Semi-supervised Speech-driven 3D Facial Animation via Cross-modal Encoding

Peiji Yang* Huawei Wei* Yicheng Zhong* Zhisheng Wang
Tencent, Shenzhen, China
{peijiyang, huaweiwei, ajaxzhong, plorywang}@tencent.com

Abstract

Existing Speech-driven 3D facial animation methods typically follow the supervised paradigm, involving regression from speech to 3D facial animation. This paradigm faces two major challenges: the high cost of supervision acquisition, and the ambiguity in mapping between speech and lip movements. To address these challenges, this study proposes a novel cross-modal semi-supervised framework, comprising a Speech-to-Image Transcoder and a Face-to-Geometry Regressor. The former jointly learns a common representation space from speech and image domains, enabling the transformation of speech into semantically-consistent facial images. The latter is responsible for reconstructing 3D facial meshes from the transformed images. Both modules require minimal effort to acquire the necessary training data, thereby obviating the dependence on costly supervised data. Furthermore, the joint learning scheme enables the fusion of intricate visual features into speech encoding, thereby facilitating the transformation of subtle speech variations into nuanced lip movements, ultimately enhancing the fidelity of 3D face reconstructions. Consequently, the ambiguity of the direct mapping of speech-to-animation is significantly reduced, leading to coherent and high-fidelity generation of lip motion. Extensive experiments demonstrate that our approach produces competitive results compared to supervised methods.

1. Introduction

Speech-driven 3D facial animation aims to automatically animate vivid facial expressions of realistic 3D avatars from the speech signal. It has become increasingly popular in many fields, such as games, virtual reality, film production, and online communication, as it enables the generation of lifelike facial expressions with minimal effort.

Recently proposed approaches, including VOCA[7], MeshTalk[28], and FaceFormer[10], have shown promising results in recovering 3D facial meshes from speech signals through the utilization of regressive networks or transformer-based autoregressive models. However, these methods, which fall under the supervised paradigm, encounter two primary challenges. Firstly, obtaining paired supervision of speech and high-fidelity 3D facial animation requires the utilization of a professional and expensive facial motion capture system, leading to considerable costs. Secondly, the mapping of low-dimensional speech signals to high-dimensional 3D facial meshes with significant variability may result in ambiguity issues, leading to suboptimal outcomes.

This study for the first time proposes a novel cross-modal semi-supervised framework to address the above issues, which consists of a Speech-to-Image Transcoder and a Face-to-Geometry Regressor. The former is designed to transform speech into semantically-consistent facial images, while the latter is responsible for reconstructing 3D facial meshes from the transformed images. The design of the Speech-to-Image Transcoder is inspired by the recent success of unsupervised image-to-image translation [26], which allows for transformation between different input image domains. Our transcoder extends this mechanism to facilitate cross-modal conversion between speech and image domains. To achieve this capability, we learn a common representation space using data from three domains: paired speech and real facial images, as well as synthetic facial images. We train the translation between the real and synthetic image domains in an unsupervised manner, and simultaneously leverage the paired relationship between real images and speech to project speech features into the image representation space. Through this joint training, we construct a common space where the representations of speech and images are tied together. With the learned common representation, we can convert speech into synthetic facial images using our domain-dependent image decoder. Furthermore, we leverage the synthetic facial data, which is paired with rendered facial images and 3D face meshes, to train the Face-to-Geometry Regressor. This allows us to infer the corresponding 3D face meshes from the synthetic faces generated from speech.

Our pipeline is capable of effortlessly acquiring the nec-

*These authors contributed equally to this work.

essary data, which includes synchronized speech and real facial images extracted from readily available video clips, as well as synthetic facial images generated through the deformation of a polygonal face mesh using a rendering engine. Notably, our approach is not bound by any prerequisites for paired speech and 3D face animation, thereby effectively addressing the issue of supervision data scarcity.

Moreover, our joint training scheme integrates the speech-to-image regression and image-to-image reconstruction into a cohesive framework, which enables the incorporation of detailed visual features into speech encoding. Leveraging the domain-dependent image decoders, which store intricate facial visual priors, we are able to convert nuanced speech variations into micro-expressions that are rich in detail, thereby facilitating the subsequent generation of high-fidelity 3D face reconstructions. Consequently, the uncertainties associated with the direct mapping of speech-to-animation in the supervised paradigm are considerably reduced, leading to the generation of lip motion with finer granularity.

In summary, the main contributions of our research are:

- We make the first attempt to build a semi-supervised framework for speech-driven 3D facial animation, which eliminates the need for paired speech and 3D animation.
- Our proposed joint training scheme leverages cross-modal translation to convert subtle speech variations into images that are rich in detail, resulting in more high-fidelity lip motion generation.
- Our extensive experiments and user studies demonstrate that our approach produces competitive results compared to supervised methods.

2. Related Work

The proposed method is related to several research fields, such as unsupervised image-to-image translation, speech-driven 2D talking head, and speech-driven 3D face animation. In the following, we review the most relevant approaches.

Unsupervised image-to-image translation. The unsupervised image-to-image translation aims to convert images from one domain to another domain with certain characteristics of the source images preserved[22, 36]. The translation involves converting from silhouettes to real images [16], from segmentation masks to RGB images[8, 23], from face to face [5, 27], and so on. This technology has wide applications in many fields, such as style transfer [40], face swapping [5, 26, 27], expression transfer [25], pose estimation [20] etc. They typically employ GAN-based networks [40] or encoder-decoder architectures [16] to achieve the goals. Notably, they only focus on image domains, our method extends this mechanism to achieve cross-modal translation.

Speech-driven 2D talking head. 2D-based talking head generation methods generally utilize different image representations to assist the conversion of speech to 2D faces, such as facial landmarks[4, 31, 37], depth maps[15] and semantic maps[6, 24]. The discriminator loss and other task-specific losses are usually employed for optimization. For example, Chen et al.[3] design an audio-visual derivative correlation loss to ensure cross-modal consistency. In addition, several works use 3D information such as 3D Morphable Model (3DMM)[2, 19, 30, 32, 39] parameters or dynamic neural radiance fields (NeRF)[12, 13, 35] to guide the generation process. Unlike these 2D-based methods, this paper focuses on high-fidelity 3D animation generation from speech.

Speech-driven 3D face animation. Recently, considerable endeavors have been undertaken in this field[7, 10, 11, 18, 28, 34, 38]. VOCA[7] casts the mapping from speech to animation as a regression issue. To train the regressor, it uses the paired data of speech and 3D face animation. FaceFormer[10] employs transformers to model the long-term dependencies of speech, leading to improved performance. Meshtalk[28] puts emphasis on addressing the scalability and realism of the model by decoupling audio-correlated and audio-uncorrelated information. All the above methods belong to the supervised paradigm, which requires costly paired data and suffers from the problem of ambiguous mapping when directly regressing 3D animation from speech. This paper develops a semi-supervised framework to convert speech into images and then reconstruct 3D faces from the transformed images, effectively alleviating the aforementioned two issues.

3. Methodolody

The proposed semi-supervised method is divided into two stages. We first employ the Speech-to-Image Transcoder to learn a common representation space for data from three domains: speech data, real facial images, and synthetic facial images. With this, we transform speech into semantically-consistent synthetic facial images. Subsequently, we infer the 3D meshes of the synthetic images via the Face-to-Geometry Regressor. This section first introduces preliminaries of the necessary data of three domains, then describes the two modules in detail.

3.1. Data Preliminaries

Unlike supervised methods that require paired data of speech and 3D facial animation, our framework only necessitates real facial videos synchronized with speech and a collection of rendered facial images.

We record the videos in an environment without background noise to ensure a clean recording of sound. The

camera is positioned to capture the speaker's front view, enabling clear capture of lip movements. In terms of rendered images, we utilize a readily available character for rendering, which is not required to have the same physiognomy as the speaker and can be easily acquired from online resources or produced by artists. Then, varying facial expressions are generated by deforming the character using performances of range-of-motion (ROM) and exercising poses from Facial Action Coding System (FACS)[9]. Consequently, each rendered image has its corresponding 3D facial mesh. More details of the dataset construction are presented in Section 4.1

3.2. Speech-to-Image Transcoder

Let $\mathcal{I}_{\mathcal{R}}^{1:T} = \left(\mathcal{I}_{\mathcal{R}}^{1}, \ldots, \mathcal{I}_{\mathcal{R}}^{T}\right)$ denotes a sequence of real facial frames. Correspondingly, the sequence of speech snippets $\mathcal{A}^{1:T} = \left(\mathbf{a}^{1}, \ldots, \mathbf{a}^{T}\right)$ is defined such that each $\mathbf{a}^{t}$ is synchronized with the respective $\mathcal{I}_{\mathcal{R}}^{t}$. $\{\mathcal{I}_{\mathcal{S}}\}$ is a collection of synthetic facial images, and each facial image has its corresponding 3D face mesh $\mathcal{Y}_{\mathcal{S}}$. These three types of data are utilized by the Speech-to-Image Transcoder to construct a common representation space for speech and images.

Network architecture. Our proposed Speech-to-Image Transcoder consists of two encoders and two decoders. We employ a speech encoder $\mathcal{E}_{\mathcal{A}}$ to transform speech into a compact representation. A shared visual encoder $\mathcal{E}_{\mathcal{V}}$ is used to encode real and synthetic faces. Additionally, two decoders, $\mathcal{D}_{\mathcal{R}}$ and $\mathcal{D}_{\mathcal{S}}$, are used to reconstruct real and synthetic faces from the encodings. The overview of our Speech-to-Image Transcoder is illustrated in Figure 1.

Specifically, the speech encoder $\mathcal{E}_{\mathcal{A}}$ consists of an audio feature extractor and a multi-layer transformer encoder[33]. The audio feature extractor is initialized with pre-trained wav2vec2.0[1] weights, responsible for converting raw waveform input into audio features. And the transformer encoder converts audio features into contextualized speech representations $\mathcal{Z}_{\mathcal{A}}^{1:T} = \left(\mathbf{z}_{\mathcal{A}}^{1}, \ldots, \mathbf{z}_{\mathcal{A}}^{T}\right)$. Simultaneously, real faces aligned to the audio are input to the visual encoder $\mathcal{E}_{\mathcal{V}}$ to obtain visual features $\mathcal{Z}_{\mathcal{R}}^{1:T} = \left(\mathbf{z}_{\mathcal{R}}^{1}, \ldots, \mathbf{z}_{\mathcal{R}}^{T}\right)$. This single visual encoder is shared for encoding synthetic facial features $\mathcal{Z}_{\mathcal{S}}^{\{\mathcal{B}\}} = \left(\mathbf{z}_{\mathcal{S}}^{\{\mathcal{B}\}}\right), \{\mathcal{B}\} \subseteq \{\mathcal{I}_{\mathcal{S}}\}$. Then, $\mathbf{z}_{\mathcal{R}}$ and $\mathbf{z}_{\mathcal{S}}$ are respectively fed into decoder $\mathcal{D}_{\mathcal{R}}$ and $\mathcal{D}_{\mathcal{S}}$ for reconstructing the corresponding facial images. In addition, due to the paired relationship between real images and speech, $\mathcal{D}_{\mathcal{S}}$ can also be utilized to reconstruct synthetic images from $\mathbf{z}_{\mathcal{A}}$. We borrow the decoder structure from [26], which consists of several upsampling blocks with an upsampling scale of 2. This architecture of a single shared visual encoder and multiple decoders for different domains has been verified to have good applicability in face swapping [27] and expression transfer [25]. In our experiments , $\mathbf{z}_{\mathcal{A}}$, $\mathbf{z}_{\mathcal{R}}$ and $\mathbf{z}_{\mathcal{S}}$ are all m-dim vectors, where m is set as 512.

The Speech-to-Image Transcoder learns a semantically consistent representation of speech, real faces, and synthetic faces. It enables the domain-dependent decoder $\mathcal{D}_{\mathcal{R}}$ or $\mathcal{D}_{\mathcal{S}}$ to reconstruct the facial image with corresponding semantics from any one of the three encodings.

Expression de-neutralization We observe that in cases where the two facial domains exhibit significant differences, such as one comprising real faces and the other consisting of cartoonized synthetic faces, the semantic distributions between the two domains may not be well aligned after encoding. This misalignment can result in semantically mismatched image-to-image translation. For instance, the decoder $\mathcal{D}_{\mathcal{S}}$ may erroneously decode the encoding of a real face exhibiting a grinning expression as a synthetic face with pouting expression.

To address this issue, we propose a simple yet effective operation that we refer to as expression de-neutralization (denoted in Equation 1). Rather than reconstructing the facial expression image directly from the encoding $\mathbf{z}_i, i \in \{\mathcal{A}, \mathcal{R}, \mathcal{S}\}$ we first subtract the neutral expression code $\overline{z}_i$ corresponding to the domain (for speech domain, we use $\overline{z}_{\mathcal{R}}$) and then recover the image with the offset expression code Δz_i. We manually select a neutral expression image for each facial domain and obtain the neutral expression code after encoding. Through expression de-neutralization, two semantically misaligned domains can be shifted to become aligned, leading to semantically-consistent translation. Our experiments profoundly verify the effectiveness of this operation.

$$\Delta z_i = z_i - \overline{z}_i. \tag{1}$$

Training objectives. We employ two commonly used loss terms for image reconstruction, combined in Equation 2. The first term denotes the mean error of all pixels between the input image and the reconstructed one ($\widetilde{\mathcal{I}}_{\mathcal{R}}$ or $\widetilde{\mathcal{I}}_{\mathcal{S}}$). The second term $\mathcal{L}_{percep}$ is the perceptual loss[17] to measure the similarity of two different images. It has obvious benefits for enhancing the fidelity of the reconstructed image.

$$\mathcal{L}_{rec} = ||\mathcal{D}_i(\Delta z_i) - \mathcal{I}_i||_2 + \mathcal{L}_{percep}, i \in \{\mathcal{R}, \mathcal{S}\} \tag{2}$$

Moreover, owing to the fact that the speech and real facial images are inherently paired, we are able to enhance temporal synchronization and consistency between speech and visual features by imposing our designed cross-modal constraint (Equation 3). This constraint is comprised of two terms, with $\mathcal{L}_{cf}$ representing the cross-modal constraint on features, and $\mathcal{L}_{cs}$ representing that on image spatial space.

$$\begin{aligned} \mathcal{L}_{cross} &= \mathcal{L}_{cf} + \mathcal{L}_{cs}, \\ \mathcal{L}_{cf} &= ||z_{\mathcal{A}} - z_{\mathcal{R}}||_2, \\ \mathcal{L}_{cs} &= ||\mathcal{D}_{\mathcal{R}}(\Delta z_{\mathcal{A}}) - \mathcal{I}_{\mathcal{R}}||_2. \end{aligned} \tag{3}$$

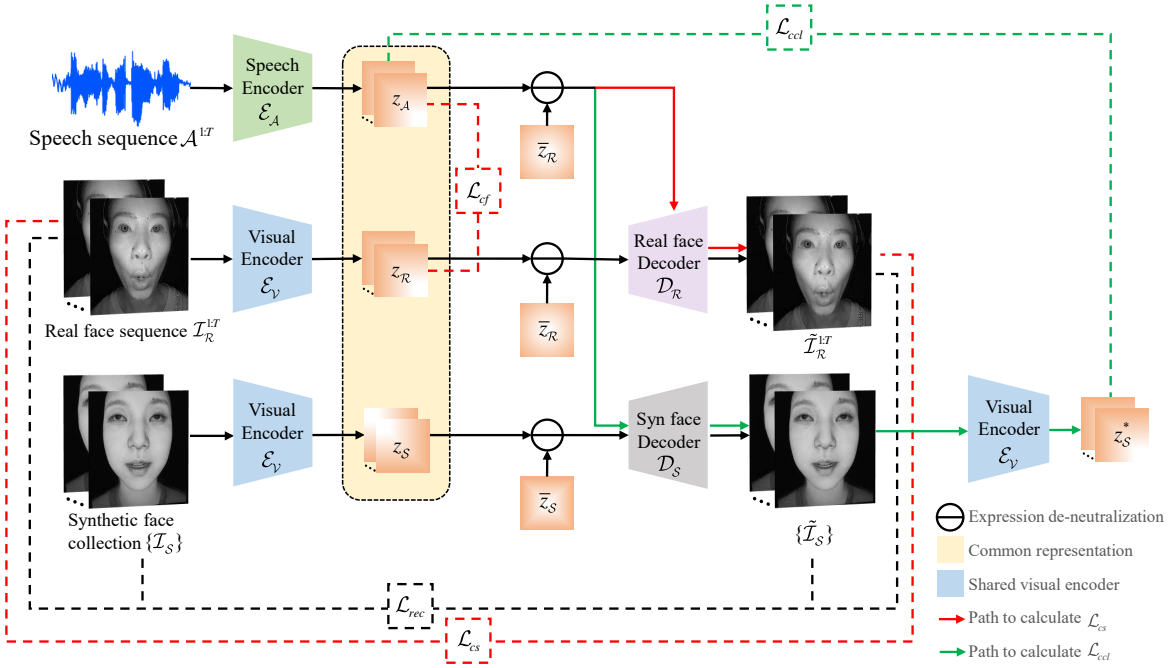

Figure 1. The architecture of the Speech-to-Image Transcoder. Firstly, the speech is converted into $\mathbf{z}_{\mathcal{A}}$ through $\mathcal{E}_{\mathcal{A}}$. The red arrow shows how we input $\mathbf{z}_{\mathcal{A}}$ into $\mathcal{D}_{\mathcal{R}}$ to obtain the reconstructed image for calculating $\mathcal{L}_{cs}$. The green arrow shows how we obtain $z^*_{\mathcal{S}}$ for calculating $\mathcal{L}_{ccl}$. We feed the Real/Synthetic face images into the $\mathcal{E}_{\mathcal{V}}$ to obtain the visual representation $\mathbf{z}_{\mathcal{R}}$ and $\mathbf{z}_{\mathcal{S}}$, which are respectively input into $\mathcal{D}_{\mathcal{R}}$ and $\mathcal{D}_{\mathcal{S}}$ to obtain reconstructed images for calculating the $\mathcal{L}_{rec}$.

To further encourage the transcoder to use a common representation across domains of speech and synthetic faces, we present a latent cycle consistency loss (Equation 4). This loss creates a loop constraint between the latent codes $z_{\mathcal{A}}$ derived from speech input and the latent codes $z^*_{\mathcal{S}}$ derived from the decoded synthetic images.

$$\mathcal{L}_{ccl} = ||z^*_{\mathcal{S}} - z_{\mathcal{A}}||_2,$$
$$z^*_{\mathcal{S}} = \mathcal{E}_{\mathcal{V}}(\mathcal{D}_{\mathcal{S}}(\Delta z_{\mathcal{A}})). \quad (4)$$

To sum up, our total loss is depicted in Equation 5.

$$\mathcal{L}_{total} = \mathcal{L}_{rec} + \mathcal{L}_{cross} + \mathcal{L}_{ccl}. \quad (5)$$

By utilizing the proposed cross-modal joint training, encodings from three distinct domains are semantically aligned. This allows for any of the three input modalities to be transformed into semantically-consistent facial images through domain-specific decoders. Specifically, we convert speech into synthetic faces. By further leveraging the Face-to-Geometry Regressor in Section 3.3, 3D facial meshes can be reconstructed from the transformed images.

3.3. Face-to-Geometry Regressor

The architecture of the Face-to-geometry Regressor $\mathcal{E}_{\mathcal{G}}$ is shown in Figure 2. It is trained in a supervised manner. We exclusively rely on information from the synthetic

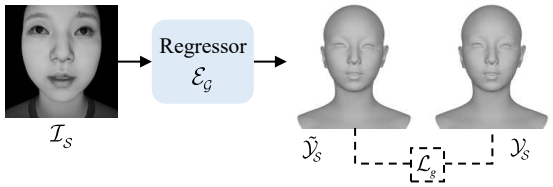

Figure 2. The architecture of the Face-to-Geometry Regressor. The synthetic face images are fed into the Regressor $\mathcal{E}_{\mathcal{G}}$, which learns to regress the 3D mesh associated with the synthetic face images using RMSE loss.

image domain, specifically synthetic facial images and the corresponding 3D face meshes. We utilize Resnet34[14] as the backbone and a single fully connected layer to regress the coordinates of the 3D mesh $\widetilde{\mathcal{Y}}_S$. A simple RMSE loss is used to optimize the network, denoted as follows:

$$\mathcal{L}_g = ||\mathcal{E}_{\mathcal{G}}(\mathcal{I}_S) - \mathcal{Y}_S||_2. \quad (6)$$

During the training phase, we employ data augmentation of color jittering and random affine transformation to enhance the robustness of our model. As a consequence, we are able to achieve satisfactory 3D face reconstruction results, even when the synthetic faces generated by the Speech-to-Image Transcoder are outside the training set.

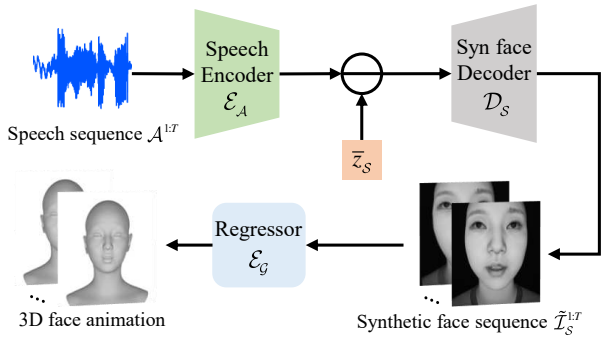

Figure 3. Inference pipeline. The raw waveform is fed to the speech encoder $\mathcal{E}_\mathcal{A}$. The output is subtracted by $\overline{z}_\mathcal{R}$ and then fed into $\mathcal{D}_\mathcal{S}$ to recover the synthetic facial images, which are then input into $\mathcal{E}_\mathcal{G}$ to generate the 3D face animation.

3.4. Inference Phase

As illustrated in Figure 3, the Speech-to-Image Transcoder and Face-to-Geometry Regressor are integrated to infer 3D face animation from speech. The raw waveform is fed into the speech encoder $\mathcal{E}_\mathcal{A}$ to generate the speech representation. Then we subtract the speech representation by $\overline{z}_\mathcal{R}$, and the offset code is fed into $\mathcal{D}_\mathcal{S}$ to recover the synthetic facial images. Due to the cross-modal constraint $\mathcal{L}_{cross}$, the transformed synthetic facial images are endowed with lip movements reflecting the content of the input speech. Finally, the face-to-geometry model is utilized to generate the 3D face animation from the transformed synthetic facial images.

3.5. Discussion

One might raise doubts about the rationale behind not utilizing the image-to-image modules to accomplish the transformation from real faces to synthetic faces, subsequently employing the Face-to-Geometry Regressor to acquire pseudo-labels, and ultimately training a speech-to-pseudo-label network in a supervised manner. (Hereafter, we refer to this pipeline as the three-stage method.) We notice that while this approach can alleviate the challenge of limited supervision data, it still suffers from the problem of ambiguous mapping between speech and animation, the same as other regression-based supervised methods.

Our cross-modal joint training scheme can identify latent patterns and correlations between speech and images. This allows us to translate subtle variations in speech into detailed lip motions, creating highly-fidelity 3D reconstructions. Therefore, our approach has a distinct advantage in mitigating the uncertainties of direct speech-to-animation mapping. We conduct detailed experiments to demonstrate the advantages of our method compared with the three-stage method in Section 4.

4. Experiments

4.1. Datasets

Our experiments are conducted on two datasets: the self-built dataset SelfCollected and the public dataset VOCASET[7]. In addition, to support the semi-supervised cross-modal training, we generate two synthetic datasets. The details are described below.

SelfCollected dataset. This dataset contains 148 speaking video clips of one speaker. Each video contains one sentence. We employ the 3D facial capture system's workflow to obtain the ground truth (GT). Specifically, a head-mounted facial capture system is used to record the videos so that detailed lip motions of the speaker can be accurately tracked. Then, the facial tracking module captures several key facial landmarks. For each video, the artists manually pair a real person's face with the CG character at several keyframes. Finally, the facial retargeting module is trained on these keyframes and then automatically directs the motions of other video frames onto the CG character. Note that the character has not the same physiognomy as the speaker, so the range of lip movements is different after face retargeting. Through the workflows, we obtain the corresponding 3D face mesh with the topology of the CG character for each video frame. In total, this dataset contains 45000 frames, each frame corresponding a face mesh with 12696 vertices. We take 128 videos for training and the remaining 20 videos for testing.

VOCASET dataset. VOCASET contains 12 speakers, each with 40 utterances. Facial meshes are scanned at 60FPS and then registered into a unified topology using the FLAME model[21], with each mesh containing 5023 vertices. In our experiments, we only use the data of one speaker for training and testing. We use 36 utterances as the training set and the remaining 4 as the test set.

SyntheticSelf dataset. This dataset is built to assist the semi-supervised training on SelfCollected. We use the same CG character in SelfCollected for rendering. During rendering, we keep the camera facing the character, and the rendering lighting remains unchanged. Animations of performances of range-of-motion (ROM) and exercising poses from the Facial Action Coding System (FACS) are utilized for rendering. Specifically, we require artists manually craft the blendshape weights of ROM and FACS by linearly combining the blendshapes of the CG character, enabling us to achieve a wide range of expressions. Finally, SyntheticSelf comprises 20000 rendered facial images, each corresponding to a facial mesh with 12696 vertices.

SyntheticVOCA dataset. Similar to SyntheticSelf, we create SyntheticVOCA to assist the semi-supervised training on VOCASET. We use FLAME model to fit expression coefficients for each rendered image in SyntheticSelf. Then expressions are transferred to the character (i.e. the

speaker's scan) provided by VOCASET. In total, we obtain 20000 rendered facial images, each corresponding to a FLAME-topology facial mesh with 5023 vertices.

4.2. Compared methods

We compare our method with several SOTA methods: VOCA[7], MeshTalk[28] and FaceFormer[10], as well as the three-stage method discussed in Section 3.5. All methods are compared on SelfCollected and VOCASET. Notably, since MeshTalk cannot be trained on a small single-speaker dataset, we only perform qualitative comparisons with its officially released model. In addition, we find VOCA cannot converge on a single-speaker dataset, so we use all the data of the other 11 speakers combined with the single speaker's 36 sentences for training. The three-stage method employs an image-to-image and face-to-geometry pipeline to obtain pseudo-labels and then experiments with network architectures of VOCA and Faceformer, respectively. Since VOCASET does not have videos of real faces, leading to that we cannot obtain the corresponding pseudo-labels, so we do not compare with the three-stage method on VOCASET. We use the symbol * later to represent the corresponding method using the three-stage scheme.

To sum up, there are **7** comparisons on SelfCollected: ours, VOCA, FaceFormer, GT, VOCA*, FaceFormer*, and pseudo GT. And **4** comparisons on SelfCollected: ours, VOCA, FaceFormer, and GT.

4.3. Quantitative Evaluation

We employ the commonly used lip vertex error[28] to evaluate the quantitative performance, which represents the maximal l_2 error of vertices in the lip region. In addition, we propose the lip landmark error to measure the lip movement difference between the generated 3D animation and real facial sequence. To obtain this error, we first estimate a similarity transformation based on the projected 2D coordinates of the predicted 3D lip vertices and the 2D lip landmarks of the real face, then align them using the estimated transformation, and finally compute the root mean squared error.

Results are illustrated in Table 1 and Table 2. Note that GT is obtained by retargeting the lip motions from the real speaker to the CG character, so it reflects the idiosyncrasies of the character instead of the speaker. Therefore, it is not objective to compare us with the supervised methods in terms of the metric of lip vertex error, and similarly, there is no comparability between supervised methods and ours in terms of the metric of lip landmark error. Nonetheless, our semi-supervised method achieves comparable performance to supervised methods in terms of lip vertex error, and even outperforms VOCA on SelfCollected. As for the lip landmark error, our method significantly outperforms the supervised approach and consistently outperforms the three-stage methods, even approaching the pseudo GT, which is the upper bound of the three-stage methods. This indicates that our joint training approach indeed promotes generating accurate and higher fidelity lip animations.

Methods	**Lip vertex error** (in mm)↓	**Lip landmark error** (in pixel)↓
VOCA	3.718	5.908
FaceFormer	**2.690**	5.845
GT	**0**	5.041
VOCA*	5.838	3.790
FaceFormer*	4.536	3.613
Pseudo GT	3.773	**2.941**
Ours	3.708	**3.241**

Table 1. Quantitative results on SelfCollected.

Methods	**Lip vertex error** (in mm)↓
VOCA	4.690
FaceFormer	**3.166**
Ours	5.102

Table 2. Quantitative results on VOCASET.

4.4. Qualitative Evaluation

Figure 4 presents qualitative results obtained by comparing three sentences. Sentences A and B come from SelfCollected-Test, and sentence C comes from VOCASET-Test. They belong to two different languages, namely Chinese and English. We evaluate sentence A with the model trained on SelfCollected and evaluate sentences B, C using the model trained on VOCASET. It should be noted that sentence B does not have its FLAME-topology GT meshes, and sentence C does not have the corresponding real faces.

From the results of sentence A, we can see that our method achieves competitive results compared to the supervised approaches, and even gains better performance in some cases (highlighted by the red box) in terms of mouth closure and lip movement amplitude. From the results of sentence C, our method achieves results that are highly synchronized with the speech even without using speech data from VOCASET for training. The results of sentence A also demonstrate that compared to the three-stage methods, our method presents results that are more consistent with the lip movements of the real face. This indicates our joint training scheme helps to generate more realistic animations.

4.5. Perceptual Evaluation

We conduct user studies to evaluate the animation quality of our method, which comprise a total of 10 participants.

User Study on SelfCollected. For SelfCollected, we obtain the results of 7 comparisons using 10 random test sen-

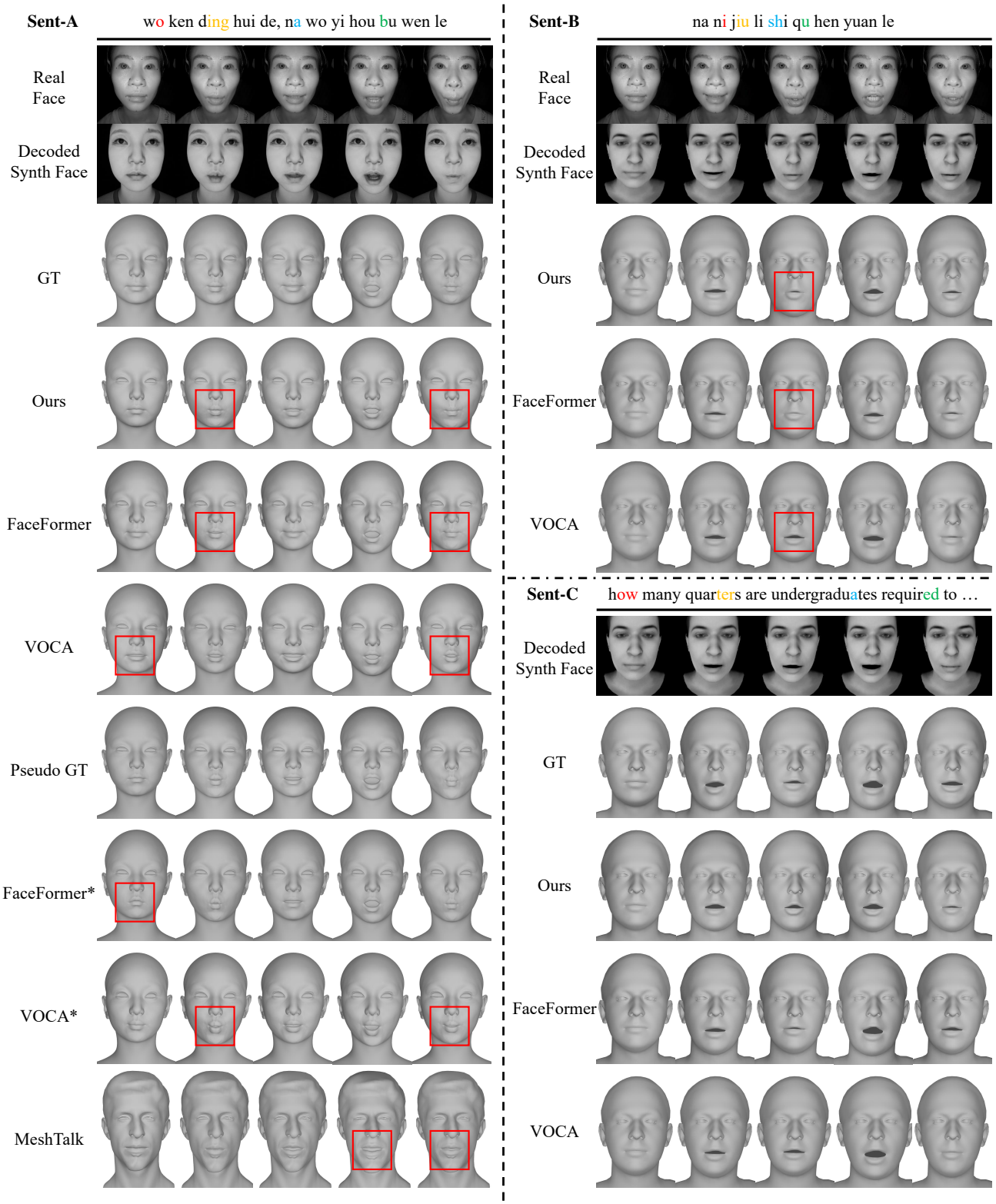

Figure 4. Qualitative results. We utilize the officially released model to produce MeshTalk results as it is not trainable on SelfCollected and VOCASET datasets. The highlighted section by red box indicates that our method has better performance in terms of mouth closure and lip movement amplitude.

Ours vs. Competitor	Realism↑	Lip Sync↑	Sim↑
Ours vs. VOCA	51.0	53.0	63.0
Ours vs. FaceFormer	45.0	49.0	57.0
Ours vs. GT	43.0	45.0	58.0
Ours vs. VOCA*	75.0	64.0	60.0
Ours vs. FaceFormer*	58.0	61.0	54.0
Ours vs. pesudo GT	52.0	51.0	48.0

Table 3. User study results on SelfCollected. We use A/B testing and report the percentage (%) of answers where A is preferred over B.

Ours vs. Competitor	Realism↑		
	VOCASET	Self.	Avg.
Ours vs. GT	17.5	—	—
Ours vs. FaceFormer	40.0	57.5	48.8
Ours vs. VOCA	45.0	95.0	70.0
Ours vs. Competitor	**Lip Sync↑**		
	VOCASET	Self.	Avg.
Ours vs. GT	5.0	—	—
Ours vs. FaceFormer	15.0	57.5	36.3
Ours vs. VOCA	60.0	80.0	70.0

Table 4. User study results (%) on VOCASET.

tences of the test set. Therefore, 60 A vs. B pairs(10 videos x 6 comparisons) are created. For each pair, we ask the participants to choose their favorite one in terms of realism and lip sync, following the user study protocol of Faceformer. Additionally, we propose a new metric "similarity" to characterize the preference for which method generates more similar lip movements to that of the real faces. For this metric, we provide both A/B pair and the real facial frames for each sentence. And then we ask the participants to choose the one that has closer lip amplitude to the real face. Table 3 shows the results. We observe that our method is comparable to the SOTA supervised methods in terms of realism and lip sync. And our method significantly outperforms the supervised approach in terms of similarity. Furthermore, the results show that our method consistently performs better than the three-stage methods. It indicates that our joint training of speech and image is indeed effective.

User Study on VOCASET. We conduct another user study on VOCASET. Since our model has never seen the speech data in VOCASET, for fairness, we additionally randomly pick 4 sentences for comparison from SelfCollected-Test. In total, 20 A vs. B pairs (there is no FLAME-topology GT in SelfCollected) are created. For each pair, participants are asked to choose their favorite one based on realism and lip sync. As shown in Table 4, we separately report the results of SelfCollected-Test and VOCASET-Test. The results show that our method consistently performs better in terms of realism. However, our model achieves worse lip sync than FaceFormer on VOCASET-Test. We claim that it is reasonable because our model is not trained with the English speech data of VOCASET. Similarly, the Faceformer's result on SelfCollected-Test is worse than ours.

Expression de-neutralization	$\mathcal{L}_{ccl}$	**Lip lmk error** (in pixel)↓
✓	—	3.407
—	✓	3.482
✓	✓	**3.241**

Table 5. Ablation study on SelfCollected.

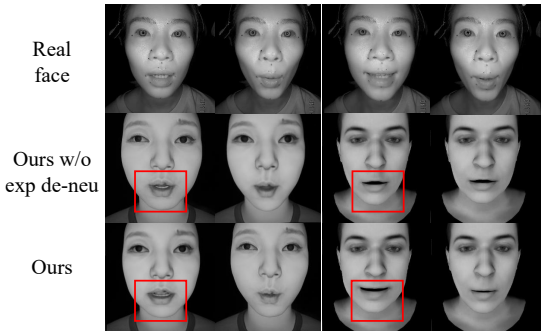

Figure 5. Ablation study on expression de-neutralization.

4.6. Ablation Study

Efffect of Expression De-neutralization . We investigate the effect of expression de-neutralization by removing it from our method. We conduct the comparison on SelfCollected. As depicted in Table 5, the operation of expression de-neutralization helps to reduce lip landmark error by 0.241 (3.482 → 3.241). We also visualize the decoded synthetic faces in Figure 5. As depicted in Figure 5, the incorporation of expression de-neutralization results in more consistent lip motions of synthetic faces with that of real faces. Without this operation, the decoded faces demonstrate obvious semantic mismatches, particularly when the disparity between the two image domains is substantial (i.e. real faces and rendered FLAME-topology faces). This phenomenon proves that expression de-neutralization helps align semantic distributions of different domains, resulting in more semantically consistent lip movement generation.

Efffect of Cycle Consistency Loss $\mathcal{L}_{ccl}$. We present the quantitative results in Table 5. As shown, $\mathcal{L}_{ccl}$ helps to reduce lip landmark error by 0.166 (3.407 → 3.241). The qualitative comparison is depicted in Figure 6, it is indicated this loss function promotes the generation of lip movements that are more consistent with the speech.

5. Conclusion

This paper aims to recover 3D animation from speech, it proposes a novel semi-supervised framework to address the

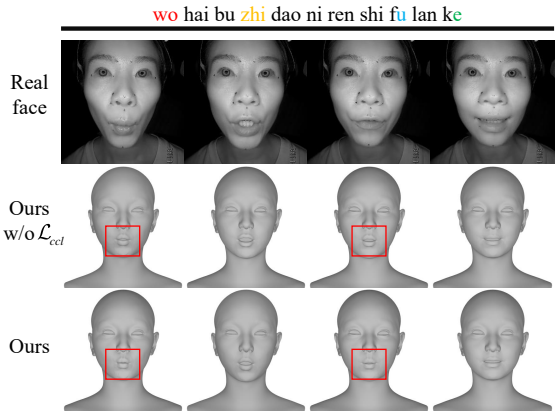

Figure 6. Ablation study on $\mathcal{L}_{ccl}$.

data scarcity problem of the supervised methods. In addition, a joint training scheme on cross-domain data, namely speech and images, is developed. Extensive experiments show the proposed method achieves satisfactory generation of lip motions. However, the major limitation of our method is its inability for multiple identities. To address this issue, we suggest two possible ways for future work. The first is expanding the decoder for each speaker, which succeeds in faceswap but increases model size with more IDs. The second is inspired by FOMM [29], which uses a universal motion module for facial motion extraction and transfer. Overall, our proposed method shows promising results and opens up opportunities for further research in the field of speech-driven 3D facial animation.

References

[1] 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.

[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] 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.

[4] 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.

[5] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. Simswap: An efficient framework for high fidelity face swapping. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2003–2011, 2020.

[6] 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.

[7] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J. Black. Capture, learning, and synthesis of 3d speaking styles. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10093–10103, 2019.

[8] Ricard Durall Lopez, Jireh Jam, Dominik Strassel, Moi Hoon Yap, and Janis Keuper. Facialgan: Style transfer and attribute manipulation on synthetic faces. In *[32nd British Machine Vision Conference]*, pages 1–14, 2021.

[9] Paul Ekman and Wallace V Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978.

[10] Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. Faceformer: Speech-driven 3d facial animation with transformers. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18749–18758, 2021.

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

[12] 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.

[13] 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.

[14] 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.

[15] Fa-Ting Hong, Longhao Zhang, Li Shen, and Dan Xu. Depth-aware generative adversarial network for talking head video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3397–3406, 2022.

[16] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[17] 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.

[18] Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)*, 36(4):1–12, 2017.

[19] 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.

[20] Siyuan Li, Semih Gunel, Mirela Ostrek, Pavan Ramdya, Pascal Fua, and Helge Rhodin. Deformation-aware unpaired image translation for pose estimation on laboratory animals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13158–13168, 2020.

[21] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017.

[22] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. *Advances in neural information processing systems*, 30, 2017.

[23] Wenshuang Liu, Wenting Chen, Zhanjia Yang, and Linlin Shen. Translate the facial regions you like using self-adaptive region translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2180–2188, 2021.

[24] 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.

[25] Lucio Moser, Chinyu Chien, Mark Williams, Jose Serra, Darren Hendler, and Doug Roble. Semi-supervised video-driven facial animation transfer for production. *ACM Transactions on Graphics (TOG)*, 40(6):1–18, 2021.

[26] Jacek Naruniec, Leonhard Helminger, Christopher Schroers, and Romann M Weber. High-resolution neural face swapping for visual effects. In *Computer Graphics Forum*, volume 39, pages 173–184. Wiley Online Library, 2020.

[27] Ivan Perov, Daiheng Gao, Nikolay Chervoniy, Kunlin Liu, Sugasa Marangonda, Chris Umé, Mr Dpfks, Carl Shift Facenheim, Luis RP, Jian Jiang, et al. Deepfacelab: Integrated, flexible and extensible face-swapping framework. *arXiv preprint arXiv:2005.05535*, 2020.

[28] Alexander Richard, Michael Zollhoefer, Yandong Wen, Fernando De la Torre, and Yaser Sheikh. Meshtalk: 3d face animation from speech using cross-modality disentanglement. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1153–1162, 2021.

[29] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in neural information processing systems*, 32, 2019.

[30] 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.

[31] 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.

[32] 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.

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

[34] Chih-Chun Yang, Wan-Cyuan Fan, Cheng-Fu Yang, and Yu-Chiang Frank Wang. Cross-modal mutual learning for audio-visual speech recognition and manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3036–3044, 2022.

[35] 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.

[36] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE international conference on computer vision*, pages 2849–2857, 2017.

[37] Lingyun Yu, Jun Yu, and Qiang Ling. Mining audio, text and visual information for talking face generation. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 787–795. IEEE, 2019.

[38] Sibo Zhang, Jiahong Yuan, Miao Liao, and Liangjun Zhang. Text2video: Text-driven talking-head video synthesis with personalized phoneme-pose dictionary. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2659–2663. IEEE, 2022.

[39] 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.

[40] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.