The talking face generator takes the combination of features of three encoders. We first concatenate the identity features and pose features along the depth dimension. Before feeding the face generator, we process these features through a convolution layer to reduce the depth. Subsequently, we concatenate the output with the audio embedding to input the face generator. Our face generator consists of consecutive transposed convolution layers. Following each transposed convolution layer, we utilize two convolution layers with a stride one. Similar to the image encoders, we employ the ReLU activation function [20, 25] and batch normalization [17] after each layer. We also apply residual connections [15] between the reciprocal layers of the face encoders (both identity and pose encoders) and the face generator.

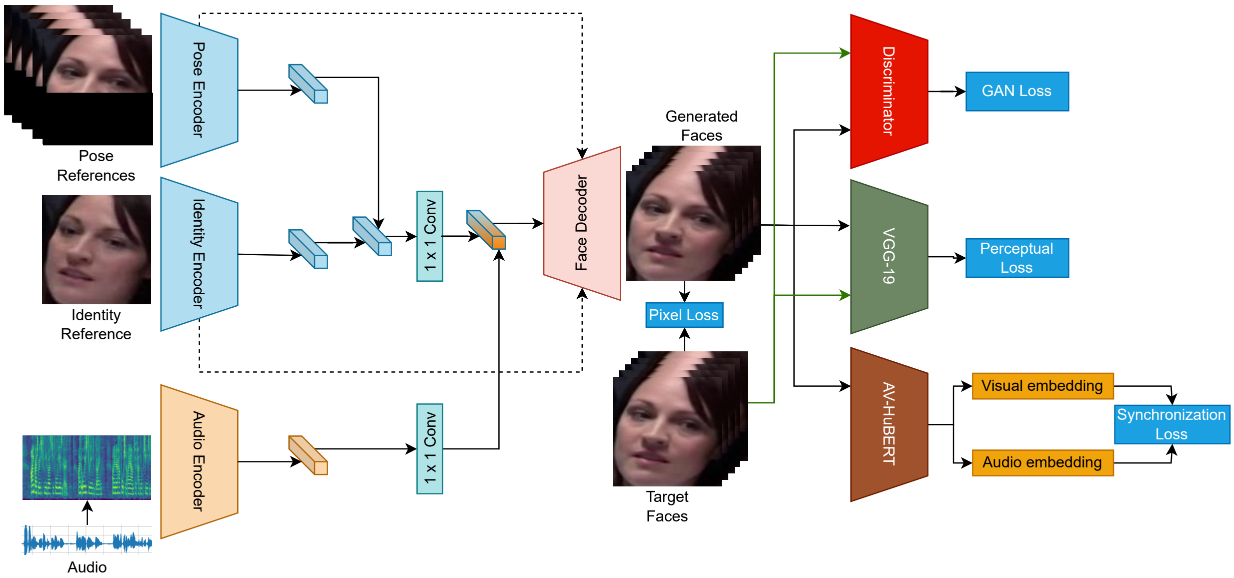


Figure 2. Illustration of the proposed audio-driven talking face generation model and employed loss functions.

This strategy yields the retention of high-level features and increased stabilization throughout the training.

3.4. Audio-Visual Speech Representation Expert

Audio-Visual Hidden Unit BERT (AV-HuBERT) [31, 32] is a self-supervised representation learning model for audio and visual data. The model processes a face sequence (the mouth region) with a modified version of ResNet-18 and the Mel-frequency cepstral coefficients (MFCC) of an audio sequence with a linear projection layer followed by a normalization step based on per-frame statistics. The fusion of audio and visual features is processed through transformer blocks to predict masked cluster assignments. Thus, AV-HuBERT learned robust audio-visual speech representation. We utilize the finetuned version of AV-HuBERT for lip reading, since it yields better lip sync as well as reading intelligibility [41]. In Fig. 1, we show the performance of AV-HuBERT on LRS2 GT data for measuring the cosine similarity and lip-sync loss. The graphs clearly show that the AV-HuBERT has more stable performance and less fluctuation compared to SyncNet [28] used in Wav2Lip [28].

3.5. Lip Synchronization Loss

We utilize the pretrained AV-HuBERT model [31, 32] and extract features from the final layer of the transformer encoder block for audio and video modalities: $F^A_{AVH} \in \mathbb{R}^{T\times768}, F^V_{AVH} \in \mathbb{R}^{T\times768}$. Along with [41], we empirically found that extracting features from the entire video instead of a short sequence yields better audio-visual feature alignment. Therefore, we specifically replace the corresponding interval of the ground-truth video with the generated face sequence (Fig. 3). We then extract features from the video using the cropped lips from the face sequence and audio sequence to calculate the lip sync between audio and visual features. However, since a major part of the video is ground-truth data, it is anticipated to have high audio-visual alignment, yielding insignificant effects of the generated samples on lip synchronization evaluation. To tackle this problem, we only consider the generated samples in the feature space for lip sync loss, as the AV-HuBERT feature extractor provides the feature representation per time step, $T \times D$. Specifically, we take the audio and visual AV-HuBERT features corresponding to the generated time interval. Subsequently, we compute cosine similarity between these two feature representations followed by binary cross-entropy loss as follows:

$$L_{sync} = -log(CS(F^{A_{t:t+k}}_{AVH}, F^{V_{t:t+k}}_{AVH})) \quad (1)$$

where CS indicates cosine similarity and $t : t+k$ represents the time interval of the generated part of the video. k is the same as the length of the face sequence generated by the talking face generation model in a single forward pass.

3.6. Implementation Details

Adversarial loss We utilize the GAN (adversarial) loss [13] to train our talking face generation model. For this, we employ a discriminator model, which is responsible for distinguishing real (target data) and fake samples (generated data) to guide the generator. Meanwhile, the generator attempts to synthesize appropriate images so that the discriminator cannot determine whether the sample is real or fake. Our discriminator network contains 7 consecutive strided-convolutional layers along with Leaky ReLU activation function and spectral normalization [23].

Perceptual loss In order to preserve the identity and textures, we utilize perceptual loss [18] as feature reconstruction loss by extracting features from the generated faces and GT faces from different layers of the pretrained VGG-19 model [33]. Afterward, we calculate the L2 distance between extracted features, as shown below. While c_i indicates weight coefficients, ϕ states the selected layers for the feature extraction. I^G and I^{GT} are the generated image and the GT image, respectively. We follow [18] for determining the coefficients and layers.

$$L_{per} = \sum_{i=1}^{5} c_i ||VGG^{\phi_i}(I^G) - VGG^{\phi_i}(I^{GT})||_2 \tag{2}$$

Pixel reconstruction loss Despite perceptual loss to capture identity and textural details, pixel-level reconstruction loss is required to capture fine-grained details and generate consistent images. Therefore, we employ a reconstruction loss in the pixel space: $L_{pixel} = ||I^G - I^{GT}||_1$

Total loss By combining all the presented loss functions, the total loss is as follows:

$$L = L_{GAN}(G, D) + \lambda_1 L_{pixel}(G) + \lambda_2 L_{per}(G) + \lambda_3 L_{sync} \tag{3}$$

where G and D denote the generator and discriminator. We determined the best coefficients through empirical analysis as follows: $(\lambda_1, \lambda_2, \lambda_3) = (10, 1, 0.5)$.

Training details In each forward pass, we generate a set of images (denoted by k) to ensure temporal consistency. Following the literature, we set $k = 5$. We use FAN [5] for face detection and obtain tight crops as input. We resize the faces to 96×96 resolution since lip sync learning in high resolution holds further challenges [24] and the LRS2 dataset [1] has low-resolution faces. Then, we can apply GFPGAN [43] face enhancement method to the output video to increase the resolution for obtaining HR videos, if necessary. Our audio encoder takes a mel-spectrogram of size 16×80 extracted from $16kHz$ audio. The hop and the window sizes are 200 and 800, respectively. We utilize the Adam optimizer [19]. The learning rate is set to 1×10^{-4}.

4. Experimental Results

Dataset We developed our talking face generator by utilizing the standard benchmark in the domain due to its diversity in terms of number of subjects: Lip Reading Sentence 2 (LRS2) [1] training set. The evaluation was conducted on the LRS2 [1], LRW [10], and HDTF [52] datasets.

4.1. Metrics

For visual quality assessment, we utilize benchmark metrics in this field: FID [16], SSIM [44], and PSNR. In evaluating lip sync, as in the literature, we employ Mouth Landmark Distance (LMD) [6] as well as LSE-C & LSE-D metrics [28], which measure the confidence and distance scores via a pretrained SyncNet [11]. We also conduct a user study to perform human evaluation. We compare our model with the state-of-the-art models that have publicly available codes and models, ensuring a fair comparison by evaluating under the same conditions.

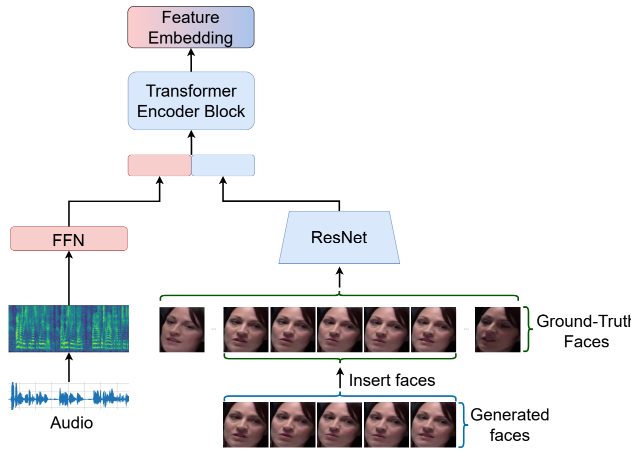


Figure 3. In the training, since extracting features from entire videos provides more informative features [41], we insert the generated faces into the corresponding part in the target video. Then, we use this video for feature extraction after cropping the mouth.

The three widely utilized synchronization metrics in the literature have crucial problems. The mouth landmark distance (LMD) measures the spatial difference between the mouth landmark points in the generated face and the target face. However, this metric faces the following three issues: (1) It is sensitive to errors in landmark detection. (2) It cannot disentangle the synchronization and the generation stability. For instance, if the model accurately generates lip movements but introduces errors in the mouth region (such as shifting the mouth/face), the landmark positions will subsequently change. Although this does not affect the synchronization, it leads to a higher landmark distance, highlighting the need for disentanglement. (3) LMD fails to consider lip movements or shapes properly. Despite having the same lip shape, variation in the lips aperture and spreading cause misleading scores.

Recently proposed LSE-C and LSE-D metrics are more informative than LMD. However, they also hold vital issues. These metrics rely on audio and lip features extracted by the pretrained SyncNet model [11], which contains audio and image encoders, and was trained with audio-lip pairs for learning lip sync. However, SyncNet is vulnerable against translations in the data due to not being properly shift invariant [24] (see Fig. 5c). Therefore, small translations in the face affects LSE-C and LSE-D metrics, resulting in not fully

Table 1. Quantitative results on the test sets of LRS2 and LRW. While green indicates the best score, yellow shows the second best.

	LRS2									LRW								
Method	SSIM ↑	PSNR ↑	FID ↓	LMD ↓	LSE-C ↑	LSE-D ↓	AVS_u ↑	AVS_m ↑	AVS_v ↑	SSIM ↑	PSNR ↑	FID ↓	LMD ↓	LSE-C ↑	LSE-D ↓	AVS_u ↑	AVS_m ↑	AVS_v ↑
Wav2Lip [28]	0.865	26.538	7.05	2.388	7.594	6.759	0.248	0.659	0.289	0.851	25.144	6.81	2.147	7.490	6.512	0.242	0.537	0.268
VRT [8]	0.841	25.584	9.28	2.612	7.499	6.824	0.361	0.763	0.426	0.873	27.110	5.30	2.390	6.598	7.123	0.383	0.758	0.557
DINet [53]	0.785	24.354	4.26	2.301	5.376	8.376	0.291	0.758	0.425	0.886	27.501	8.17	1.963	5.249	9.099	0.276	0.638	0.360
TalkLip [41]	0.860	26.112	4.94	2.344	8.530	6.086	0.570	0.895	0.702	0.868	26.349	15.73	1.836	7.281	6.485	0.581	0.813	0.604
IPLAP [55]	0.877	29.670	4.10	2.119	6.495	7.165	0.331	0.711	0.479	0.917	30.456	8.40	1.641	5.949	7.767	0.272	0.593	0.331
Ours	0.947	31.273	4.51	1.188	7.958	6.301	0.508	0.939	0.879	0.919	30.185	6.21	1.487	7.738	6.456	0.554	0.856	0.762

disentangled lip synchronization evaluation. Moreover, the margin around the faces has also impact on extracted features by SyncNet, yielding inconsistent LSE-C & D scores. To tackle these problems, we introduce three novel lip synchronization evaluation metrics. We employ AV-HuBERT, a robust audio-visual speech representation learning model, to obtain superior audio and visual feature representations to measure synchronization.

Unsupervised Audio-Visual Synchronization ($\mathbf{AVS}_u$) In this metric, we measure the lip sync by only considering the given audio and the generated video. Specifically, we extract audio and visual features (lips) from the transformer encoder block of AV-HuBERT. Subsequently, we compute the cosine similarity between these two feature representations, as several works demonstrate the superiority of cosine similarity for audio-visual alignment [28, 41] and speaker recognition [9] . This metric does not require GT data and thus can flexibly be applied to any data, as LSE-C & D. As AV-HuBERT provides more robust feature representation than SyncNet [11], this metric is more consistent, reliable, and not vulnerable to translation (see Fig. 1 and Fig. 5).

$$\text{AVS}_u = CS(TE(V^G_{1:T}, \mathbf{0}), TE(\mathbf{0}, A_{1:T})) \quad (4)$$

where CS indicates the cosine similarity and TE represents the transformer encoder block from which we extract features. V^G and A state the lips of the generated faces and audio sequence, respectively. Please note that we give lips and audio to the AV-HuBERT model individually as we need to extract separate features to calculate the similarity between them. However, the audio-visual transformer encoder of AV-HuBERT requires both modalities together. Therefore, we follow the approach proposed by Shi et al. [31] and provide a placeholder of zeros with equivalent dimensions.

Multimodal Audio-Visual Synchronization ($\mathbf{AVS}_m$) In this metric, we employ the generated and GT videos as pairs to measure the synchronization. We first extract features from the generated video using the mouth region of the faces and audio. We then repeat the same procedure with the GT video. In the end, we calculate the cosine similarity between these two embeddings. Intuitively, this metric considers the similarity between the alignment of the generated lips-audio and GT lips-audio pairs.

$$\text{AVS}_m = CS(TE(V^G_{1:T}, A_{1:T}), TE(V^{GT}_{1:T}, A_{1:T})) \quad (5)$$

Visual-only Lip Synchronization ($\mathbf{AVS}_v$) We only employ the lips of the generated faces and GT faces without involving audio. Therefore, this metric only focuses on the visual shape similarity of the lips. As AV-HuBERT is finetuned for lip-reading purpose, the extracted visual features hold the information for lip reading, yielding meaningful representation for lip synchronization and intelligibility.

$$\text{AVS}_v = CS(TE(V^G_{1:T}, \mathbf{0}), TE(V^{GT}_{1:T}, \mathbf{0})) \quad (6)$$

4.2. Quantitative Evaluation

We demonstrate quantitative results on the test data of LRS2 and LRW in Tab. 1 and HDTF in Tab. 2. On LRS2 and HDTF, we surpass all other methods in the visual quality, except FID on LRS2, where we are slightly behind IPLAP and DINet. On LRW, we achieve similar visual quality scores with IPLAP for SSIM and PSNR, and with VRT for FID. In summary, we surpass other methods with a large margin in most of the visual quality metrics.

In LMD, we achieve state-of-the-art results on all three datasets. This outcome does not only show the synchronization performance but also demonstrate the visual stability of our generated outputs. In LSE-C and LSE-D metrics, we have state-of-the-art performance on LRW dataset. While TalkLip has better scores on LRS2, Wav2Lip achieves the highest score on HDTF. However, when considering our user study (Sec. 4.4), Wav2Lip is far behind our method on HDTF in synchronization performance, as in LRS2 and LRW datasets. This shows the vulnerability and inconsistency of LSE-C & D metrics. Fig. 5c illustrates the sensitive performance of LSE-C & D metrics when the visual data is horizontally shifted [24]. It clearly proves that these metrics are extremely sensitive to translation in the data, measuring poor lip synchronization performance when the face is shifted while preserving the same lip shape. Moreover, SyncNet [11] demonstrates similar fluctuation on GT data as in Fig. 1a. All these analyses and outcomes validate the motivation of our proposed new lip synchronization metrics.

According to our AV-HuBERT-based lip sync metrics, we achieve state-of-the-art results on all three datasets in AVS_m and AVS_v. On the other hand, TalkLip surpasses us with a small difference on all three datasets in the AVS_u metric. In Fig. 1c, we illustrate the analysis of cosine simi-

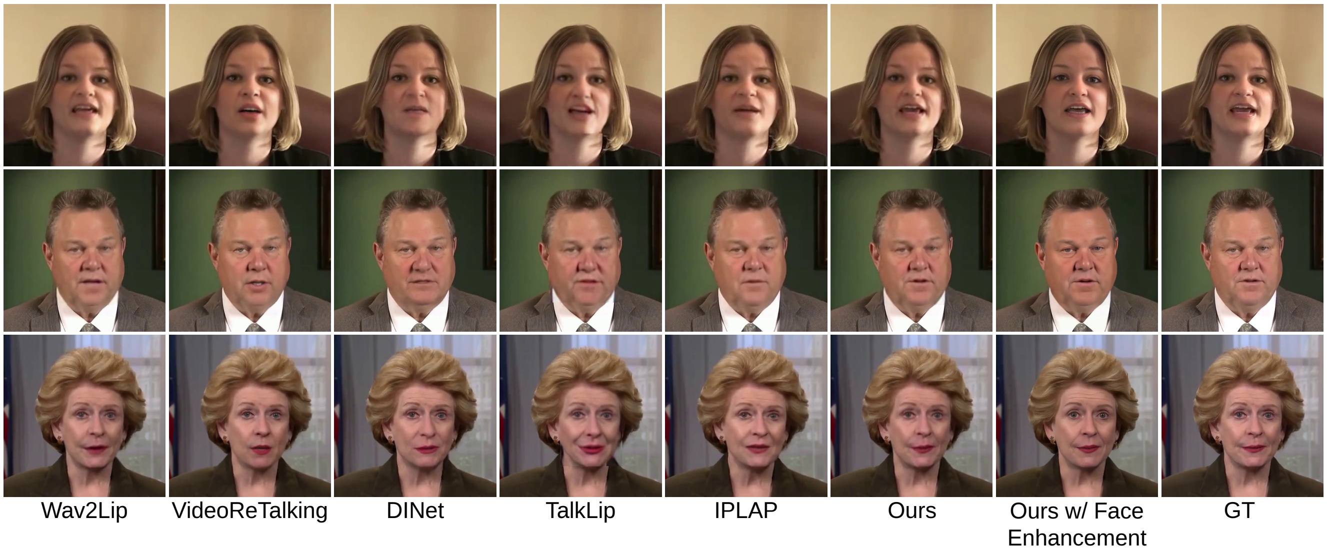
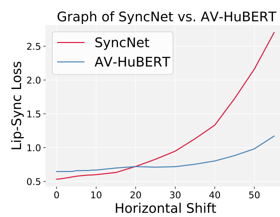


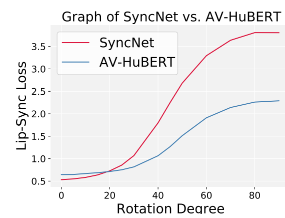
Figure 4. Qualitative comparison of our approach with state-of-the-art models and ground-truth data on HDTF

Table 2. Quantitative results on HDTF [52].

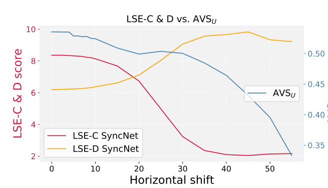
Method	SSIM ↑	PSNR ↑	FID ↓	LMD ↓	LSE-C ↑	LSE-D ↓	AVS_u ↑	AVS_m ↑	AVS_v ↑
Wav2Lip	0.841	24.812	35.41	1.341	9.054	6.414	0.297	0.514	0.358
VideoReTalking	0.830	24.551	29.77	3.085	6.121	7.368	0.384	0.677	0.570
TalkLip	0.820	25.229	25.10	2.981	6.189	7.276	0.591	0.823	0.730
IPLAP	0.869	27.801	22.09	2.217	5.563	8.495	0.459	0.661	0.528
Ours	0.933	30.579	16.76	1.292	8.106	6.765	0.538	0.892	0.783



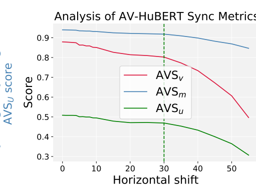
(a) Horizontal shift analysis



(b) Rotation analysis



(c) Comparison of evaluation metrics



(d) Analyses of our metrics

Figure 5. **(a,b)** shows the performance analyses of SyncNet [28] and AV-HuBERT features for lip-sync loss on GT LRS2 data with the horizontal shift and rotation in spatial space. It clearly shows that SyncNet [28] is not shift invariant and vulnerable to the affine transformation, while AV-HuBERT demonstrates robust performance. **(c)** compares the LSE-C & D metrics with our AVS_u metric while applying horizontal shifting in the spatial space. While LSE-C & D scores are aligned with the left axis, AVS_u is alifned with the right one. **(d)** analyses our three metrics under the shifting conditions. Since AVS_v and AVS_m require generated data-GT pairs, we use GT LRS2 data and our model's output. On the other hand, for **(a,b,c)**, we only use GT LRS2 data.

larity between AV-HuBERT lip and audio features on LRS2 GT data (similar to Fig. 1a). The graph shows that the AV-HuBERT features are more stable than SyncNet features. Similarly, Fig. 5c and Fig. 5d also confirm the stability of our three novel metrics compared to LSE-C & D. Considering these analyses as well as the results of our user study, it is clear that our proposed lip synchronization evaluation metrics provide more insight about the lip synchronization and also show more reliable performance.

4.3. Qualitative Evaluation

In Fig. 4, we present a qualitative comparison with SOTA models. We employ the respective publicly available models of the compared methods while generating videos. We choose the HDTF dataset to present results from an unseen dataset, except for DINet since it was trained on the HDTF dataset. The results clearly show that our model surpasses all other methods in terms of having the most similar lip shapes with the GT face. Besides, mouth region and teeth have lower quality in Wav2Lip. Moreover, TalkLip has a severe pose stability issues and shows artifacts around the face. This clearly degrades the naturalness of a video. Our user study in Tab. 3 validates the poor visual quality of TalkLip, DINet, and Wav2Lip. In summary, the qualitative analysis and user study demonstrate the superiority of our model in terms of the lip synchronization and visual quality, resulting in natural talking face generation. Moreover, we present results in Fig. 4 by employing a face enhancement model, GFPGAN, to show that our model's results can be improved with a post-processing step to produce high resolution faces

Table 3. User study on randomly selected HDTF videos [52].

Method	Sync ↑	Visual ↑	Overall ↑
Wav2Lip [28]	2.91	2.88	2.73
VideoReTalking w/ FR [8]	3.05	3.70	3.46
DINet [53]	2.50	2.34	2.48
TalkLip [41]	3.32	2.05	2.08
IPLAP [55]	2.62	3.86	3.27
Ours	**3.92**	**4.02**	**3.95**

Table 4. Ablation study on LRS2 dataset for AV-HuBERT-based synchronization losses.

Method	SSIM ↑	PSNR ↑	FID ↓	LMD ↓	LSE-C ↑	LSE-D ↓	AVS$_u$ ↑	AVS$_m$ ↑	AVS$_v$ ↑
Baseline	0.864	26.424	12.25	2.423	7.116	7.396	0.301	0.637	0.423
Visual-visual	0.905	29.248	15.14	1.798	7.481	6.556	0.381	0.765	0.545
Multimodal	0.910	30.014	6.11	1.774	6.998	6.794	0.395	0.789	0.575
Unsupervised	**0.947**	**31.273**	**4.51**	**1.188**	**7.958**	**6.301**	**0.508**	**0.939**	**0.879**

Figure 6. Qualitative samples from the ablation study

for high-resolution talking face video generation.

4.4. User Study

We conduct a user study to explore how the generated videos look to humans. We randomly select ten videos from the HDTF dataset and generate these videos with each model to use in the user study along with the GT videos. Users were shown aligned videos of multiple faces generated by different methods and asked to rank them on lip synchronization, visual quality and overall quality. In total, ten different participants joined the user study and we present the results in Tab. 3. The scores are scaled between 1 (worst) and 5 (best). According to the results, we outperform all other models in lip synchronization, visual quality, and overall quality. TalkLip demonstrates the second-best synchronization performance. However, its visual quality issues and artifacts yield the lowest visual quality score.

4.5. Ablation Study

We conduct an ablation study to show the effect of using AV-HuBERT features in lip-sync loss throughout training. For this, we first train our model with *lip-expert* [28] features for calculating lip-sync loss. We call this model *baseline* in Tab. 4. *Unsupervised* represents the method that we propose to use. Specifically, we extract features from the audio and lip sequences using the AV-HuBERT transformer encoder. Afterward, we calculate lip-sync loss to explicitly measure the synchronization performance of the model in the training. We also employ AV-HuBERT in two more ways to compare with our approach, following the AVS$_v$ and AVS$_m$ metrics. In the *visual-visual* approach, we extract only visual features by feeding the AV-HuBERT model with the generated lips and GT lips, individually. Then, we apply lip-sync loss between these two embeddings without involving audio. This method obtains better scores than *baseline*. We further apply the *multimodal* strategy, extracting features from generated lips-audio pairs and also from GT lips-audio pairs, and then applying lip-sync loss thereupon. The *multimodal* approach enhances the visual quality compared to *baseline* and *visual-visual* approaches but decreases the LSE-C & D scores even below the *baseline*. On the other hand, in our synchronization metrics, *multimodal* outperforms the *baseline* as well as *visual-visual* method. Finally, the best results in visual quality and lip synchronization are achieved by employing the *unsupervised* approach according to the ablation study.

In Fig. 6, we share the sample images from the approaches presented in Tab. 4. The *baseline* has distinguishable face borders, artifacts in the mouth region, and not fully aligned lip shape. Although *visual-visual* shows enhanced visual quality and improved lip shape, *multimodal* approach generates more appropriate lip shapes. Finally, *unsupervised* is able to generate the best fitting lip shapes as well as enhanced visual quality. Specifically, the teeth have better visual quality and are more similar to the GT in terms of the characteristic features of the subject (identity).

5. Conclusion

We propose to use the pretrained audio-visual speech representation expert AV-HuBERT for training a talking face generation network with high-quality audio-lip synchronization. Furthermore, we utilize this network to obtain three complementary and robust metrics for evaluating lip synchronization. Our experimental results demonstrate the effectiveness of our approach. We also analyze the proposed metrics for robustness and validate their alignment with human preferences through a user study.

Limitations AV-HuBERT and its features should be investigated further to employ them more efficiently for lip synchronization, despite increased performance and stability in the training. Furthermore, the sample size of our user study could be increased to gain more statistical power.

Ethics & Social Impact Talking face generation is essential for a wide range of applications. However, its vulnerability and potential for misuse (e.g., deepfake) pose significant risks. We will apply Watermarking and take necessary precautions to prevent unauthorized usage of our model.

Acknowledgement This work was supported in part by the European Commission project Meetween (101135798) under the call HORIZON-CL4-2023-HUMAN-01-03.

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. 2, 5

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

[3] James Booth, Anastasios Roussos, Allan Ponniah, David Dunaway, and Stefanos Zafeiriou. Large scale 3d morphable models. *International Journal of Computer Vision*, 126(2): 233–254, 2018. 3

[4] Matthew Brand. Voice puppetry. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 21–28, 1999. 2

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

[6] 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, 3, 5

[7] 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

[8] 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. 2, 6, 8

[9] JS Chung, J Huh, S Mun, M Lee, HS Heo, S Choe, C Ham, S Jung, BJ Lee, and I Han. In defence of metric learning for speaker recognition. arxiv 2020. *arXiv preprint arXiv:2003.11982*, 2003. 6

[10] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13*, pages 87–103. Springer, 2017. 5

[11] 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. 1, 2, 3, 5, 6

[12] 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

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 4

[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. 3

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[16] 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. 5

[17] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015. 3

[18] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016. 5

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

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 3

[21] Borong Liang, Yan Pan, Zhizhi Guo, Hang Zhou, Zhibin Hong, Xiaoguang Han, Junyu Han, Jingtuo Liu, Errui Ding, and Jingdong Wang. Expressive talking head generation with granular audio-visual control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3387–3396, 2022. 2

[22] Xian Liu, Yinghao Xu, Qianyi Wu, Hang Zhou, Wayne Wu, and Bolei Zhou. Semantic-aware implicit neural audio-driven video portrait generation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*, pages 106–125. Springer, 2022. 3

[23] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018. 5

[24] Urwa Muaz, Wondong Jang, Rohun Tripathi, Santhosh Mani, Wenbin Ouyang, Ravi Teja Gadde, Baris Gecer, Sergio Elizondo, Reza Madad, and Naveen Nair. Sidgan: High-resolution dubbed video generation via shift-invariant learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7833–7842, 2023. 3, 5, 6

[25] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010. 3

[26] Foivos Paraperas Papantoniou, Panagiotis P Filntisis, Petros Maragos, and Anastasios Roussos. Neural emotion director: Speech-preserving semantic control of facial expressions in" in-the-wild" videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18781–18790, 2022. 3

[27] Se Jin Park, Minsu Kim, Joanna Hong, Jeongsoo Choi, and Yong Man Ro. Synctalkface: Talking face generation with precise lip-syncing via audio-lip memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2062–2070, 2022. 2

[28] 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. 1, 2, 3, 4, 5, 6, 7, 8

[29] Max Ritter, Uwe Meier, Jie Yang, and Alex Waibel. Face translation: A multimodal translation agent. In *AVSP'99-International Conference on Auditory-Visual Speech Processing*. Citeseer, 1999. 3

[30] 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 *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XII*, pages 666–682. Springer, 2022. 3

[31] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. *arXiv preprint arXiv:2201.02184*, 2022. 2, 3, 4, 6

[32] Bowen Shi, Wei-Ning Hsu, and Abdelrahman Mohamed. Robust self-supervised audio-visual speech recognition. *arXiv preprint arXiv:2201.01763*, 2022. 2, 3, 4

[33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[34] 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. 3

[35] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017. 2

[36] 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

[37] 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. 3

[38] Alex Waibel, Tanja Schultz, Michael Bett, Matthias Denecke, Robert Malkin, Ivica Rogina, Rainer Stiefelhagen, and Jie Yang. Smart: The smart meeting room task at isl. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, pages IV–752. IEEE, 2003. 3

[39] Alexander Waibel, Moritz Behr, Dogucan Yaman, Fevziye Irem Eyiokur, Tuan-Nam Nguyen, Carlos Mullov, Mehmet Arif Demirtas, Alperen Kantarci, Stefan Constantin, and Hazim Kemal Ekenel. Face-dubbing++: Lip-synchronous, voice preserving translation of videos. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 1–5. IEEE, 2023. 3

[40] Duomin Wang, Yu Deng, Zixin Yin, Heung-Yeung Shum, and Baoyuan Wang. Progressive disentangled representation learning for fine-grained controllable talking head synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17979–17989, 2023. 3

[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, 3, 4, 5, 6, 8

[42] Jiayu Wang, Kang Zhao, Shiwei Zhang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Lipformer: High-fidelity and generalizable talking face generation with a pre-learned facial codebook. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13844–13853, 2023. 2

[43] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 5

[44] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5

[45] Haozhe Wu, Jia Jia, Haoyu Wang, Yishun Dou, Chao Duan, and Qingshan Deng. Imitating arbitrary talking style for realistic audio-driven talking face synthesis. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1478–1486, 2021. 3

[46] 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.

[47] 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. 3

[48] Hani Yehia, Philip Rubin, and Eric Vatikiotis-Bateson. Quantitative association of vocal-tract and facial behavior. *Speech Communication*, 26(1-2):23–43, 1998. 2

[49] Fei Yin, Yong Zhang, Xiaodong Cun, Mingdeng Cao, Yanbo Fan, Xuan Wang, Qingyan Bai, Baoyuan Wu, Jue Wang, and Yujiu Yang. Styleheat: One-shot high-resolution editable talking face generation via pre-trained stylegan. In *Computer Vision–ECCV 2022: 17th European Conference, Tel*

Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII, pages 85–101. Springer, 2022. 3

[50] Fangneng Zhan, Yingchen Yu, Rongliang Wu, Jiahui Zhang, and Shijian Lu. Multimodal image synthesis and editing: A survey. *arXiv preprint arXiv:2112.13592*, 2021. 1

[51] 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. 3

[52] 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. 3, 5, 7, 8

[53] 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

[54] Rui Zhen, Wenchao Song, Qiang He, Juan Cao, Lei Shi, and Jia Luo. Human-computer interaction system: A survey of talking-head generation. *Electronics*, 12(1):218, 2023. 1

[55] 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, 8

[56] 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. 3

[57] 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

[58] 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, 3