Speech4Mesh: Speech-Assisted Monocular 3D Facial Reconstruction for Speech-Driven 3D Facial Animation

Shan He[1,2], Haonan He[2], Shuo Yang[2], Xiaoyan Wu[2], Pengcheng Xia[2],
Bing Yin[2], Cong Liu[2], Lirong Dai[1], Chang Xu[3]
[1]University of Science and Technology of China, [2]iFLYTEK Research, [3]University of Sydney
{shanhe2, hnhe, shuoyang7, xywu10, pcxia, bingyin, congliu2}@iflytek.com
lrdai@ustc.edu.cn, c.xu@sydney.edu.au

Abstract

Recent audio2mesh-based methods have shown promising prospects for speech-driven 3D facial animation tasks. However, some intractable challenges are urgent to be settled. For example, the data-scarcity problem is intrinsically inevitable due to the difficulty of 4D data collection. Besides, current methods generally lack controllability on the animated face. To this end, we propose a novel framework named Speech4Mesh to consecutively generate 4D talking head data and train the audio2mesh network with the reconstructed meshes. In our framework, we first reconstruct the 4D talking head sequence based on the monocular videos. For precise capture of the talking-related variation on the face, we exploit the audio-visual alignment information from the video by employing a contrastive learning scheme. We next can train the audio2mesh network (e.g., FaceFormer) based on the generated 4D data. To get control of the animated talking face, we encode the speaking-unrelated factors (e.g., emotion, etc.) into an emotion embedding for manipulation. Finally, a differentiable renderer guarantees more accurate photometric details of the reconstruction and animation results. Empirical experiments demonstrate that the Speech4Mesh framework can not only outperform state-of-the-art reconstruction methods, especially on the lower-face part but also achieve better animation performance both perceptually and objectively after pre-trained on the synthesized data. Besides, we also verify that the proposed framework is able to explicitly control the emotion of the animated talking face.

1. Introduction

Speech-driven facial animation is aimed at guiding the face model (either 2D image or 3D mesh) to have perceptually rational motion (especially for lip sync) only according to the input audio signal. It is drawing increasingly more attention with the development of virtual reality, game and film production, etc. There are currently two mainstreams to deal with this task, either from 2D or 3D perspectives. By learning from video data [9, 41, 55, 71], 2D solutions try to generate the talking head images frame by frame. Although sufficient data is available to be used, the synthesized images cannot directly be applied to 3D scenes. Thus, in this paper, we mainly focus on 3D facial animation driven by speech. In recent years, a handful of 3D speech-driven facial animation methods [20, 47, 39, 13, 7, 34] have shown promising performance by directly learning a mapping from audio to 3D avatar mesh deformation with deep neural networks (hereinafter called audio2mesh methods). In this way, some subtle variations on the face mesh can be captured. Generally speaking, for the audio2mesh methods, data quality is one of the most important factors to be considered [20] since they generally demand high-precision

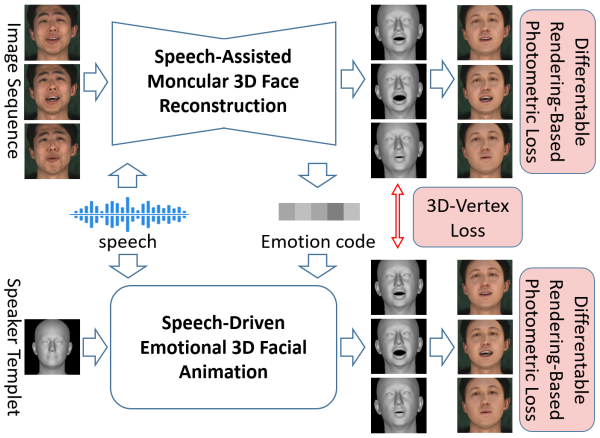

Figure 1: Overview of the proposed Speech4Mesh framework. In this framework, the speech signal is leveraged for both the reconstruction module and the audio2mesh module. An emotion code is embedded into the audio2mesh module for emotional manipulation. Differentiable rendering is employed for recovering the details of the face.

4D talking head data to learn the correspondence from audio embedding to mesh movement in a supervised manner. However, in practice, real 4D data is extremely hard to collect. It is generally reconstructed and registered from synchronized multi-view images or 3D scan frame by frame, which is extraordinarily costly [13, 21]. Therefore, how to deal with the data scarcity problem becomes an urgent agenda to be discussed.

In contrast, the accessibility of monocular videos is much higher. The scarcity of 4D data may be mitigated through the use of 3D reconstruction techniques applied to 2D videos. The majority of current monocular 3D face reconstruction methods (e.g., 3DDFA-V2 [27], DECA [22]) are based on the self-supervised training scheme, whose essential guidances are landmark loss and photometric loss. Nonetheless, these losses can only provide 2D information from the image plane, which is relatively weak to capture the complex nonrigid deformation of the lower face region [51, 62]. Therefore, the reconstructed results always demonstrate obvious artifacts and ambiguities between utterance and visual perception, such as mouth funnel, pucker, and rounded vowels, which may considerably limit the precision of the audio2mesh methods [20].

To address the problems mentioned above, we propose a novel framework named Speech4Mesh to consecutively synthesize 4D talking head data and train the audio2mesh network with the reconstructed mesh sequences. An overview of the Speech4Mesh framework is illustrated in Fig. 1. Firstly, to obtain better synthesis results, we propose a novel speech-assisted monocular 3D face reconstruction module to exploit multi-modal correlation from the video data rather than merely from the RGB image, since the mouth geometry and expression are naturally correlated to the speech. Specifically, this module is built based on a backbone reconstruction model DECA [22], one of the state-of-the-art monocular 3D face reconstruction methods. By employing a speech-visual contrastive loss, it can learn the correspondence between audio and visemes to make the reconstructed 3D face to be more reliable and perceptually reasonable.

Secondly, once sufficient 4D data has been acquired, we can train an audio2mesh network to model the deformation of a face mesh based on input speech. For our work, we employed FaceFormer [20] as the backbone for the audio2mesh network. Compared with the vanilla training scheme of FaceFormer, the primary benefit of Speech4Mesh is data sufficiency. In fact, Speech4Mesh provides a low-cost self-supervised approach to generate "infinite" 4D training data from infinite RGB videos. Besides, owing to the diversity of the talking head videos, the reconstructed data empowers our model to further disentangle other auxiliary properties of the face (e.g., emotion, etc.) for manipulation. To be specific, we first pre-train FaceFormer on the reconstructed emotional 4D dataset with the corresponding emotion codes. At the fine-tuning stage, different emotion codes can be embedded into the FaceFormer backbone to realize the emotion control of the animated talking head.

The main contributions of our work can be summarized as follows:

- We propose Speech4Mesh, a novel framework that tackles the data scarcity issue in speech-driven 3D facial animation by consecutively reconstructing pseudo-4D data and training an audio2mesh network. This framework also provides a referable scheme to alleviate the data scarcity problem of similar 4D tasks.
- This framework enables emotional control in the absence of high-quality expressive 4D data by pre-training with the emotional pseudo-4D talking head data reconstructed from 2D videos. And the synthesized emotional 4D talking head dataset can be found in https://github.com/haonanhe/MEAD-3D/tree/main.
- To the best of our knowledge, we are the first to utilize speech information as complement for monocular 3D face reconstruction. Empirical experiments show that the exploitation of multi-modal information yields more natural and precise results.

2. Related Work

Speech-driven 3D Facial Animation: Speech-driven facial animation has been widely explored over the years. Existing methods could be categorized into 2D-based approaches which attempt to operate in 2D pixel space [72, 70, 45, 15, 11, 8], and 3D-based approaches that we focus on. Some works animate 3D models based on visemes through complex rules of coarticulation [19, 73, 68, 57, 12]. Though they could capture accurate lip movements and achieve explicit control over the entire process, such methods usually involve plenty of manual adjustments. Data-driven approaches have been proposed to automatically learn a map from acoustic features to 3D models. Most of them were trained in a supervised manner [20, 13, 47, 6, 34] with multi-view motion capture data [4, 6] or high-resolution 4D scans [34, 13]. [6] searches the matching mouth animation for phonemes based on an Anime Graph structure. [34] proposes an end-to-end convolutional network and uses a latent code to control the emotional state of animated mesh. VOCA [13] achieves speaker-independent facial animation by leveraging a subject-specific latent code. MeshTalk [47] learns a categorical latent space to disentangle audio-correlated and audio-uncorrelated information, which enables it to animate the entire face. FaceFormer [20] first applies Transformer to 3D facial animation to make an autore-

gressive prediction of the animated 3D face mesh sequence, which claims to be able to capture accurate lip movements.

However, 4D scans are usually too expensive to access, which leads to the data scarcity issue. There are methods trying to address this problem by using 4D data generated from videos [56, 65, 44]. [44] trains an LSTM-RNN model to regress the shape parameter of the parametric face model tracked from input video frames. [56] uses a sliding window predictor to learn mappings from phoneme sequences to lower facial shape and appearance that an Active Appearance Model (AAM) tracks from video frames. [65] aims to synthesize 3D talking-head with rich emotion. However, their 3D faces tracked from videos are only optimized by 2D landmarks loss, which may lead to inaccurate reconstruction and animation results. In contrast, our method leverages speech features to capture 3D faces with more accurate facial movements. In addition, pre-training on emotion datasets enables our reconstructed 3D face meshes to have emotion states, which leads to 3D facial animation with rich emotion.

Monocular 3D face reconstruction: Monocular 3D face reconstruction aims to track and recover the 3D shape and texture of the human face from monocular 2D data. Early methods often perform an optimization procedure in an analysis-by-synthesis framework, predicting landmarks[75, 62, 30, 5] or local features[29, 50] to regress parameters of 3D Morphable Models (3DMM) [2, 43, 3, 35]. Although optimization-based approaches could recover high-quality 3D faces, they usually involve costly optimization processes. In recent years, learning-based methods were proposed for more efficient reconstruction. Some of them predict 3DMM coefficients [32, 74, 48, 37, 49] directly. To alleviate constraints caused by models' limited geometry, there are also model-free methods directly regressing 3D meshes [16, 18, 23, 33, 53], voxels [31] or Signed Distance Function (SDF) [42].

However, the greatest limitation of learning-based methods is the lack of paired training data. Some methods address this problem by generating synthetic data or using image-model pairs fitted by traditional optimization-based methods [63, 66, 28, 38, 74, 48, 53]. Others have tried to solve it in an unsupervised manner. Most of unsupervised methods try to regress 3DMM parameters by predicting 2D landmarks [17, 22, 40, 52, 54, 58, 59, 60, 69], photometric constraints [17, 22, 25, 54, 59, 60, 69] or multi-image constraints[17, 22, 25, 52, 58]. Perceptual constraints such as face recognition features [17, 22, 25, 54] or emotion features [14] are also utilized to constrain models' learning objects. While unsupervised methods provide a large amount of training data, reconstructing 3D faces from merely 2D information is still underconstrained. To capture more accurate expressions from monocular videos, our framework leverages speech information as another source of constraints, which has a strong relationship with facial movements, especially the mouth area.

3. Preliminaries

Current monocular 3D face reconstruction methods generally leverage the self-supervised scheme for training. In such a way, a parametric face model and a differentiable renderer are required to recover the 3D face mesh and render the 3D face back to the image for training, respectively. In this section, we review the basic knowledge of the parametric face model and differentiable renderer for reference.

3.1. FLAME

In this work, we used FLAME [36] as the prior parametric face model for reconstruction, which is a generic 3D head model using standard vertex-based linear blend skinning (LBS) with corrective blendshapes. Specifically, it consists of $N = 5023$ vertices and could be defined as:

$$M(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi}) \in \mathbb{R}^{3N}, \tag{1}$$

which contains shape parameters $\boldsymbol{\beta} \in \mathbb{R}^{|\beta|}$ accounting for shape variation, expression $\boldsymbol{\psi} \in \mathbb{R}^{|\psi|}$ to capture facial expressions, and pose parameters $\boldsymbol{\theta} \in \mathbb{R}^{3K+3}$ to rotate vertices around $K = 4$ joints (neck, jaw, and eyeballs) as well as correcting pose deformations. FLAME provides a geometric constraint and a low-dimensional latent space, making our reconstruction process much easier. Besides, highly disentangled parameters allow us to control the face model's expression or pose independently, which benefits us in both the reconstruction and animation phases.

We obtain FLAME's texture map through an appearance model $A(\boldsymbol{\alpha}) \in \mathbb{R}^{d \times d \times 3}$ with albedo parameters $\boldsymbol{\alpha} \in \mathbb{R}^{|\alpha|}$ as input. This is achieved by converting Basel Face Model's linear albedo space to FLAME's UV layout. More details can be found in [35].

3.2. Differentiable Rendering

3D rendering is a function that maps the 3D object to a 2D image. Differentiable rendering, as the name implies, provides a differentiable rendering function where the gradient of the function with respect to the 3D object parameters can be effectively computed. In this work, we use the differentiable rasterizer from Pytorch3D [46] to render parametric FLAME model $M(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi})$ to 2D images I_R as

$$R(M(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi}), \boldsymbol{\alpha}, \mathbf{l}, \mathbf{c}) \rightarrow I_R, \tag{2}$$

with albedo parameters $\boldsymbol{\alpha}$, Spherical Harmonics (SH) lighting parameters $\mathbf{l} = \left[\mathbf{l}_1^T, \cdots, \mathbf{l}_9^T\right]^T, \mathbf{l}_k \in \mathbb{R}^3$ to create shaded textures and camera parameters $\mathbf{c} \in \mathbb{R}^3$ which is the concatenation of isotropic scale $\mathrm{s} \in \mathbb{R}$ and 2D translation $\mathrm{t} \in \mathbb{R}^2$.

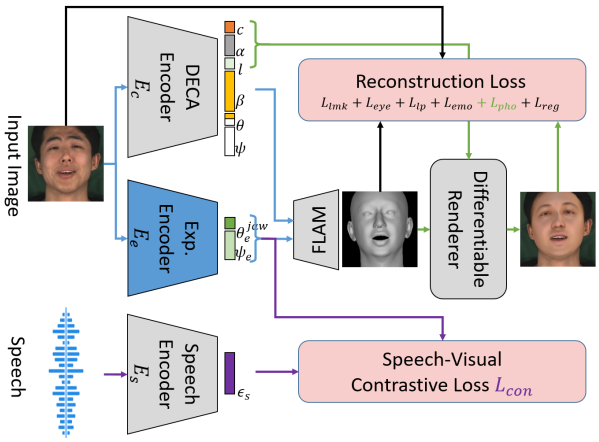

Figure 2: Speech-Assisted Monocular 3D Face Reconstruction. The block in grey color are fixed models or networks, and the Expression Encoder E_e in blue color is the learnable network.

4. Speech4Mesh

4.1. Speech-Assisted Monocular 3D Face Reconstruction

In the proposed Speech4Mesh framework, 3D face sequences are first reconstructed from RGB videos. The detailed structure of our reconstruction module is illustrated in Fig. 2. The overall structure is built on DECA [22], the state-of-the-art method in monocular 3D face reconstruction. DECA learns animatable detailed 3D face models from static RGB images I in a two-stage scheme. The coarse stage first predicts parameters to reconstruct a rough face shape based on the FLAME model. The detail stage then generates a UV displacement map conditioned on expression and jaw pose parameters to achieve fine-grained details. Since the detail stage may bring noise to the reconstructed mesh, we only borrow the coarse part as our backbone. In DECA's coarse part, an encoder (ResNet-50) learns to regress 3DMM coefficients merely under visual constraints, including landmark-related losses, photometric loss, perceptual-level loss, and shape consistency loss. We will introduce some of the losses we have used later in the text.

Expression encoder: Since DECA already can capture the overall facial features well, we fix the pre-trained coarse encoder E_c in DECA for shape, pose (except jaw), camera, albedo, and light parameters. We then train an additional expression encoder E_e to regress expression parameters ψ_e and jaw parameters θ_e^{jaw} as follows:

$$\boldsymbol{\theta_e^{jaw}}, \boldsymbol{\psi}_e = E_e(I). \tag{3}$$

Note that, for computational efficiency, we use MobileNetV2 instead of ResNet-50 as the backbone network.

Only training E_e simplifies the learning goal and enables us to remove losses that are unrelated to lip motion such as identity loss L_{id} and shape consistency loss L_{sc} in DECA, which leads to faster training speed and fewer training resources. Also, speech content would mainly affect mouth-related movements, which are most relevant to expression and jaw pose. The rendering output of the reconstruction part is then denoted as:

$$R\left(M\left(\boldsymbol{\beta}, [\boldsymbol{\theta}, \boldsymbol{\theta_e^{jaw}}], \boldsymbol{\psi}_e\right), \boldsymbol{\alpha}, \mathbf{l}, \mathbf{c}\right) \rightarrow I_{Re}, \tag{4}$$

in which $[\boldsymbol{\theta}, \boldsymbol{\theta_e^{jaw}}]$ means we replace the original vectors representing jaw pose in $\boldsymbol{\theta}$ with $\boldsymbol{\theta_e^{jaw}}$.

Speech-visual contrastive loss: As discussed in previous sections, the majority of current works only utilize the visual information from utterance videos. While these methods have achieved acceptable reconstruction results, it is still an underconstrained problem to retrieve 3D faces merely from 2D planes. Ignoring information contained in audio can lead to a considerable loss in reconstruction. Thus, we propose to leverage speech constraints here to achieve more precise 3D face reconstruction in a contrastive manner. We calculate the contrastive loss using a minibatch of N speech-visual representation pairs (ϵ_s, ϵ_v), where ϵ_v is the hidden state before the last 4-layer MLP in expression encoder, which projects ϵ_v to the target 3DMM space, and ϵ_s is the audio embedding. In specific, ϵ_s is obtained by the pre-trained wav2vec 2.0 [1] model as follows:

$$\epsilon_s = E_s(S), \tag{5}$$

which can extract powerful representations from speech audio. Since we interpolate the frequency of audio features to be two times the video frame rate, we then adopt a linear combination of audio features from time steps $(t - \lfloor T/2 \rfloor) \cdot 2$ to $(t + \lfloor T/2 \rfloor) \cdot 2$ for frame t as the speech-visual positive pair, where T is set to 5 in our experiment. Minimizing the contrastive loss leads to encoders that maximally preserve the mutual information between positive speech-visual pairs. The contrastive loss consists of two terms, in which the first term calculates the speech-to-visual contrastive cost for the i-th pair:

$$\ell_i^{(s \rightarrow v)} = -\log \frac{\exp\left(\langle \epsilon_{\mathbf{i}}^{\mathbf{s}}, \epsilon_{\mathbf{i}}^{\mathbf{v}} \rangle / \tau\right)}{\sum_{k=1}^{N} \exp\left(\langle \epsilon_{\mathbf{i}}^{\mathbf{s}}, \epsilon_{\mathbf{k}}^{\mathbf{v}} \rangle / \tau\right)} \tag{6}$$

where $\langle \epsilon_{\mathbf{i}}^{\mathbf{s}}, \epsilon_{\mathbf{i}}^{\mathbf{v}} \rangle$ calculates cosine similarity between two vectors, i.e., $\langle \mathbf{v}, \mathbf{u} \rangle = \mathbf{v}^\top \mathbf{u} / \|\mathbf{v}\| \|\mathbf{u}\|$. $\tau \in \mathbb{R}^+$ is the temperature parameter, which is fixed to 0.07 in our experiment. While the second term is similar to the first term, it calculates cosine similarity as $\langle \epsilon_{\mathbf{i}}^{\mathbf{v}}, \epsilon_{\mathbf{i}}^{\mathbf{s}} \rangle$. The final contrastive loss is the weighted sum of the two terms:

$$L_{con} = \frac{1}{N} \sum_{i=1}^{N} \left(\lambda \ell_i^{(s \rightarrow v)} + (1 - \lambda) \ell_i^{(v \rightarrow s)}\right), \tag{7}$$

where we set λ to 0.5 in our experiment as a scalar weight.

Weighted landmark loss: We use weighted landmark loss to minimize the L_1 distance between ground-truth 2D key points and corresponding landmarks projected into the image plane from mesh.

$$L_{lmk} = \sum_{i=1}^{68} \lambda_i \left\| \mathbf{k}_i - s\Pi\left(M_i\right) + t \right\|_1 \tag{8}$$

in which λ_i is the weight of the ith landmark, $\Pi\left(M_i\right)$ represents projecting 3D meshes M_i to the 2D pixel space. We set weights of landmarks in the mouth, jaw, and nose area higher than others to capture more accurate geometry representations.

Eye closure loss: We calculate the eye closure loss

$$L_{eye} = \sum_{(i,j)\in E} \left\| \mathbf{k}_i - \mathbf{k}_j - s\Pi\left(M_i - M_j\right) \right\|_1 \tag{9}$$

to force the relative offset of vertices M_i and M_j on the upper and lower eyelid to be close to the offset of ground-truth eyelid landmarks pairs $\mathbf{k}_i$ and $\mathbf{k}_j$. In this way, the reconstructed mesh could capture the blink of eye.

Lip closure loss: The lip closure loss L_{lp} is similar to the eye closure loss. It measures the relative distance between the upper/lower lip and left/right lip corners keypoint pairs.

Photometric loss: The photometric loss

$$L_{pho} = \left\| V_I \odot \left(I - I_{Re}\right) \right\|_{1,1} \tag{10}$$

measures the pixel-to-pixel differences between ground-truth training images and rendered images. V_I is the face skin mask with a value 1 in the face skin area and $\odot$ denotes the Hadamard product.

Emotion consistency loss: To capture more vivid expressions, we borrow the idea from EMOCA [14] that introduces an emotion consistency loss with a fixed emotion recognition network. We compute the L_2 distance between emotion features of input images and output rendered images as follows:

$$L_{emo} = \left\| \epsilon_I^{emo}, \epsilon_{Re}^{emo} \right\|_2, \tag{11}$$

where ϵ_I^{emo} and ϵ_{Re}^{emo} are produced by a fixed emotion recognition network.

Regularization term: We constrain the expression parameters $\boldsymbol{\psi}_e$ and jaw parameters $\boldsymbol{\theta}_{\boldsymbol{e}}^{\boldsymbol{jaw}}$ produced by our expression encoder to be close to the original parameters produced by pre-trained DECA by minimizing their L_2 distance as performed in [24]: $\left\|\psi_e - \psi\right\|_2$ and $\left\|\theta_e^{jaw} - \theta^{jaw}\right\|_2$.

Loss function: To sum up, our optimization goal is:

$$\begin{aligned} L_{\text{recon}} = \lambda_{con} L_{con} + \lambda_{lmk} L_{lmk} + \lambda_{eye} L_{eye} + \lambda_{lp} L_{lp} \\ \lambda_{pho} L_{pho} + \lambda_{emo} L_{emo} + \lambda_{reg} L_{reg}, \end{aligned} \tag{12}$$

where λ with different subscripts are hyper-parameters to balance the corresponding losses. The detailed values of λs can be found in the supplemental material.

4.2. Speech-Driven Emotional 3D Facial Animation

After acquiring sufficient synthesized 4D training data, we can train the audio2mesh module for animation. We use FaceFormer [20], a Transformer-based model to drive 3D face vertices with speech embeddings. FaceFormer [20] is the first work to apply Transformer to 3D facial animation. It has achieved better performances in both realism and lip synchronization than prior methods owing to the Transformer's ability to learn long-term audio dependencies. Specifically, FaceFormer predicts the vertices' movement in an autoregressive manner as follows:

$$\hat{\mathbf{y}}_{\mathbf{t}} = \mathbf{FaceFormer}(\hat{\mathbf{y}}_{<\mathbf{t}}, \mathbf{i_n}, \mathbf{S}), \tag{13}$$

where $\hat{\mathbf{y}}_{\mathbf{t}}$ represents the predicted vertices offset at time step t, and $\mathbf{i_n}$ represents the identity code. Although FaceFormer has achieved remarkable animation results, its performance highly depends on the data quality, since it is trained in a fully supervised scheme where the only supervision is the MSE loss between predicted and ground-truth meshes:

$$\mathcal{L}_{\text{MSE}} = \sum_{\mathbf{t}=1}^{\mathbf{T}} \sum_{\mathbf{v}=1}^{\mathbf{V}} \left\| \hat{\mathbf{y}}_{\mathbf{t},\mathbf{v}} - \mathbf{y}_{\mathbf{t},\mathbf{v}} \right\|^2. \tag{14}$$

In such a way, it's hard to capture some facial details due to the registration loss of the scanned 4D training data. To this end, during pre-training with the synthesized data, we further introduce a pixel-level constraint by computing the photometric loss between the rendered images and corresponding real images as follow:

$$R\left(M_T + \hat{\mathbf{y}}_{\mathbf{t}}, \boldsymbol{\theta}_{\boldsymbol{t}}, \boldsymbol{\alpha}_{\boldsymbol{t}}, \mathbf{l_t}, \mathbf{c_t}\right) \rightarrow I_{Re}^t, \tag{15}$$

where M_T is the template face mesh (the mean of all faces in the training set as in FaceFormer). Note that, when performing differentiable rendering, since the output meshes of the FaceFormer network are all aligned, i.e., without any rotation or camera transform, we borrow $\boldsymbol{\theta}_{\boldsymbol{t}}$[1], $\boldsymbol{\alpha}_{\boldsymbol{t}}, \mathbf{l_t}, \mathbf{c_t}$ estimated by our 3D reconstruction encoders to render the vertices to 2D images. Except for the pixel-to-pixel loss, we also calculate landmark-related losses as in the reconstruction part.

Furthermore, due to the insufficiency of emotional 4D data, current audio2mesh methods generally lack emotion controllability. To make our meshes more lifelike, we introduce a one-hot emotion code to explicitly control the emotional state. Specifically, during pre-training on the emotional 2D video dataset, we embed the emotion category into the FaceFormer network as follows:

$$\hat{\mathbf{y}}_{\mathbf{t}} = \mathbf{FaceFormer}(\hat{\mathbf{y}}_{<\mathbf{t}}, \mathbf{i_n}, \mathbf{e_m}, \mathbf{S}), \tag{16}$$

[1]With slight abuse of notation, the θ_t here does not include the jaw rotation.

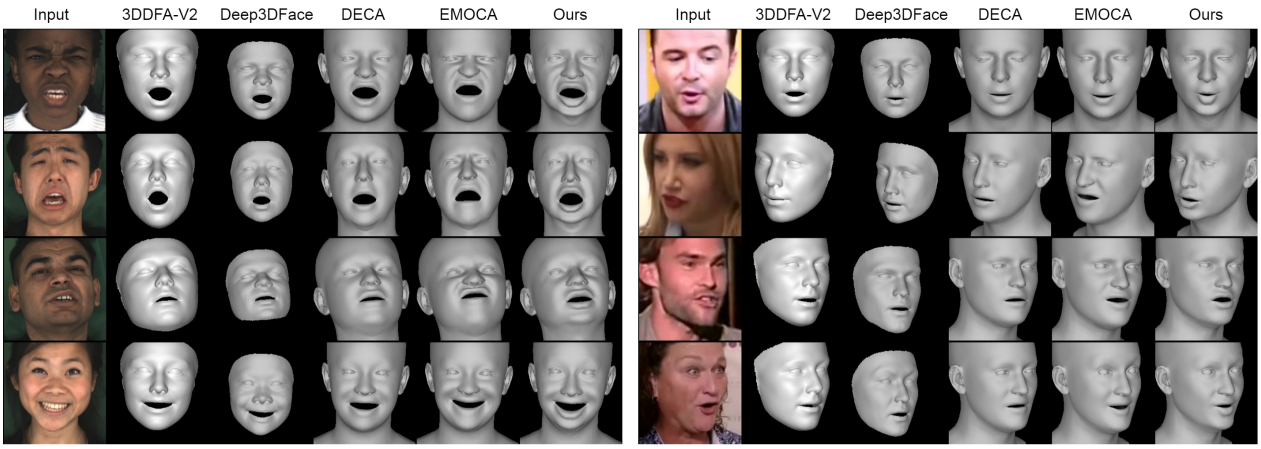

Figure 3: Comparison of the reconstruction results on MEAD (left) and VoxCeleb2 (right) datasets. The first column on the left is the original RGB images. The reconstructed faces from the second to the sixth column are from 3DDFA-V2, Deep3DFace (Pytorch), DECA (coarse), EMOCA (coarse), and ours, respectively. More qualitative results could be found in the supplementary materials.

where $\mathbf{e_m}$ is the one-hot vector projected to a higher dimension space. In this way, during the inference stage, we can achieve speaker-independent emotion-controllable 3D facial animation by embedding different emotion vectors into FaceFormer.

To sum up, by employing the self-supervised differentiable rendering scheme and emotion code, our modified FaceFormer model can generate not only more realistic but also emotion-controllable animated face meshes.

5. Experiments

5.1. Dataset

In this paper, we leveraged RGB video data and 4D mesh data for training the reconstruction module and fine-tuning the audio2mesh module, respectively. A brief introduction of the datasets we used is as follows.

2D Training Data: MEAD [64], VoxCeleb2 [10]. The MEAD dataset is a talking head corpus with different intensities of emotions. It has a total of 281,400 video clips collected from 60 English-speaking actors and actresses under 8 common emotions with 3 levels of intensity. Note that although the dataset contains videos from different views, we only took the front-face data for training to keep more visible details. VoxCeleb2 is a large-scale talking head video dataset collected from open-source media which covers a huge range of identities, races, languages, and head poses. In practice, we randomly sampled 2,000 identities for training. The images were extracted under 30 and 25 fps for MEAD and VoxCeleb2 videos, respectively. We used wav2vec 2.0 [1] to exact the audio embeddings, in which the output frequency is 49 Hz. To ensure temporal alignment of the image and audio, we interpolate the audio embeddings to 50 Hz for the VoxCeleb2 dataset and to 60 Hz for the MEAD dataset.

4D Fine-Tuning Data: VOCASET [13]. VOCASET is a 4D dataset composed of consecutive scanned 3D heads which are all registered to a 5023 vertices mesh. It contains a total of 480 talking motion sequences from 12 subjects under the frame rate of 60 fps. The training, validation, and testing splits are directly adopted from [20].

5.2. Training Setup

The reconstruction part was trained for a maximum of 20 epochs, using the Adam optimizer and a learning rate of $5e-5$. In one iteration, we used 4 mini-batches of data, each of which contains 16 frames randomly sampled from a single video. We further implied distributed data-parallel technique to collect data from 4 GPUs to enlarge the pool of training pairs for contrastive learning.

The audio2mesh network was trained using the Adam optimizer with a learning rate of $1e-4$ for a total of 100 epochs. The parameters of the encoder were initialized with the pre-trained wav2vec 2.0 weights and the parameters of TCN were fixed during training. The period p for the biased causal multi-head self-attention was set to 30.

We pre-trained our audio2mesh network on 4D data produced by our monocular 3D reconstruction network from the MEAD dataset. While it contains 3 levels of emotional intensity, we only use the mildest first level to train. We created 8 one-hot emotion codes (angry, contempt, disgusted, fear, happy, sad, surprised, or neutral) and 42 one-hot speaker identity codes. For each pair of image and its 3D reconstruction in the MEAD dataset, we embedded its corresponding emotion code and speaker ID code to a higher dimension and then concatenated and projected it to the audio2mesh network.

We fine-tuned the audio2mesh network on VOCASET

which contains high-resolution 4D scans. During training, we randomly chose 8 kinds of speaker ID codes defined in the pre-train phase and set the emotion code to neutral.

5.3. Analysis of 3D Face Reconstruction Results

Subjective Evaluation: Visualization of the reconstruction results is demonstrated in Fig. 3. We randomly sampled several frames from MEAD and VoxCeleb2 datasets for comparison. For clear observation of the expression details, we did not apply texture on the face. The sampled frames contain a variety of talking heads with various lip motions, extreme emotions, different genders, head poses, ages, and races. From Fig. 3, we can find that our method apparently outperforms the 3DDFA-V2 and DECA regarding the lip motion as well as the overall appearance. The curly lips and emotional expressions cannot be recovered by these two baselines. EMOCA is an expert in capturing extreme emotions. However, some mild expressions will be significantly exaggerated by EMOCA, which causes unnatural artifacts and inaccurate lip shapes in the reconstruction results. More comparison results with EMOCA could be found in supplementary materials. Note that, we used the Pytorch version of Deep3DFace for comparison, which is claimed to be better than the Tensorflow implementation. As shown in the third column in Fig. 3, the recovered faces by Deep3DFace reveal impressive identity consistency compared with the original RGB image. This may attribute to the identity aggregation mechanism and a different parametric face model (i.e., Basel Face Model [43]). However, the lower face expressiveness is still lower than our proposed method.

We further conducted a user study to make a perceptual evaluation between our and the competitors' reconstruction results. Specifically, 150 participants were asked to judge a side-by-side image composed of the original image, our reconstruction result, and the competitor's result. The participants can select their favorable one or choose the equal option if they cannot distinguish the difference. 50 samples were randomly selected from the MEAD and Voxceleb2 datasets. The result of the study is summarized in Tab. 1.

As we can see from Tab. 1, the result demonstrates an obvious inclination to support our results. It is worth mentioning that about 80 percent of our reconstructed samples are deemed better than or equal with 3DDFA-V2 and DECA methods regarding both full and lower face. EMOCA improves the perception of the results due to its emotion prior. Deep3DFace also achieves better performance as we discussed above. However, owing to the audio-visual alignment, our method achieves better performance at both full-face and the low-face region according to the user study.

Objective Evaluation: Our reconstruction model's ability to capture lip shapes is measured using the lip vertex error as the objective metric, which has been extensively leveraged to evaluate the precision of lip synchronization

Ours vs. Competitor	Face region	Favorability(%) ↑			Ours better or equal
		Equal	Comp.	Ours	
Ours vs. 3DDFA-V2	Full	5.4	19.4	**75.2**	80.6
	Lower	5.4	21.3	**73.3**	78.7
Ours vs. DECA	Full	11.4	18.8	**69.8**	81.2
	Lower	9.8	18.5	**71.7**	81.5
Ours vs. EMOCA	Full	8.4	38.3	**53.3**	58.9
	Lower	8.3	37.2	**54.5**	62.8
Ours vs. Deep3DFace	Full	5.9	41.1	**53.1**	58.9
	Lower	8.7	34.2	**57.0**	65.8

Table 1: A user study of the reconstruction results. We compare the participants' favorability of our reconstructions with the competitors' w.r.t. different regions of the face.

Metric	Ours	EMOCA
Lip Vertex Error ↓	$\mathbf{0.995}e^{-2}$	$1.359e^{-2}$

Table 2: Reconstruction error on VOCASET.

[20, 47]. To calculate this error, we determine the maximal L_2 distance between the lip vertices of the reconstructed meshes and their corresponding ground truth in the VOCASET. We compared Speech4Mesh with EMOCA, which performs the best in the subjective evaluation. As presented in Table 2, our method outperforms EMOCA in terms of lower lip vertex error, demonstrating our superior ability to capture accurate lip shapes. On the contrary, EMOCA performs poorly on this metric due to its exaggerated facial expression with contorted mouth movements.

5.4. Analysis of Speech-Driven Facial Animation Results

Subjective Evaluation: We also performed a user study on the speech-driven animation results. Similar to the reconstruction user study, we recruited 150 participants to judge the animation quality of our method and the vanilla FaceFormer from the realism and lip synchronization dimensions. The speeches for evaluation are from FaceFormer's test data (the demo video of FaceFormer) and the VoxCeleb2 dataset. For each dataset, we sampled 56 speeches (7 ids, 8 sentences for each id). Each participant evaluates 8 videos selected at random. The statistic of the user study is shown in Tab. 3. From Tab. 3 we can find that the favorability rating of our Speech4Mesh is consistently higher than that of FaceFormer, especially on the more challenging VoxCeleb2 speeches. We also visualize several representative syllables and corresponding animations in Fig. 4 for demonstration. In Fig. 4, our animated faces are more expressive and perceptually reasonable. We would like to note that MeshTalk cannot be well-trained on VOCASET which in a different topology, as mentioned in [61]. We have refrained from presenting any quantitative or qualitative comparison results here. However, we do incorporate a straightforward comparison in the supplementary videos. It is highly recommended to see those videos for more details.

Objective Evaluation: We evaluate the lip vertex error of our proposed method and the vanilla FaceFormer on the test set of VOCASET. For simplicity, the scale we used for eval-

Test Speech Data	Criterion	Favorability(%) ↑			Ours better
		Equal	FaceFormer	Speech4Mesh	or equal
Speech from	Realism	24.8	31.8	**43.4**	68.2
FaceFormer Demo	Lip Sync	21.9	36.4	**41.8**	63.6
Speech from	Realism	16.3	26.1	**57.7**	73.9
VoxCeleb2	Lip Sync	16.7	30.7	**52.5**	69.3

Table 3: A user study of the speech-driven results. We compare the participants' favorability of our results with the FaceFormer with respect to realism and lip synchronization.

Method	Lip Vertex Error ↓
VOCA	$6.428e^{-3}$
FaceFormer	$5.355e^{-3}$
Speech4Mesh	$\mathbf{4.863}e^{-3}$

Table 4: Comparison of the lip-sync error between VOCA, the vanilla FaceFormer, and Speech4Mesh.

uation is the same as the FLAME model.

Tab. 4 displays the lip vertex errors of the vanilla FaceFormer method and our Speech4Mesh method. It can be seen that more than a 9% improvement in lip synchronization is brought by our training framework, which restates the importance of data sufficiency for 4D tasks.

5.5. Emotion Control

As introduced in previous sections, our audio2mesh module can be trained on the reconstructed meshes with different emotions. By employing an emotion code, we can manipulate the originally neural 4D talking heads to have different emotions. As shown in Fig. 5, different emotions (e.g., happy, surprised, sad) are embedded into the talking face driven by the same speech. The result reveals the potentiality of more diverse applications by exploiting the 2D videos under the Speech4Mesh framework. More emotion control results could be found in supplementary videos.

5.6. Ablation Studies

To understand the impact of the contrastive loss and the photometric loss on speech-driven animation, we conducted quantitative experiments as shown in Tab. 5. More ablation studies could be found in supplementary materials. For the ablation study of contrastive loss, we pre-trained our audio2mesh module using 4D meshes reconstructed from the MEAD dataset, without incorporating contrastive loss during training. For the ablation study of photometric loss, we simply removed it in the pre-training stage. The quantitative results in Tab. 5. show that pre-training without contrastive loss or photometric loss results in a higher lip vertex error. These results underscore the importance of high-quality reconstructed 4D meshes in our method. They also support the notion that the photometric loss can encourage the audio2mesh network to learn detailed information which can not be captured solely from 4D meshes.

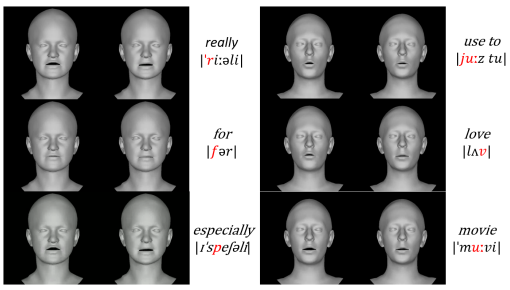

Figure 4: Representative syllables and corresponding animations. The animation results of our Speech4Mesh and FaceFormer are displayed in the first and second columns.

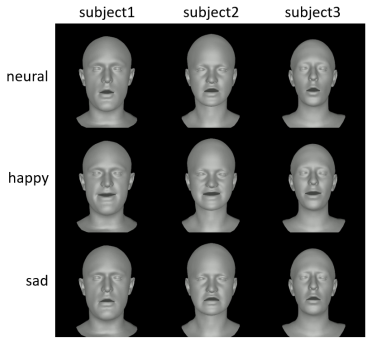

Figure 5: Speech4Mesh samples with different emotions.

Metric	Full	wo. Pho.	wo. Con.	FaceFormer
Lip Vertex Error ↓	$\mathbf{4.863}e^{-3}$	$5.233e^{-3}$	$5.271e^{-3}$	$5.355e^{-3}$

Table 5: Ablation study on the photometric loss and contrastive loss w.r.t. the lip-sync error.

6. Conclusions and Future Work

In this paper, we propose a novel framework named Speech4Mesh to train the audio2mesh network for speech-driven 3D facial animation. In Speech4Mesh, the speech-assisted monocular 3D face reconstruction module can provide significantly more accurate 4D synthesized data for pre-training by leveraging the audio-visual contrastive alignment. By employing the differentiable rendering loss and an emotion code, the modified audio2mesh network can achieve more realistic animation performance and unprecedentedly realize emotion manipulation. Empirical experiments verify the effectiveness and superiority of the Speech4Mesh framework both quantitatively and perceptually. We hope that the proposed Speech4Mesh framework can bring inspiration for similar 4D tasks to address the data scarcity problem.

In the future, we plan to replace the image-based reconstruction backbone DECA with a video-based reconstruction backbone such as MICA [76], neural head avatar [26] or [67]. We will also leverage more high-quality 2D video data for better animation performance and devise a more effective emotion modeling scheme to improve the expressiveness of expressions.

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. 4, 6

[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, Stefanos Zafeiriou, Allan Ponniah, and David Dunaway. A 3d morphable model learnt from 10,000 faces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5543–5552, 2016. 3

[4] Carlos Busso, Zhigang Deng, Michael Grimm, Ulrich Neumann, and Shrikanth Narayanan. Rigid head motion in expressive speech animation: Analysis and synthesis. *IEEE transactions on audio, speech, and language processing*, 15(3):1075–1086, 2007. 2

[5] Chen Cao, Qiming Hou, and Kun Zhou. Displaced dynamic expression regression for real-time facial tracking and animation. *ACM Trans. Graph*, 33(4):1–10, 2014. 3

[6] Yong Cao, Wen C Tien, Petros Faloutsos, and Frédéric Pighin. Expressive speech-driven facial animation. *ACM Transactions on Graphics (TOG)*, 24(4):1283–1302, 2005. 2

[7] Yujin Chai, Yanlin Weng, Lvdi Wang, and Kun Zhou. Speech-driven facial animation with spectral gathering and temporal attention. *Frontiers of Computer Science*, 16(3):1–10, 2022. 1

[8] 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

[9] 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. 1

[10] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *Proc. Interspeech 2018*, pages 1086–1090, 2018. 6

[11] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Asian conference on computer vision*, pages 251–263. Springer, 2016. 2

[12] Michael M Cohen, Rashid Clark, and Dominic W Massaro. Animated speech: Research progress and applications. In *AVSP 2001-International Conference on Auditory-Visual Speech Processing*, 2001. 2

[13] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J Black. Capture, learning, and synthesis of 3d speaking styles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10101–10111, 2019. 1, 2, 6

[14] Radek Daněček, Michael J Black, and Timo Bolkart. Emoca: Emotion driven monocular face capture and animation. In *CVPR*, pages 20311–20322, 2022. 3, 5

[15] Dipanjan Das, Sandika Biswas, Sanjana Sinha, and Brojeshwar Bhowmick. Speech-driven facial animation using cascaded gans for learning of motion and texture. In *European conference on computer vision*, pages 408–424. Springer, 2020. 2

[16] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5203–5212, 2020. 3

[17] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 3

[18] Pengfei Dou, Shishir K Shah, and Ioannis A Kakadiaris. End-to-end 3d face reconstruction with deep neural networks. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5908–5917, 2017. 3

[19] Pif Edwards, Chris Landreth, Eugene Fiume, and Karan Singh. Jali: an animator-centric viseme model for expressive lip synchronization. *ACM Trans. Graph*, 35(4):1–11, 2016. 2

[20] Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. Faceformer: Speech-driven 3d facial animation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18770–18780, 2022. 1, 2, 5, 6, 7

[21] Gabriele Fanelli, Juergen Gall, Harald Romsdorfer, Thibaut Weise, and Luc Van Gool. A 3-d audio-visual corpus of affective communication. *IEEE Transactions on Multimedia*, 12(6):591–598, 2010. 2

[22] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Trans. Graph.*, 40(4):1–13, 2021. 2, 3, 4

[23] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *Proceedings of the European conference on computer vision*, pages 534–551, 2018. 3

[24] Panagiotis P Filntisis, George Retsinas, Foivos Paraperas-Papantoniou, Athanasios Katsamanis, Anastasios Roussos, and Petros Maragos. Visual speech-aware perceptual 3d facial expression reconstruction from videos. *arXiv preprint arXiv:2207.11094*, 2022. 5

[25] Kyle Genova, Forrester Cole, Aaron Maschinot, Aaron Sarna, Daniel Vlasic, and William T Freeman. Unsupervised training for 3d morphable model regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8377–8386, 2018. 3

[26] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural head avatars from monocular rgb videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18653–18664, 2022. 8

[27] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3d dense face alignment. In *European Conference on Computer Vision*, pages 152–168. Springer, 2020. 2

[28] Yudong Guo, Jianfei Cai, Boyi Jiang, Jianmin Zheng, et al. Cnn-based real-time dense face reconstruction with inverse-rendered photo-realistic face images. *IEEE transactions on pattern analysis and machine intelligence*, 41(6):1294–1307, 2018. 3

[29] Patrik Huber, Zhen-Hua Feng, William Christmas, Josef Kittler, and Matthias Rätsch. Fitting 3d morphable face models using local features. In *2015 IEEE international conference on image processing*, pages 1195–1199. IEEE, 2015. 3

[30] Patrik Huber, Guosheng Hu, Rafael Tena, Pouria Mortazavian, P Koppen, William J Christmas, Matthias Ratsch, and Josef Kittler. A multiresolution 3d morphable face model and fitting framework. In *Proceedings of the 11th international joint conference on computer vision, imaging and computer graphics theory and applications*, 2016. 3

[31] Aaron S Jackson, Adrian Bulat, Vasileios Argyriou, and Georgios Tzimiropoulos. Large pose 3d face reconstruction from a single image via direct volumetric cnn regression. In *Proceedings of the IEEE international conference on computer vision*, pages 1031–1039, 2017. 3

[32] Amin Jourabloo and Xiaoming Liu. Large-pose face alignment via cnn-based dense 3d model fitting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4188–4196, 2016. 3

[33] Harim Jung, Myeong-Seok Oh, and Seong-Whan Lee. Learning free-form deformation for 3d face reconstruction from in-the-wild images. In *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2737–2742. IEEE, 2021. 3

[34] 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 Trans. Graph.*, 36(4):1–12, 2017. 1, 2

[35] 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. 3

[36] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. 3

[37] Feng Liu, Dan Zeng, Qijun Zhao, and Xiaoming Liu. Joint face alignment and 3d face reconstruction. In *European Conference on Computer Vision*, pages 545–560. Springer, 2016. 3

[38] Feng Liu, Ronghang Zhu, Dan Zeng, Qijun Zhao, and Xiaoming Liu. Disentangling features in 3d face shapes for joint face reconstruction and recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5216–5225, 2018. 3

[39] Jingying Liu, Binyuan Hui, Kun Li, Yunke Liu, Yu-Kun Lai, Yuxiang Zhang, Yebin Liu, and Jingyu Yang. Geometry-guided dense perspective network for speech-driven facial animation. *IEEE Transactions on Visualization and Computer Graphics*, 2021. 1

[40] Yaojie Liu, Amin Jourabloo, William Ren, and Xiaoming Liu. Dense face alignment. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1619–1628, 2017. 3

[41] Gaurav Mittal and Baoyuan Wang. Animating face using disentangled audio representations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3290–3298, 2020. 1

[42] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 3

[43] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *2009 sixth IEEE international conference on advanced video and signal based surveillance*, pages 296–301. Ieee, 2009. 3, 7

[44] Hai X Pham, Samuel Cheung, and Vladimir Pavlovic. Speech-driven 3d facial animation with implicit emotional awareness: a deep learning approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 80–88, 2017. 3

[45] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 484–492, 2020. 2

[46] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020. 3

[47] Alexander Richard, Michael Zollhöfer, Yandong Wen, Fernando De la Torre, and Yaser Sheikh. Meshtalk: 3d face animation from speech using cross-modality disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1173–1182, 2021. 1, 2, 7

[48] Elad Richardson, Matan Sela, and Ron Kimmel. 3d face reconstruction by learning from synthetic data. In *2016 fourth international conference on 3D vision (3DV)*, pages 460–469. IEEE, 2016. 3

[49] Elad Richardson, Matan Sela, Roy Or-El, and Ron Kimmel. Learning detailed face reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1259–1268, 2017. 3

[50] Sami Romdhani and Thomas Vetter. Estimating 3d shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 986–993. IEEE, 2005. 3

[51] Shunsuke Saito, Tianye Li, and Hao Li. Real-time facial segmentation and performance capture from rgb input. In *European conference on computer vision*, pages 244–261. Springer, 2016. 2

[52] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael J Black. Learning to regress 3d face shape and expression from an image without 3d supervision. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7763–7772, 2019. 3

[53] Matan Sela, Elad Richardson, and Ron Kimmel. Unrestricted facial geometry reconstruction using image-to-image translation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1576–1585, 2017. 3

[54] Jiaxiang Shang, Tianwei Shen, Shiwei Li, Lei Zhou, Mingmin Zhen, Tian Fang, and Long Quan. Self-supervised monocular 3d face reconstruction by occlusion-aware multi-view geometry consistency. In *European Conference on Computer Vision*, pages 53–70. Springer, 2020. 3

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

[56] Sarah Taylor, Taehwan Kim, Yisong Yue, Moshe Mahler, James Krahe, Anastasio Garcia Rodriguez, Jessica Hodgins, and Iain Matthews. A deep learning approach for generalized speech animation. *ACM Transactions on Graphics (TOG)*, 36(4):1–11, 2017. 3

[57] Sarah L Taylor, Moshe Mahler, Barry-John Theobald, and Iain Matthews. Dynamic units of visual speech. In *Proceedings of the 11th ACM SIGGRAPH/Eurographics conference on Computer Animation*, pages 275–284, 2012. 2

[58] Ayush Tewari, Florian Bernard, Pablo Garrido, Gaurav Bharaj, Mohamed Elgharib, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. Fml: Face model learning from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10812–10822, 2019. 3

[59] Ayush Tewari, Michael Zollhöfer, Pablo Garrido, Florian Bernard, Hyeongwoo Kim, Patrick Pérez, and Christian Theobalt. Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2549–2559, 2018. 3

[60] Ayush Tewari, Michael Zollhofer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1274–1283, 2017. 3

[61] Balamurugan Thambiraja, Ikhsanul Habibie, Sadegh Aliakbarian, Darren Cosker, Christian Theobalt, and Justus Thies. Imitator: Personalized speech-driven 3d facial animation. *arXiv preprint arXiv:2301.00023*, 2022. 7

[62] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016. 2, 3

[63] Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gérard Medioni. Regressing robust and discriminative 3d morphable models with a very deep neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5163–5172, 2017. 3

[64] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *European Conference on Computer Vision*, pages 700–717. Springer, 2020. 6

[65] Qianyun Wang, Zhenfeng Fan, and Shihong Xia. 3d-talkemo: Learning to synthesize 3d emotional talking head. *arXiv preprint arXiv:2104.12051*, 2021. 3

[66] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J Cashman, and Jamie Shotton. Fake it till you make it: face analysis in the wild using synthetic data alone. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3681–3691, 2021. 3

[67] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Matthew Johnson, Jingjing Shen, Nikola Milosavljević, Daniel Wilde, Stephan Garbin, Toby Sharp, Ivan Stojiljković, et al. 3d face reconstruction with dense landmarks. In *European Conference on Computer Vision*, pages 160–177. Springer, 2022. 8

[68] Yuyu Xu, Andrew W Feng, Stacy Marsella, and Ari Shapiro. A practical and configurable lip sync method for games. In *Proceedings of Motion on Games*, pages 131–140, 2013. 2

[69] Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 601–610, 2020. 3

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

[71] 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. 1

[72] 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

[73] Yang Zhou, Zhan Xu, Chris Landreth, Evangelos Kalogerakis, Subhransu Maji, and Karan Singh. Visemenet: Audio-driven animator-centric speech animation. *ACM Transactions on Graphics (TOG)*, 37(4):1–10, 2018. 2

[74] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 146–155, 2016. 3

[75] Xiangyu Zhu, Zhen Lei, Junjie Yan, Dong Yi, and Stan Z Li. High-fidelity pose and expression normalization for face recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 787–796, 2015. 3

[76] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. In *European Conference on Computer Vision*, pages 250–269. Springer, 2022. 8