Pose-Aware 3D Talking Face Synthesis using Geometry-guided Audio-Vertices Attention

Bo Li, Xiaolin Wei, Bin Liu, Zhifen He, Junjie Cao, Yu-Kun Lai

***Abstract*—Most of the existing 3D talking face synthesis methods suffer from the lack of detailed facial expressions and realistic head poses, resulting in unsatisfactory experiences for users. In this paper, we propose a novel pose-aware 3D talking face synthesis method with a novel geometry-guided audio-vertices attention. To capture more detailed expression, such as the subtle nuances of mouth shape and eye movement, we propose to build hierarchical audio features including a global attribute feature and a series of vertex-wise local latent movement features. Then, in order to fully exploit the topology of facial models, we further propose a novel geometry-guided audio-vertices attention module to predict the displacement of each vertex by using vertex connectivity relations to take full advantage of the corresponding hierarchical audio features. Finally, to accomplish pose-aware animation, we expand the existing database with an additional pose attribute, and a novel pose estimation module is proposed by paying attention to the whole head model. Numerical experiments demonstrate the effectiveness of the proposed method on realistic expression and head movements against state-of-the-art methods.**

***Index Terms*—Audio-driven, 3D Facial Animation, Pose-Aware, Hierarchical Features, audio-vertices Attention.**

I. Introduction

3D talking face synthesis aims to create virtual life-like visual simulations of human conversation. There are various applications of 3D talking faces, such as virtual customer service agents and digital avatars for gaming. Benefiting from the strong geometric expression capacity, 3D audio-driven facial animation is richer and more vibrant than 2D methods, can be viewed from arbitrary directions, and accurately replicates natural head motion and facial expressions with sufficient fidelity.

Generally, 3D talking face techniques aim to establish a correlation between input audio and realistic 3D facial expressions along with head movements [1]. Nevertheless, the current methods [2]–[4] encounter a common limitation in that they frequently lack intricate facial expressions and precise head poses, leading to synthesized results that do not consistently resemble authentic human face animation. The inconsistency can be traced back to the following issues. Among existing methods [2]–[4], encoder-decoder structured networks are commonly used to establish the relationship between audio and facial spaces. Their common drawback is that the encoders only focus on global features of audio or face meshes. Hence it is difficult for these methods to capture subtle nuances of mouth shape, eye movement and forehead wrinkles which are essential for creating refined and realistic expressions. In addition, despite the popularity of datasets such as VOCASET [2] and Multiface [5], they lack the attribute of head poses corresponding to audio. As a result, most existing methods can only produce facial animation with a fixed "zero pose", and cannot produce pose-aware realistic animation.

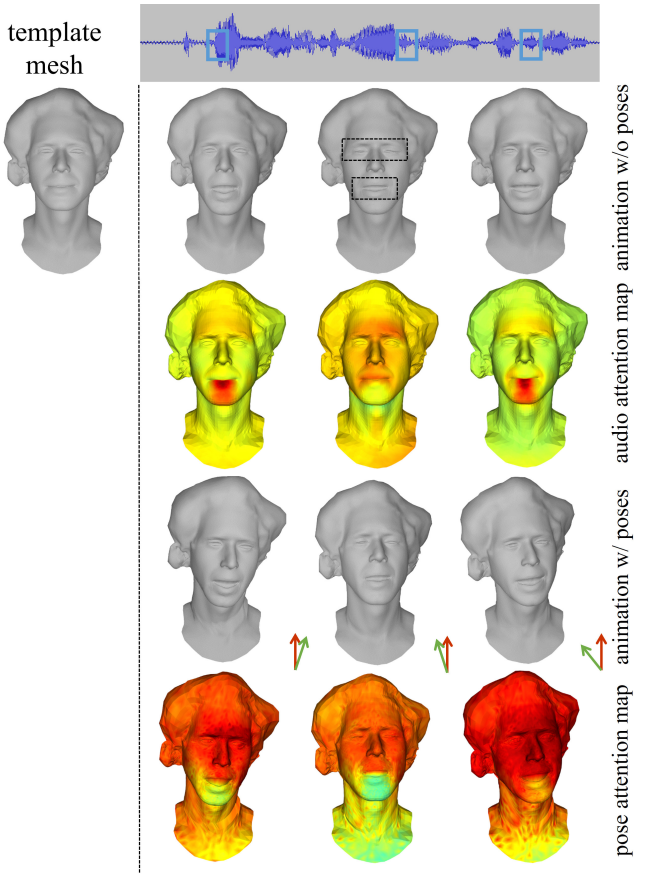


Fig. 1. Results of audio-driven facial animations generated by the proposed method. Given a template mesh and an audio clip, our method produces a mesh deformation sequence with reasonable mouth and eye movements, which also contains smooth head movements with realistic poses. To better demonstrate how the method works, we visualize the audio and pose attention maps. To illustrate the pose variation, we plot the vertex normal at the top of the nose. The arrows in red and green indicate the orientation of the original template mesh and the synthesized mesh, respectively. This visualization also applies to subsequent figures in this paper.

To address the challenges stated above, a novel pose-aware and realistic 3D talking face synthesis algorithm is proposed in this paper. On the one hand, to enhance the level of details in facial animation synthesis, a novel hierarchical audio feature

Bo Li, Xiaolin Wei, Bin Liu and Zhifen He is with School of Mathematics and Information Science, Nanchang Hangkong University, Nanchang, China. Junjie Cao is with the School of Mathematics, Dalian University of Technology, Dalian, China. Yu-Kun Lai is with the School of Computer Sciences and Informatics, Cardiff University, Cardiff, UK.

is designed. In contrast to the global features employed in existing works [2]–[4], the proposed hierarchical feature is composed of a global attribute feature and a series of vertex-wise local latent movement features. In addition, in order to fully exploit the topology of facial models, we further propose a novel geometry-guided audio-vertices attention (GAVA) module to get geometry-consistent movements of individual vertices by using vertex connectivity relations to take full advantage of corresponding hierarchical audio features. On the other hand, to create life-like talking face animation with realistic pose movements, we propose to expand the existing 3D talking head datasets with a novel 3D pose attribute estimation algorithm, then an attention-based head pose prediction module is designed to generate realistic and pose-aware head movements according to the input audio. Numerical experiments are conducted to validate the effectiveness of the proposed method on realistic expression and head movements against state-of-the-art methods.

In summary, the main contributions of the work include:

- A novel geometry-guided audio-vertices attention method is proposed to predict detailed and geometry-consistent facial expressions by taking full advantage of inherent geometric structure constraints.
- A hierarchical feature composed of a global attribute feature and a series of vertex-wise local latent movement features is proposed to achieve detailed facial expressions.
- A novel 3D pose estimation method is proposed to add complementary pose attributes to the popular datasets, including VOCASET [2] and Multiface [5].
- Qualitative and quantitative experiments demonstrate that the proposed approach outperforms state-of-the-art methods.

II. Related Work

As a key technology of human-computer interaction in the virtual environment, audio-driven facial animation has attracted a lot of research [1], [6]. Based on the representation of talking heads, the majority of existing methods can be divided into two primary groups: 2D synthesis methods and 3D synthesis methods.

A. 2D synthesis methods

In 2D synthesis methods, facial animation is generated mainly using facial landmarks, semantic maps, 3D parametric models or image translation as bridges to solve the problem. For example, methods [7], [8] utilize facial landmarks as an intermediate layer for mapping from low-dimensional audio to high-dimensional video. Another class of methods [9], [10] uses image-to-image translation to generate facial animation, where convolutional neural networks or generative adversarial networks are used to learn the joint embedding of face and audio. Unlike the above methods, Ye et al. [11] propose a new image encoding structure, where they extract features from the audio input and reshape these features as dynamic convolution kernels of the encoder network. By modularizing the representations of talking human faces into the spaces of speech content, head pose, and identity respectively, Zhou et al. [12] achieved results with more accurate lip synchronization. Wang et al. [13] designed an audio-visual correlation transformer that takes phonemes and facial keypoint-based motion fields as input to enable single-speaker training, while Huang et al. [14] performed audio-driven facial video synthesis via neural rendering from tri-planes [15] to produce realistic frames. All of the 2D methods described above operate in the pixel space and cannot be easily generalized to producing 3D animation sequences.

B. 3D synthesis methods

3D synthesis methods can be divided into parametric and non-parametric methods in general.

1) Parametric methods: The main idea of parametric methods [16]–[21] is to learn the mapping between speech features and semantic coefficients represented by parameterized face models [22]–[24]. The main differences among these methods are the speech encoder and coefficient regression model. For example, Zhang *et al*. [20] use a convolution-based generative adversarial network to produce head poses for a given audio, and do not adopt the multi-layer perceptron (MLP) architecture utilized in most methods. In addition, to better model the head movement of faces, other researchers [25]–[29] attempt to predict the semantic parameters of head poses from captured face videos. Limited by the linear assumption of face-parameterized representations such as FLAME [24] and 3DMM [22], the reconstruction accuracy and flexibility of parametric methods are not good enough especially for facial details, and it is unable for these methods to control the local semantics of lips, eyes and wrinkles.

2) Non-parametric methods: Non-parametric methods directly use the geometric representation of the 3D head and aim to learn the movement of each vertex. To improve the generalization capability, VOCA [2] employs principal component analysis for the initial setup of the face representation latent space and then uses a neural network to subsequently update and improve this representation. GDPnet [3] starts by utilizing an autoencoder network to learn the geometric prior of the face representation latent space from a facial mesh dataset. Subsequently, they apply the learned geometric prior to constraining the face representation space, which is subsequently learned from the speech. Lahiri *et al*. [30] introduce a new model that incorporates personalized information from videos to improve the realism of 3D face animation. FaceFormer [31] employs a transformer-based model to analyze and capture the mapping between audio space and facial movements. MeshTalk [4] introduces a two-stage talking face algorithm. In the first stage, a latent expression space is learned with aligned audio and facial mesh. In the second stage, an audio-conditioned autoregressive network is employed to synthesize the facial animation. CodeTalker [32] is also a two-step process. Instead of regarding the audio-vertices mapping as a continuous regression task as in MeshTalk, CodeTalker conducted cross-modal mapping in a learned quantified latent space. Nevertheless, the non-parametric methods discussed above solely focus on the global feature of audio, neglecting the essential local spatial attention in audio features. As a consequence, this oversight

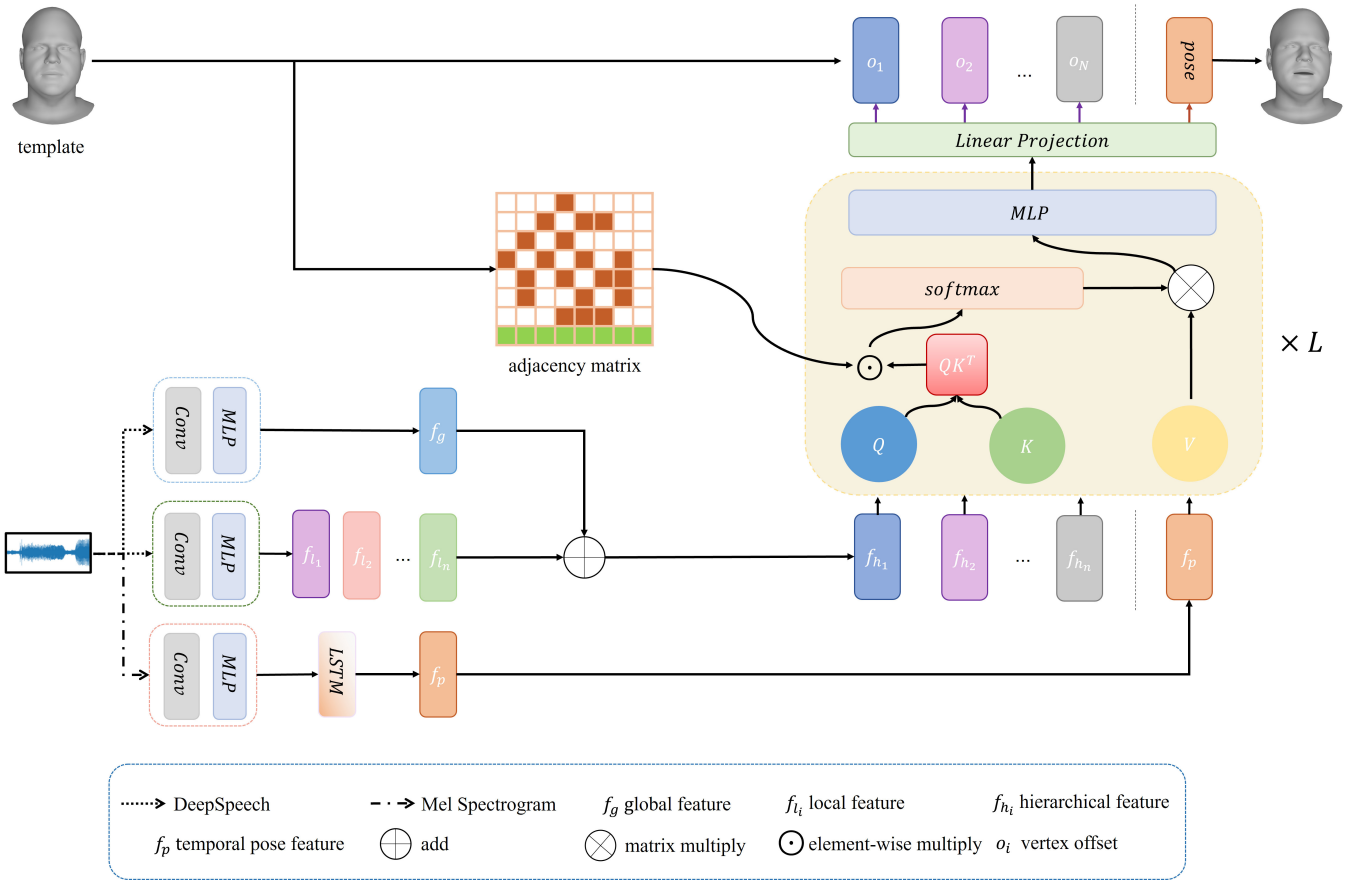


Fig. 2. The pipeline of the proposed pose-aware 3D facial animation synthesis method.

can lead to a lack of intricate and detailed animation in the final output. In addition, as the existing public talking head datasets including VOCASET [2] and Multiface [5] only provide "zero-pose" data, most methods cannot produce realistic talking faces with fluent pose movement.

III. Pose-aware 3D Facial Animation Synthesis

Given a template mesh and input audio, our goal is to generate a realistic and detailed 3D facial expression animation with fluent poses consistent with the input audio. In this section, a novel geometry-guided audio-vertices attention method is proposed for pose-aware 3D facial animation synthesis. The pipeline of the proposed method is illustrated in Fig. 2. First, instead of merely using global features as done in related work [2]–[4], we propose to extract both global facial attribute feature and local vertex-wise latent movement feature based on the DeepSpeech [33] encoding, and then both features are fused to produce the hierarchical feature. To get the underlying sentiment of the input audio which is crucial to the head pose, we extract the Mel spectrogram feature of the audio and utilize a Long Short-Term Memory (LSTM) network to encode the temporal audio feature. Then a novel geometry-guided audio-vertices attention module is designed to predict both the vertex-wise movement and the global pose transformation. Finally, the animated facial model will be generated by performing the predicted transformation on the template mesh.

The remainder of this section is organized as follows. Section III-A illustrates the symbol definition, while Section III-B introduces the proposed method to augment existing datasets with the pose attribute, followed by the description of the proposed hierarchical feature extracted from the input audio signal and defined holistically as well as at individual vertices in Section III-C. Finally, the geometry-guided audio-vertices attention method is described in Section III-D.

A. Symbol Definition

In this paper, we organize the training data in the following form, $\{(\mathbf{I}, \mathbf{y}_i, \mathbf{p}_i, \mathbf{d}_i, \mathbf{m}_i)\}_{i=1}^{T}$. $\mathbf{I} \in \mathbb{R}^{N\times 3}$ denotes the template mesh and each row of $\mathbf{I}$ contains the x, y, z coordinates of a vertex. N is the number of vertices of the mesh. $\mathbf{y}_i \in \mathbb{R}^{N\times 3}$ and $\mathbf{p}_i \in \mathbb{R}^3$ denote the ground truth spatial coordinates and head pose of the ith frame. $\mathbf{d}_i \in \mathbb{R}^{W\times D}$ is the speech feature window at the ith frame generated by DeepSpeech [33], where D is the number of phonemes in the alphabet plus an extra one for a blank label and W is the window size. $\mathbf{m}_i \in \mathbb{R}^{F\times L}$ represents the Mel spectrogram transformed from the raw waveform at the ith frame, where F is the number of Mel filter banks and L is the length. T is the total number of frames.

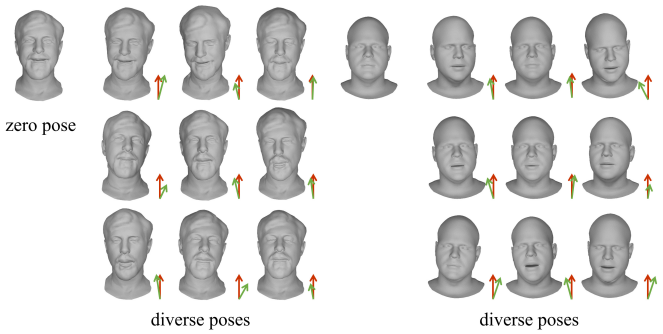


Fig. 3. The diverse poses produced by the proposed method.

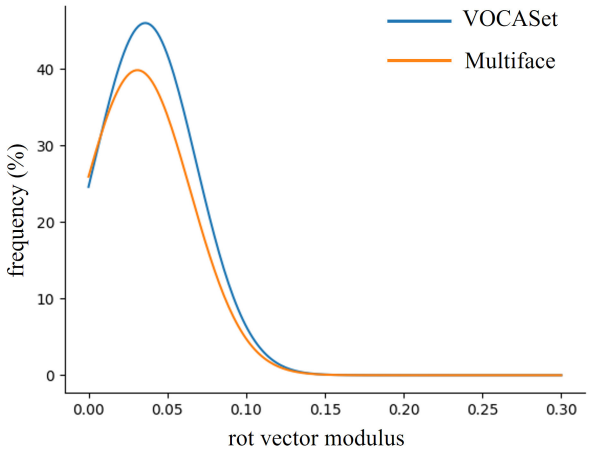


Fig. 4. The statistics of pose movement on both datasets. The x-axis denotes the modulus of the pose transformation vector, while the y-axis means the frequency of the corresponding movement.

B. Pose Attribute Augmentation

Existing public 3D talking mesh datasets including VOCASET [2] and Multiface [5] only provide the animated head meshes with "zero pose", and do not have the flexible pose attribute with respect to audios with various sentiments. Therefore, most of the existing methods [2]–[4] trained on the above two datasets cannot provide pose-aware 3D facial animation.

In order to generate realistic facial animation with audio-related pose variations, we propose to expand the popular VOCASET and Multiface datasets with an additional "pose" attribute. The motivation for the proposed pose augmentation method is as follows. First, 2D talking face animation methods can produce facial images with consistent head poses benefiting from the huge amounts of training videos with real poses. Second, although 2D face animation methods cannot render realistic facial images as well as 3D synthesis algorithms, especially in cases with obvious occlusions, the predicted 2D poses are reliable and can be utilized to help predict a proper 3D head pose.

The proposed pose augmentation method is composed of two stages. First, given the rendered image of a subject in VOCASET (or Multiface) and its corresponding audio, we use the image-based audio-driven talking head method [34] to synthesize a facial video with realistic pose variations. Then, we utilize the method [35] to predict the 3D pose parameters of the head object from the video frames, which are consistent with the FLAME model [24]. To get more consistent poses, a Gaussian filter with a standard deviation of 1 and a window size of 15 is used to smooth the estimated pose parameters along the time axis. Some examples of the generated head models with audio-related poses are shown in Fig. 3, and the statistics of pose movement computed on both datasets are shown in Fig. 4.

In order to predict the proper latent pose feature f_p from the given audio, we propose to use the Mel spectrogram rather than the DeepSpeech feature, since the Mel spectrogram feature is more responsive to the speaker's emotions and is thus more related to head poses. An ablation study is conducted in the supplementary material to compare the performance of the pose prediction by the above two features. An LSTM network is then utilized to extract the latent pose feature $f_p(m_j)$ from the Mel spectrogram features of the input audio clip (Fig. 2). The detailed network architecture is illustrated in the supplementary material.

C. Hierarchical Feature Construction

Most of the existing related works [2]–[4] merely extract global features from speech and then map them to the motion of each vertex. Although the global audio feature has strong relations to vertex-wise movement, it encodes all the information of the input audio including sentiment, global head poses, expressions and vertex movements, etc. As a result, the animated meshes predicted only by the global feature suffer from a lack of detailed expressions, such as the movement around the mouth regions (Fig. 9).

In this paper, we propose to build hierarchical features extracted from the audio signal including a global attribute feature f_g and a series of vertex-wise local latent movement features $\{f_{l_j}\}_i$, where $j = 1, \cdots, N$ denotes the index of each vertex and i is the frame index. Global feature $f_g(d_i)$ encodes the holistic audio feature d_i, such as facial expressions with the input audio, while local features $\{f_{l_j}(d_i)\}$ indicate the localized vertex-wise movement complemented with the global attribute f_g. Then the hierarchical feature $\{f_{h_j}\}_i$ is constructed by concatenating or summing up both global and local features. Numerical experiments show that the performance of the two operations is similar and we choose to sum up global and local features in this paper. The framework of hierarchical feature extraction is shown in Fig. 2, and the network details are illustrated in the supplementary material. An ablation study is designed in Section IV to validate the effectiveness of the proposed hierarchical feature.

D. Geometry-guided Audio-vertices Attention Mechanism

Most of the related talking mesh synthesis methods [2]–[4] utilize MLPs as the decoder to predict the vertex-wise movement. However, MLP is redundant for the task of talking mesh as it computes the relationships among any vertices

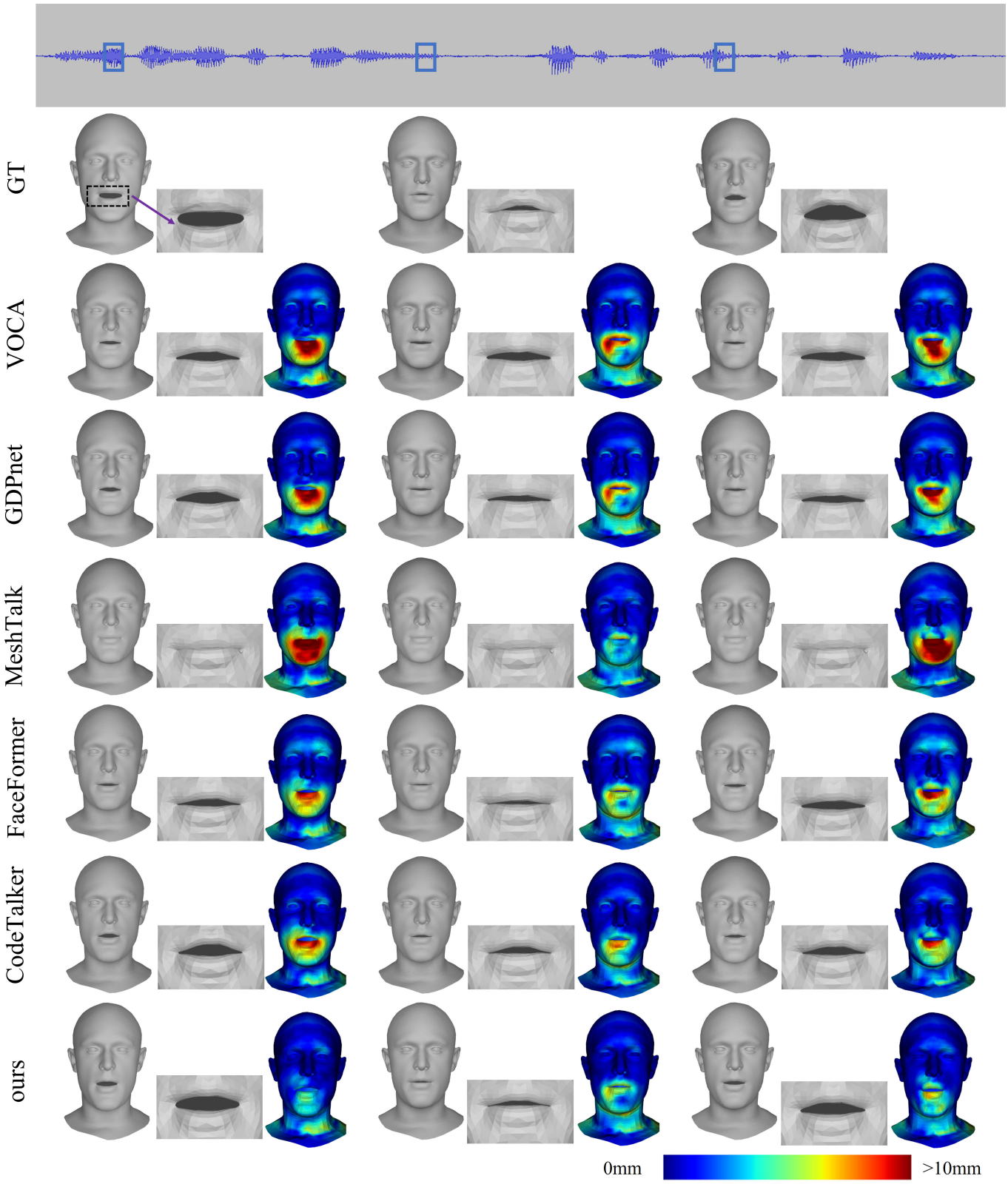


Fig. 5. Experimental results on VOCASET [2]. For clearer visualization, we enlarge the mouth area, and the color maps give the distribution of vertex-to-vertex distance errors (unit: millimeter). We can find that our method has a better ability to preserve details, such as the lips and chin.

blindly and ignores the intrinsic geometric prior for registered head meshes.

In order to find the most salient vertices with respect to the input audio and fully utilize the inherent geometry prior, a novel geometry-guided audio-vertices attention module (GAVA) is proposed to predict the realistic facial animation. The pipeline of GAVA is shown in Fig. 2. The hierarchical geometric features f_{h_j} along with the global pose feature f_p are set as the input tokens of an encoder-only transformer. Each unit of the proposed geometry-driven audio-vertices transformer is composed of two modules: geometric attention and pose attention. Geometric attention is designed to estimate the detailed movement of each vertex according to the hierarchical audio features and intrinsic head model geometry, while pose attention is proposed to estimate the consistent and realistic global pose by analyzing the features of all vertices.

Traditional transformers with self-attention (SA) take a global approach, i.e., the interactions between all pairs of vertices (geometric tokens) are computed irrespective of their local topology relationship. Hence, the architecture does not incorporate the intrinsic geometric prior to a standard base head model. To fully exploit the topology of facial models, a novel geometry-guided audio-vertices attention module is proposed to predict the realistic movement of each vertex with the geometric prior regularization. Specifically, we limit the range that the hierarchical features of each vertex can "see". Each vertex can only pay attention to those vertices that are directly connected to it in each attention module. The motivation is inspired by traditional heat diffusion on meshes [36], and the attention map will learn to propagate to the final meaningful regions with the geometric prior regularization. An experiment is designed to validate the intuition in Sec. IV, and the results showed that our design yielded better results.

In addition to the vertex-wise geometry tokens f_{h_j}, we set the global pose feature f_p as the last token to predict the corresponding pose transformation. We assume that the the prediction of global pose attribute can "see" the features of the whole mesh while the vertex-wise detailed facial geometric features are independent of the global head pose.

Based on the above analysis, the geometry-guided audio-vertices attention is designed as follows. First, we extract the adjacency matrix of the template head model, $\mathcal{M} \in \mathbb{R}^{N\times N}$, where $\mathcal{M}_{i,j}$ represents the connection relationship between the ith and jth vertices. If there is an edge between these two vertices, $\mathcal{M}_{i,j}$ will be set to 1, otherwise, it will be assigned as negative infinity. In order to predict the global pose transformation, one more row and one more column are augmented on $\mathcal{M}$, corresponding to the global head pose. Therefore, we can get an augmented adjacency matrix $\tilde{\mathcal{M}} \in \mathbb{R}^{(N+1)\times(N+1)}$, where the values of the last row are set to 1, and the first N elements in the last column are set to negative infinity, i.e., the head pose depends on detailed features of all the vertices, but vertex features are independent of the head pose.

Then, the geometry-guided audio-vertices attention can be formulated as follows,

$$\text{softmax}(\frac{\mathbf{F}\mathbf{W}_Q(\mathbf{F}\mathbf{W}_K)^T}{\sqrt{d_k}} \odot \mathcal{M})(\mathbf{F}\mathbf{W}_V). \tag{1}$$

where $\mathbf{F} = \{f_{h_1}, \cdots, f_{h_N}, f_p\} \in \mathbb{R}^{(N+1)\times d}$, d is the feature dimension. $\mathbf{W}_Q$, $\mathbf{W}_K$ and $\mathbf{W}_V$ are three trainable linear projection layers, corresponding to queries, keys and values. d_k is the dimension of the queries and keys. As the first N values of the last row of $\mathcal{M}$ are set to 1, the pose prediction module is computed by integrating the movements of the whole head model.

Finally, The output of the final transformer layer is further encoded into the vertex-wise displacement $\mathbf{O} = \{o_1, o_2, ..., o_N\} \in \mathbb{R}^{N\times 3}$ and the global head pose transformation parameters $\mathbf{T_p}$. The ith frame of the predicted facial model can be computed by

$$\hat{\mathbf{y}}_i = \mathbf{T}_p(\mathbf{I} + \mathbf{O}) \tag{2}$$

E. Loss function

The training loss $\mathscr{L}$ is composed of four items, including reconstruction loss $\mathscr{L}_r$, velocity loss $\mathscr{L}_v$, eye loss $\mathscr{L}_e$, and pose loss $\mathscr{L}_p$:

$$\mathscr{L} = \mathscr{L}_r + \lambda_1\mathscr{L}_v + \lambda_2\mathscr{L}_p + \lambda_3\mathscr{L}_e, \tag{3}$$

where λ_1, λ_2 and λ_3 are weight parameters.

The reconstruction loss $\mathscr{L}_r$ between the ground truth $\mathbf{y}_i$ and the predicted 3D model $\hat{\mathbf{y}}_i$ is defined as the mean squared error of vertex-to-vertex displacements,

$$\mathscr{L}_r = \frac{1}{N}\sum_{j=1}^{N} \|\mathbf{y}_i^j - \hat{\mathbf{y}}_i^j\|_2, \tag{4}$$

where $||\cdot|||_2$ is the L_2-norm.

Velocity loss is used to induce temporal stability, which measures the smoothness of the prediction in the context of the sequence. It is formulated as

$$\mathscr{L}_v = \frac{1}{N}\sum_{j=1}^{N} \|(\mathbf{y}_i^j - \mathbf{y}_{i-1}^j) - (\hat{\mathbf{y}}_i^j - \hat{\mathbf{y}}_{i-1}^j)\|_2. \tag{5}$$

Since eye movements have a limited correlation with audio, further eye loss is required for the prediction of the eye movements like blinking. We propose to calculate the KL divergence between the anticipated and actual movements of the eye region, treating them as random variables. Specifically, the loss $\mathscr{L}_e$ is defined as the following:

$$\mathscr{L}_e = \underset{j\in M_{eye}}{KL}(\hat{y}^j, y^j), \tag{6}$$

where $\hat{y}^j$ denotes the coordinates of the j-th vertex in all frames of batches, y^j is the corresponding ground-truth position, and M_{eye} refers to the mask of eye regions. It is crucial to note that we no longer distinguish between batches while computing the loss $\mathscr{L}_e$, because we assume that the way human eyes move follows a similar pattern.

The pose loss function $\mathscr{L}_p$ is utilized to constrain the predicted pose to be similar to the ground truth obtained in Sec. III-B.

$$\mathscr{L}_p = \|\mathbf{p}_i - [\mathbf{T}_p(\mathbf{p}_0)]_i\|_2\,. \tag{7}$$

where $\mathbf{p}_0$ is the initial "zero-pose" of template head model.

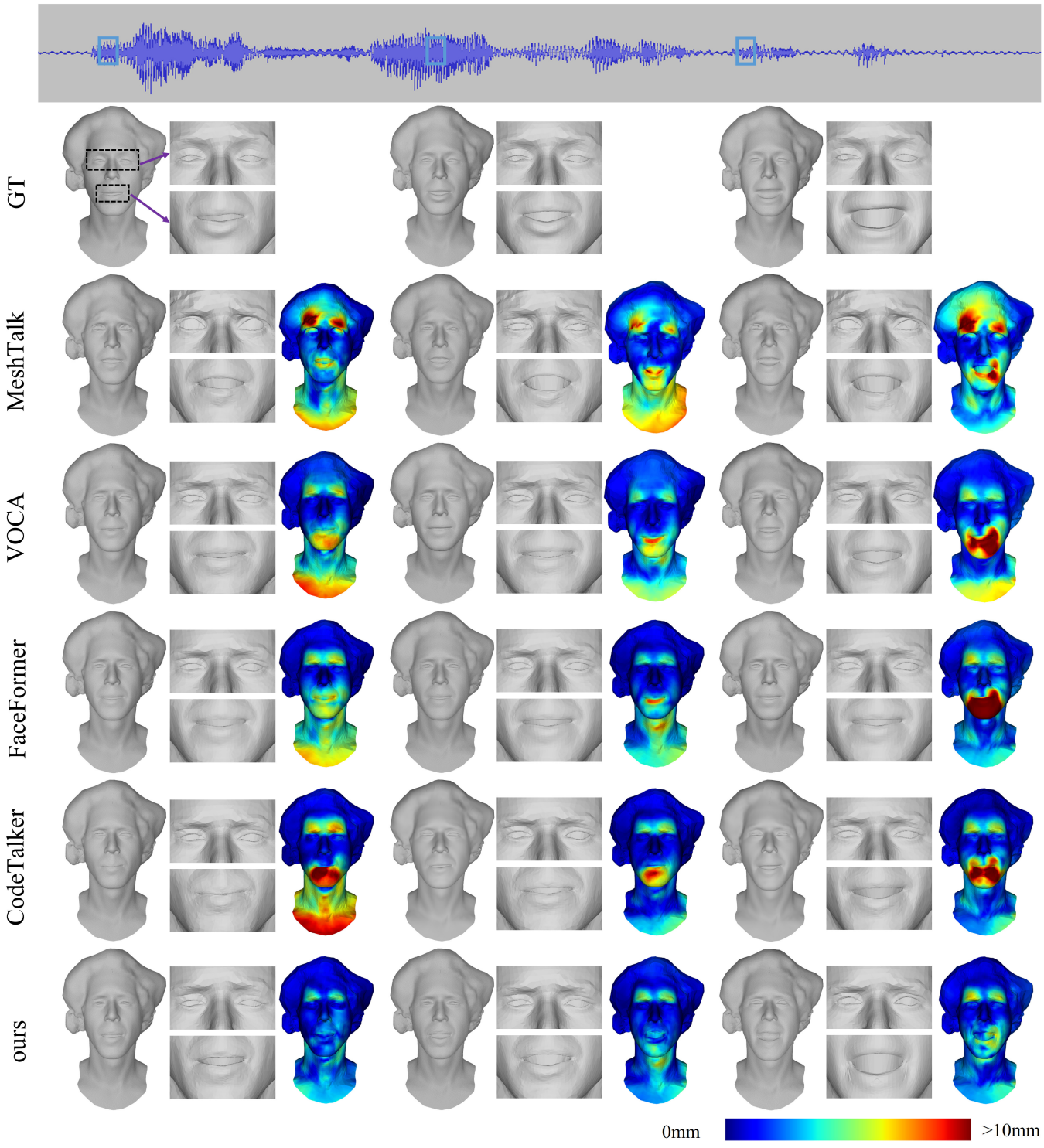


Fig. 6. Experimental results on Multiface [5]. For clearer visualization, we enlarge the mouth and eye areas, and the color maps give the distribution of vertex-to-vertex distance errors (unit: millimeter). We can find that our method has a better ability to preserve details, such as the lips and eyes.

TABLE I
QUANTITATIVE COMPARISON WITH STATE-OF-THE-ART METHODS.

	VOCASET [2]				Multiface [5]			
	E_{vl}	E_{ve}	FDD	$AFDD$	E_{vl}	E_{ve}	FDD	$AFDD$
VOCA [2]	6.384	1.993	1.927	1.624	5.075	3.224	3.353	2.514
GDPnet [3]	6.062	1.986	1.929	1.570	-	-	-	-
MeshTalk [4]	6.449	2.097	2.657	2.226	6.385	4.462	3.355	2.369
Faceformer [31]	5.506	2.013	1.673	1.349	5.285	**3.098**	4.655	3.221
CodeTalker [32]	5.278	2.060	1.544	1.098	6.139	4.124	2.507	1.711
Ours	**5.081**	**1.932**	**1.407**	**1.085**	**4.567**	3.171	**2.222**	**1.699**

IV. EXPERIMENT

In this section, we first introduce the implementation details and datasets used in the proposed method. Then qualitative and quantitative experiments are conducted to demonstrate the effectiveness of our method. Finally, two ablation study experiments are conducted to validate the effectiveness of the proposed modules.

Implementation details. We train the network using the Adam optimization algorithm on an NVIDIA A100 GPU. The values of hyperparameters λ_1, λ_2 and λ_3 in the loss function (Eq. 3) have been assigned as 10, 0.001, and 0.005, respectively. The batch sizes are set as 16 and 8 for the VOCASET and Multiface datasets respectively, and the training process consists of 70,000 iterations. The parameter W in DeepSpeech is set to 16 and the F in the Mel spectrogram is set to 80.

Datasets. The VOCASET dataset [2] comprises a comprehensive collection of audio-4D scan pairs obtained from 6 female and 6 male subjects. Each subject delivers 40 sentences for the recordings. The 3D facial movements are captured at a frame rate of 60 FPS and are accurately registered using the publicly available generic FLAME model [24]. All facial meshes in the dataset are in a standardized "zero pose" state.

Public Multiface dataset [5] contains a collection of audio-4D scan pairs captured from 13 subjects, one subject speaks 12 sentences and others speak 50 sentences, and 3D facial movements are captured at a frame rate of 30FPS. However, neither of these datasets has variations in the head posture.

A. Audio-driven 3D facial animation

In this section, we compare the performance of audio-driven 3D facial animation with "zero pose" against state-of-the-art methods, VOCA [2], GDPnet [3], MeshTalk [4], FaceFormer [31] and CodeTalker [32] on both VOCASET [2] and Multiface [5] datasets. Note that, we randomly choose speaking styles for VOCA, GDPnet, FaceFormer and CodeTalker methods. As done in the baselines [2]–[4], we use the mean of the maximum error in all frames for the lip and eye regions as the evaluation metrics, denoted as E_{vl} and E_{ve}, respectively. Specifically, E_{vl} can be written in the following form:

$$E_{vl} = \frac{1}{T}\sum_{i=1}^{T} \underset{j \in M_{lip}}{max} (\|\hat{y}_i^j - y_i^j\|_2), \tag{8}$$

where $\hat{y}_i^j$ denotes the jth vertex coordinates in the predicted ith frame, y_i^j denotes the ground true position, $||\cdot|||_2$ denotes the Euclidean distance, M_{lip} denotes the mask of lips, and T denotes the total number of frames. E_{ve} is similarly defined.

As discussed in related work, CodeTalker [32], L2 distance with ground truth vertices E_{ve} cannot assess the accuracy of expressions within eye regions completely. Therefore, we also used the Upper Face Dynamics Deviation (FDD) metric proposed in CodeTalker to evaluate the performance of eye movements.

$$\mathrm{FDD}(\mathbf{M}_{1:T}, \hat{\mathbf{M}}_{1:T}) = \frac{\sum_{v \in \mathcal{S}_U} (\mathrm{dyn}(\mathbf{M}_{1:T}^v) - \mathrm{dyn}(\hat{\mathbf{M}}_{1:T}^v))}{|\mathcal{S}_U|}, \tag{9}$$

where $\mathbf{M}_{1:T}^v, \hat{\mathbf{M}}_{1:T}^v \in \mathbb{R}^{3\times T}$ denote the ground truth and predicted motions of the v-th vertex respectively, and $\mathcal{S}_U$ is the index set of upper-face vertices. $\mathrm{dyn}(\cdot)$ denotes the standard deviation along the temporal axis.

FDD is proposed to measure the distribution consistency between the predicted and ground truth eye motion space. However, we found some potential issues in Eq. 9 may limit the effectiveness of FDD. First, the deviation $\mathrm{dyn}(\cdot)$ in Eq. 9 is defined based on the element-wise L2 norm of each motion vector $\mathbf{M}_{1:T}^v \in \mathbb{R}^3$, which result in the loss of the direction of the motion. Second, the motion deviation $\mathrm{dyn}(\mathbf{M}_{1:T}^v) - \mathrm{dyn}(\hat{\mathbf{M}}_{1:T}^v)$ is summed directly within the eye regions. As the motion deviation is a signed value, some real motion statistics cannot be computed accurately (as positive and negative values may unintentionally cancel each other). Based on the above analysis, we proposed an improved axis-based upper-face dynamics deviation (AFDD). First, we propose to compute the standard deviation along each axis x, y and z, and then we summarize the absolute value of the difference of the standard deviation within the eye regions along each axis. The formula of the proposed AFDD is defined as follows,

$$\mathrm{AFDD} = \frac{\sum_{v \in \mathcal{S}_U} \sum_{i \in \{x,y,z\}} (|\,\mathrm{dyn}(\mathbf{M}_{1:T}^{v_i}) - \mathrm{dyn}(\hat{\mathbf{M}}_{1:T}^{v_i})|)}{3 * |\mathcal{S}_U|}. \tag{10}$$

Furthermore, we also employ qualitative visual perception as a criterion. Qualitative results are shown in Fig. 5 and Fig. 6, while quantitative errors are shown in Table I. Note that, the results of GDPnet [3] for the Multiface dataset are not included as the authors have not released the training code. In Fig. 5, we can find that other methods produce more obvious and larger errors than the proposed method around the mouth regions. Compared with VOCASET, the Multiface dataset provides training data with more detailed expressions, and the experimental results as shown in Fig. 6

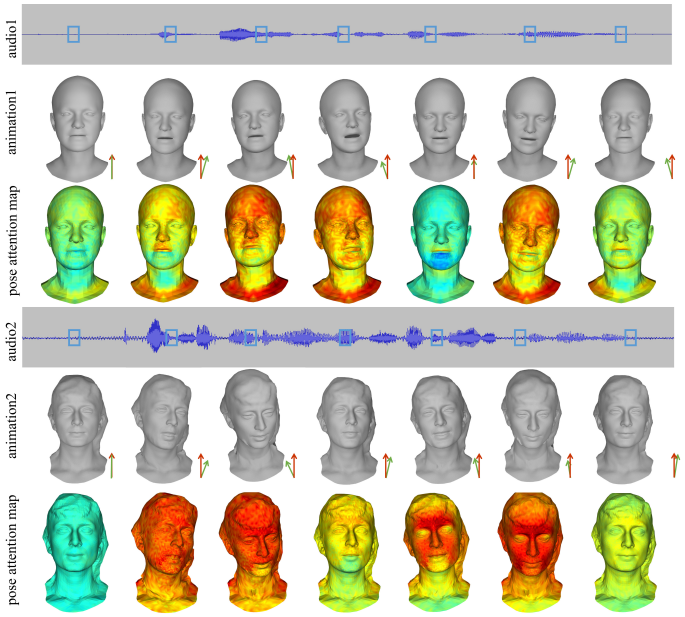


Fig. 7. Synthesized head meshes with audio-related poses and the corresponding pose-vertices attention maps. The visual pose attention map is obtained from the last row of the attention map output by the third attention module, where each value in the last row is assigned to the corresponding vertex.

highlight the superiority of the proposed method in predicting movement details. Benefiting from the hierarchical feature extraction strategy and audio-vertices attention mechanism, the accuracy of the proposed method is significantly superior to other methods in detailed expression synthesis and eye movements. See the supplementary video for a more detailed comparison. An experiment is designed in Sec. IV-C to validate the effectiveness of the proposed hierarchical feature and attention mechanism respectively.

Table I further illustrates the quantitative errors around the lip and eye regions. Compared with the previous methods, the error of our approach is reduced by at least 0.19mm in terms of E_{vl}. The proposed method also gains competitive performance in the prediction of eye regions. The L2 error E_{ve} quantifies the absolute variance in comparison to the actual movement, whereas FDD and AFDD evaluate how well the predicted motion space aligns with the actual motion space in terms of distribution consistency. From Table I, we can find that the proposed method achieved superior performance in most cases.

In addition to the closed-form metrics, we also conducted subjective user study experiments to evaluate lip synchronization and realism on both datasets. We randomly select 20 pieces of audio, and generate the corresponding facial animation using all of the comparison methods. Then, a random pair of the synthesized videos are shown to the user to determine which method performs better on high-quality animation generation with lip synchronization and realism. 80 participants were invited to join the user study, and each user gave their judgment on each pair of videos. The final quantitative results are shown in Table II. It is easy to find that the proposed method outperformed other methods in terms of both lip synchronization and realism.

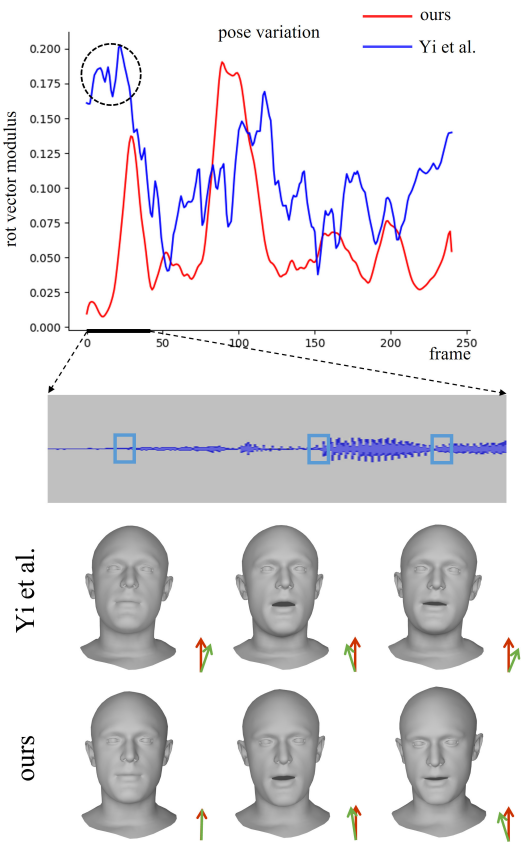


Fig. 8. Comparison of pose variation amplitude between ours and Yi *et al*. [37]. The horizontal axis represents the frames and the vertical axis represents the norm of the head pose rotation vector.

B. Pose attribute evaluation

Benefiting from the proposed pose attribute augmentation approach and the attention-based pose estimation algorithm, our method can predict realistic and diverse pose movements rather than merely "zero pose" as done in baselines [2]–[4]. From Fig. 7, we can find that the proposed method can generate realistic and fluent head pose movements with the given audio, and the changes of pose tend to align with tone transitions. In addition, the pose-vertices attention map is also demonstrated to evaluate the importance of each vertex to the final pose estimation. It is obvious that the pose estimation network has learned to pay appropriate attention to the vertices which play more important roles in pose prediction. For example, when there is an obvious pose variation, the estimation network will attend to most of the vertices on the head (2nd column in the last row). It implies that the global pose can be estimated by the salient vertices with larger attention values. When there is only a small pose variation around some regions, the network will pay less attention to these vertices, such as the chin regions in the 4th column of the last row of Fig. 7. The phenomenon is in line with human intuition and validates the effectiveness of the proposed pose estimation strategy to some extent.

Recently, Yi *et al*. [37] proposed a parametric method to synthesize talking face videos. Their intermediate process is

TABLE II
USER STUDY RESULTS FOR LIP SYNCHRONIZATION AND REALISM.

Competitors	VOCASET [2]		Multiface [5]	
	Lip Sync	Realism	Lip Sync	Realism
Ours $vs.$ VOCA	78.06	83.79	75.88	80.20
Ours $vs.$ GDPnet	79.11	82.61	-	-
Ours $vs.$ MeshTalk	95.35	96.96	71.47	72.75
Ours $vs.$ FaceFormer	57.77	61.08	94.55	96.76
Ours $vs.$ CodeTalker	57.42	59.04	80.29	83.14
Ours $vs.$ GT	47.42	41.34	41.96	43.24

TABLE III
POSE USER STUDY RESULTS.

Competitors	VOCASET [2]		Multiface [5]	
	Smoothness	Realism	Smoothness	Realism
Ours $vs.$ Yi et al.	81.13	71.17	84.91	73.58

able to predict the 3D head poses of 3DMM. Therefore, we compare the performance of head pose estimation of the proposed non-parametric method against the parametric method [37] in this section. The pose estimation results of both methods on a speech of about one second (with audio content 'Severe myopia') are shown in Fig. 8. We can find that the variation of the head poses in Yi *et al*. [37] is drastic in a very short interval as highlighted by the marked circle and its corresponding facial animations in Fig. 8. The head turns from right to left several times in less than 1 second, leading to an unacceptable experience. In contrast, our method produces a more consistent, downward facial motion to the left.

In addition to visual inspection, a subjective user study was conducted to compare the performance of both methods. Twenty video clips were randomly selected from each dataset. 53 users between the ages of 20 and 50 were invited to participate in the user study. For each individual, 40 pairs of short clips were randomly shown. Participants were asked to evaluate which video clip demonstrated superior smoothness and realism in head pose. The results of the study are presented in Table III. We can find that over 70% of participants preferred the smooth and realistic poses produced by the proposed technique, while others favored those of Yi *et al*. [37].

C. Ablation experiments

There are two key modules in the proposed framework, i.e., hierarchical feature and geometry-guided audio-vertices attention. In this section, we conduct experiments to validate the effectiveness of each module. To eliminate the influence of some unnecessary factors, we removed the pose module during the ablation experiments. See Supplementary Material for more details of ablation experiments on the pose audio.

As discussed in Sec. III, global feature G encodes the holistic attribute of audio, and is adopted in most of the previous work [2]–[4]. Local features L indicate the localized vertex-wise movement under the global attribute G. In this paper, we propose to build hierarchical audio features (LG) by combining the global attribute feature G and the vertex-wise local latent movement features L. We conduct two groups of experiments for various features with global MLP and the proposed geometry-guided audio-vertices attention (GAVA) module. On the one hand, we evaluate the performance of different features under the same mapping network. On the other hand, we compare the performance of MLP and the proposed GAVA with the same features.

TABLE IV
ABLATION EXPERIMENTAL RESULTS.

	VOCASET [2]		Multiface [5]	
	E_{vl}	E_{ve}	E_{vl}	E_{ve}
L+MLP	5.739	1.972	4.962	3.111
G+MLP	6.571	2.112	5.785	3.219
LG+MLP	5.814	1.969	5.081	3.177
L+GAVA	<u>5.602</u>	<u>1.915</u>	<u>4.662</u>	<u>3.045</u>
G+GAVA	6.393	2.084	5.596	3.572
LG+SA	5.718	1.997	4.849	3.138
LG+GAVA	**5.322**	**1.864**	**4.495**	**3.030**

The qualitative results are shown in Fig. 9. From the visual inspection and error maps, we can find that method G+MLP can capture the rough global movement but cannot produce detailed expressions, especially around mouth regions as shown in magnified views. [1]. The global feature is copied N times in the proposed GAVA module, and cannot produce discriminative features for each vertex. Therefore, G+GAVA cannot generate meaningful expressions. Compared with global features, local features (L+MLP) can produce more detailed expressions, such as mouth movements. L+GAVA can alleviate the artifacts generated by global MLP with the proposed attention to geometry prior, such as the wrinkle in the cheek. However, local features cannot capture the global semantics well, such as the range of mouth opening in the first frame. Compared with the local feature or global feature, the proposed hierarchical feature (LG+MLP) can produce audio-related global movement with detailed expressions. Benefiting from the geometry-prior attention mechanism, LG+GAVA can generate more realistic expressions with geometry consistency, such as the corner of mouth regions in the third frame.

The quantitative measurements are shown in Table IV. We calculate the maximum value of the vertex-to-vertex squared error in the lip and eye regions per frame and use the average of the maximum values across all frames to evaluate the error. The quantitative results are in line with the visual analysis and demonstrate the effectiveness of the proposed hierarchical feature and GAVA module.

In this experiment, we also evaluate the performance of the proposed GAVA against traditional self-attention (SA). In contrast to SA, our proposed GAVA uses the adjacency matrix of the mesh to constrain the span of attention, while SA does not take into account the local manifold structure of 3D models. Fig. 10 demonstrates the synthesized results by GAVA and SA with the same hierarchical feature input. The method based on SA achieves similar performance as LG+MLP in both visual inspections (corners of the mouth in the third frame) and quantitative metrics (Table IV) due

[1]It is noted that VOCA and GDPNet can generate smooth expressions with global features and MLP due to that they both use some prior knowledge of facial movements. VOCA uses the PCA coefficient of facial movements to initialize the decoder, and GDPnet constrains the consistency of intermediate features with those of the facial mesh autoencoder reconstruction network.

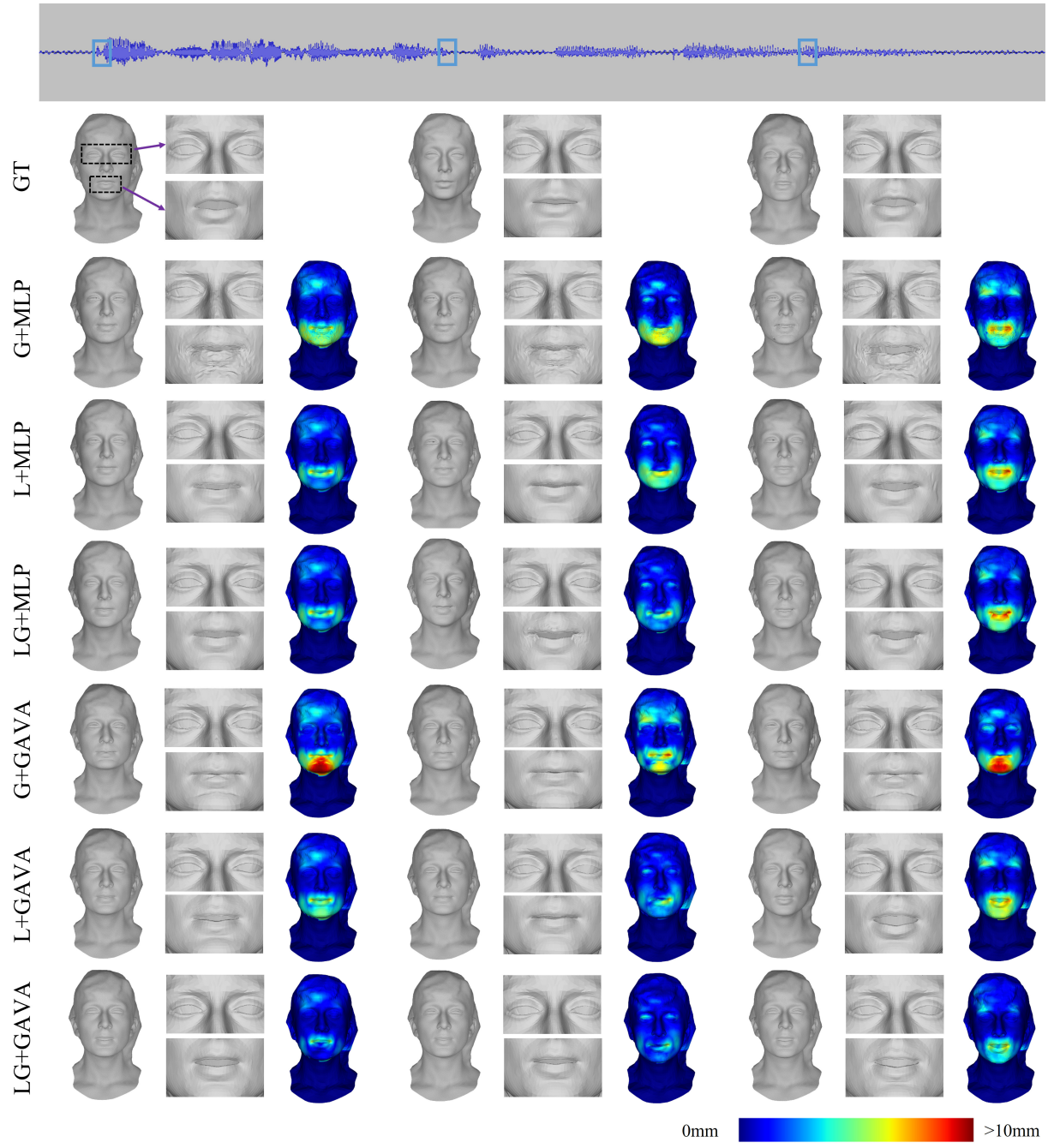


Fig. 9. Effectiveness of the proposed hierarchical feature and GAVA module. The color maps give the distribution of vertex-to-vertex distance errors (unit: millimeter).

to its global mechanism. Instead, the proposed GAVA can produce more consistent animation (1st frame) and detailed and realistic geometric deformation (corners of mouth and chin regions) by incorporating geometry priors into the attention module. From the attention map, we can also find that GAVA can capture more accurate and localized attention against the blind SA. To further validate the performance of GAVA, we show the attention maps at different layers of both methods in Fig. 11. We can find that the proposed GAVA is indeed a geometry-guided propagation process.

D. Generalization

To validate the robust performance of the proposed method, the synthesized results based on noisy audio signals and generalization across unseen subjects and across languages are shown in the supplementary document.

V. Conclusion and Future Work

In this paper, we present a novel approach for pose-aware life-like 3D talking face synthesis. A novel hierarchical feature integrated by a global attribute feature and a series of vertex-wise local latent movement features is proposed to capture

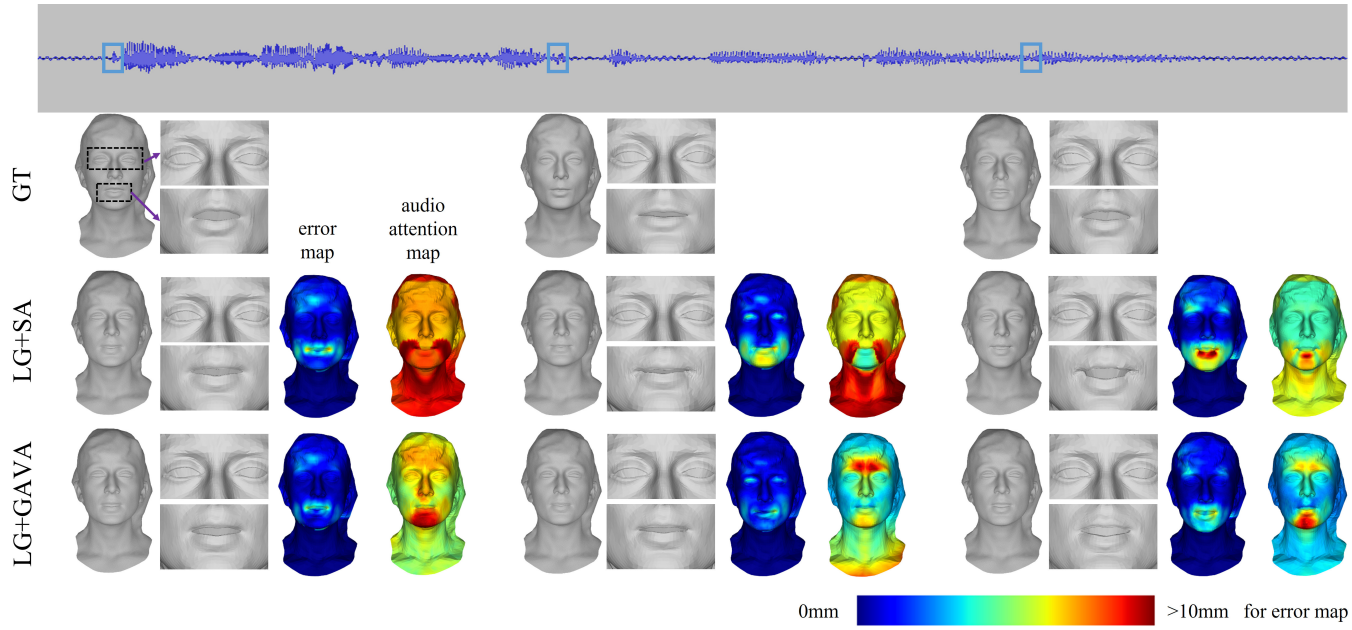


Fig. 10. Comparisons of SA and GAVA modules. The color maps give the distribution of vertex-to-vertex distance errors (unit: millimeter) and the attention maps show the motivation amplitude distribution of correspondence mesh vertices. The audio attention map is derived from the first N rows and N columns of the attention map generated by the final attention module. In this process, the maximum values of each column are attributed to their corresponding vertices, which are then depicted in different hues corresponding to these values.

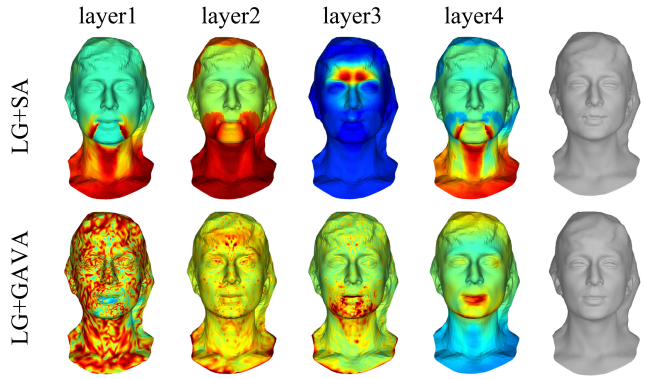


Fig. 11. Audio attention output in different layers.

more intricate facial expressions. To generate more realistic and geometry-consistent facial animation, we propose a novel geometry-guided audio-vertices attention module. Finally, to accomplish pose-aware animation, we expand the existing database with an additional pose attribute by the proposed 3D pose estimation method. Quantitative and qualitative experiments validate the effectiveness of the proposed method on realistic expression and head movements against state-of-the-art methods. A limitation of the proposed method is that it does not explicitly incorporate emotion as input to guide the generation of emotional animation. This might cause the creation of fixed sequences of actions with a given audio, leading to the lack of diverse outcomes. In the future, we anticipate making significant advancements in the field of audio-driven emotional 3D face animation.

ACKNOWLEDGMENTS

The work was funded by Natural Science Foundation of China (NSFC) under Grant 62172198, 61762064, Key Project of Jiangxi Natural Science Foundation 20224ACB202008, Key R&D Plan of Jiangxi Province 20232BBE50022, Project of Academic and Technical Leaders in Major Disciplines in Jiangxi Province 20232BCJ22001, and the Opening Project of Nanchang Innovation Institute, Peking University.

REFERENCES

[1] R. Zhen, W. Song, Q. He, J. Cao, L. Shi, and J. Luo, "Human-computer interaction system: A survey of talking-head generation," *Electronics*, vol. 12, no. 1, p. 218, 2023.

[2] D. Cudeiro, T. Bolkart, C. Laidlaw, A. Ranjan, and M. J. Black, "Capture, learning, and synthesis of 3D speaking styles," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 101–10 111.

[3] J. Liu, B. Hui, K. Li, Y. Liu, Y.-K. Lai, Y. Zhang, Y. Liu, and J. Yang, "Geometry-guided dense perspective network for speech-driven facial animation," *IEEE Transactions on Visualization and Computer Graphics*, 2021.

[4] A. Richard, M. Zollhöfer, Y. Wen, F. De la Torre, and Y. Sheikh, "MeshTalk: 3D face animation from speech using cross-modality disentanglement," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1173–1182.

[5] C.-h. Wuu, N. Zheng, S. Ardisson, R. Bali, D. Belko, E. Brockmeyer, L. Evans, T. Godisart, H. Ha, A. Hypes, T. Koska, S. Krenn, S. Lombardi, X. Luo, K. McPhail, L. Millerschoen, M. Perdoch, M. Pitts, A. Richard, J. Saragih, J. Saragih, T. Shiratori, T. Simon, M. Stewart, A. Trimble, X. Weng, D. Whitewolf, C. Wu, S.-I. Yu, and Y. Sheikh, "Multiface: A dataset for neural face rendering," in *arXiv*, 2022. [Online]. Available: https://arxiv.org/abs/2207.11243

[6] C. Sheng, G. Kuang, L. Bai, C. Hou, Y. Guo, X. Xu, M. Pietikäinen, and L. Liu, "Deep learning for visual speech analysis: A survey," *arXiv preprint arXiv:2205.10839*, 2022.

[7] L. Chen, R. K. Maddox, Z. Duan, and C. Xu, "Hierarchical cross-modal talking face generation with dynamic pixel-wise loss," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[8] X. Wang, Q. Xie, J. Zhu, L. Xie, and O. Scharenborg, "AnyoneNet: Synchronized speech and talking head generation for arbitrary persons," *IEEE Transactions on Multimedia*, 2022.

[9] H. Zhou, Y. Liu, Z. Liu, P. Luo, and X. Wang, "Talking face generation by adversarially disentangled audio-visual representation," 2018.

[10] A. Jamaludin, J. S. Chung, and A. Zisserman, "You said that?: Synthesising talking faces from audio," *International Journal of Computer Vision*, 2019.

[11] Z. Ye, M. Xia, R. Yi, J. Zhang, Y.-K. Lai, X. Huang, G. Zhang, and Y.-j. Liu, "Audio-driven talking face video generation with dynamic convolution kernels," *IEEE Transactions on Multimedia*, 2022.

[12] H. Zhou, Y. Sun, W. Wu, C. C. Loy, X. Wang, and Z. 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*, 2021, pp. 4176–4186.

[13] S. Wang, L. Li, Y. Ding, and X. Yu, "One-shot talking face generation from single-speaker audio-visual correlation learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, 2022, pp. 2531–2539.

[14] R. Huang, P. Lai, Y. Qin, and G. Li, "Parametric implicit face representation for audio-driven facial reenactment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12 759–12 768.

[15] E. R. Chan, C. Z. Lin, M. A. Chan, K. Nagano, B. Pan, S. De Mello, O. Gallo, L. J. Guibas, J. Tremblay, S. Khamis *et al.*, "Efficient geometry-aware 3D generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 123–16 133.

[16] S. Zhang, Z. Wu, H. M. Meng, and L. Cai, "Facial expression synthesis based on emotion dimensions for affective talking avatar," in *Modeling machine emotions for realizing intelligence*. Springer, 2010, pp. 109–132.

[17] X. Li, Z. Wu, H. M. Meng, J. Jia, X. Lou, and L. Cai, "Expressive speech driven talking avatar synthesis with DBLSTM using limited amount of emotional bimodal data," in *Interspeech*, 2016, pp. 1477–1481.

[18] H. X. Pham, Y. Wang, and V. Pavlovic, "End-to-end learning for 3D facial animation from speech," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, 2018, pp. 361–365.

[19] Y. Kim, S. An, Y. Jo, S. Park, S. Kang, I. Oh, and D. D. Kim, "Multi-task audio-driven facial animation," in *ACM SIGGRAPH 2019 Posters*, 2019, pp. 1–2.

[20] C. Zhang, S. Ni, Z. Fan, H. Li, M. Zeng, M. Budagavi, and X. Guo, "3D talking face with personalized pose dynamics," *IEEE Transactions on Visualization and Computer Graphics*, 2021.

[21] C.-Y. Wu, K. Xu, C.-C. Hsu, and U. Neumann, "Voice2mesh: Cross-modal 3d face model generation from voices," *arXiv preprint arXiv:2104.10299*, 2021.

[22] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 1999, pp. 187–194.

[23] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3D face model for pose and illumination invariant face recognition," in *IEEE international conference on advanced video and signal based surveillance*. Ieee, 2009, pp. 296–301.

[24] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, "Learning a model of facial shape and expression from 4D scans," *ACM Trans. Graph.*, vol. 36, no. 6, pp. 194–1, 2017.

[25] A. Hussen Abdelaziz, B.-J. Theobald, P. Dixon, R. Knothe, N. Apostoloff, and S. Kajareker, "Modality dropout for improved performance-driven talking faces," in *Proceedings of the 2020 International Conference on Multimodal Interaction*, 2020, pp. 378–386.

[26] Y. Liu, F. Xu, J. Chai, X. Tong, L. Wang, and Q. Huo, "Video-audio driven real-time facial animation," *ACM Transactions on Graphics (TOG)*, vol. 34, no. 6, pp. 1–10, 2015.

[27] H. X. Pham, S. Cheung, and V. 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*, 2017, pp. 80–88.

[28] D.-Y. Huang, E. Chandra, X. Yang, Y. Zhou, H. Ming, W. Lin, M. Dong, and H. Li, "Visual speech emotion conversion using deep learning for 3D talking head," in *Proceedings of the Joint Workshop of the 4th Workshop on Affective Social Multimedia Computing and first Multi-Modal Affective Computing of Large-Scale Multimedia Data*, 2018, pp. 7–13.

[29] A. Richard, C. Lea, S. Ma, J. Gall, F. De la Torre, and Y. Sheikh, "Audio- and gaze-driven facial animation of codec avatars," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 41–50.

[30] A. Lahiri, V. Kwatra, C. Frueh, J. Lewis, and C. Bregler, "LipSync3D: Data-efficient learning of personalized 3D talking faces from video using pose and lighting normalization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 2755–2764.

[31] Y. Fan, Z. Lin, J. Saito, W. Wang, and T. Komura, "FaceFormer: Speech-driven 3D facial animation with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 770–18 780.

[32] J. Xing, M. Xia, Y. Zhang, X. Cun, J. Wang, and T.-T. Wong, "CodeTalker: Speech-driven 3D facial animation with discrete motion prior," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12 780–12 790.

[33] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates *et al.*, "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.

[34] Y. Zhou, X. Han, E. Shechtman, J. Echevarria, E. Kalogerakis, and D. Li, "MakeItTalk: speaker-aware talking-head animation," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 6, pp. 1–15, 2020.

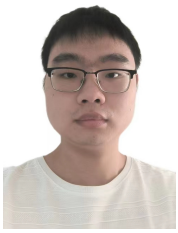
[35] Y. Feng, H. Feng, M. J. Black, and T. Bolkart, "Learning an animatable detailed 3D face model from in-the-wild images," *ACM Transactions on Graphics (ToG)*, vol. 40, no. 4, pp. 1–13, 2021.

[36] G. Patané and M. Spagnuolo, "Heat diffusion kernel and distance on surface meshes and point sets," *Computers & Graphics*, vol. 37, no. 6, pp. 676–686, 2013.

[37] R. Yi, Z. Ye, J. Zhang, H. Bao, and Y.-J. Liu, "Audio-driven talking face video generation with learning-based personalized head pose," *arXiv preprint arXiv:2002.10137*, 2020.

VI. Biography Section

Bo Li received the Ph.D. degree in 2008 in computational mathematics, Dalian University of Technology (DUT), Dalian, China. Now he is the professor in School of Mathematics and Information Science of Nanchang Hangkong University. His current research interests include the areas of image processing and computer graphics.



Xiaolin Wei received the B.S. degree in Computer Science and Technology from Sichuan University, China, in 2016. He is now a Master's degree student at Nanchang Hangkong University, China. His current research interests include computer graphics and deep learning.

Bin Liu received the Ph.D. degree in computational mathematics from Dalian University of Technology, Dalian, China, in 2021. He is currently a lecturer with the School of Mathematics and Information Science, Nanchang Hangkong University, NanChang, China. His current research interests include geometric processing and deep learning.

Zhifen He received the Ph.D. degree in the School of Mathematical Sciences from Nanjing Normal University, Nanjing, China, in 2015. She is currently an associate professor with the School of Mathematics and Information Science, Nanchang Hangkong University, Nanchang, China. Her current research interests include machine learning and computer vision.

Junjie Cao is an associate professor in School of Mathematical Sciences at Dalian University of Technology, P.R. China. He received the Ph.D. degree in computational mathematics from Dalian University of Technology in 2010. His research interests include point cloud processing, human body modeling, and related computer graphics and vision problems.

Yu-Kun Lai received his bachelor and Ph.D. degrees in computer science from Tsinghua University in 2003 and 2008, respectively. He is currently a Professor in the School of Computer Science & Informatics, Cardiff University. His research interests include computer graphics, geometry processing, image processing and computer vision. He is on the editorial boards of *IEEE Transactions on Visualization and Computer Graphics* and *The Visual Computer*.