AU-Aware Dynamic 3D Face Reconstruction from Videos with Transformer

Chenyi Kuang[1], Jeffrey O. Kephart[2], Qiang Ji[1]
[1] Rensselaer Polytechnic Institute, [2] IBM Thomas J. Watson Research Ctr.
{kuangc2,jiq}@rpi.edu, kephart@us.ibm.com

Abstract

In spite of the significant progresses in monocular or multi-view image based 3D face reconstruction research, recovering 3D faces from videos, which contains rich dynamic information of facial motions, still remains as a highly challenging problem. First, most prior works fail to generate accurate and stable 3D faces on videos, especially for recovering subtle expression details. Furthermore, existing dynamic reconstruction approaches have not fully considered the temporal dependency of facial expression transitions, which is based on the dynamic muscle activation system under a local region of the skin. To tackle the aforementioned challenges, we present a framework for dynamic 3D face reconstruction from monocular videos, which can accurately recover 3D facial geometrical representations for facial action unit (AU). Specifically, we design a coarse-to-fine framework, where the "coarse" 3D face sequences are generated by a pre-trained static reconstruction model; and the "refinement" is performed through a Transformer-based network. We design 1) a Temporal Module used for modeling temporal dependency of facial motion dynamics; 2) an Spatial Module for modeling AU spatial correlations from geometry-based AU tokens; 3) feature fusion for simultaneous dynamic facial AU recognition and 3D expression capturing. Experimental results show the superiority of our method in generating AU-aware 3D face reconstruction sequences both quantitatively and qualitatively.

1. Introduction

Human facial motion analysis has been an emerging field of study in computer vision, psychology and cognitive science, which has a great potential in a wide range of application fields, such as human-computer interaction, movie animation, video games, communication, etc. A plethora of computer vision studies have been conducted over the past decades to capture and analyze facial motions from images or videos. With the first 3D morphable model (3DMM) proposed by Blanz et al. [6], 3DMM-based face modeling and reconstruction have gained sustained attention, which focus on recovering 3D facial meshes with plausible identity and expression details given 2D images or videos. In general, 3D face reconstruction from 2D data via a pre-constructed 3DMM can be treated as a regression problem of estimating 3DMM parameters of camera, head pose, identity, expression, texture and illumination for different people and environment. In earlier researches, 3DMM fitting algorithms are proposed to optimize the 3D parameters sequentially using detected 2D features, such as [19,21,44]. As the development of deep learning models and algorithms, many regression-based methods have been designed for solving the 3D parameters simultaneously, such as 3DDFA [20, 63, 64], RingNet [47] and Deep3DFaceRecon [11]. Such deep-learning based methods can produce well-aligned 3D faces to the given image, but still lack sufficient expressiveness to capture subtle or extreme expressions. Responding to the increasing demand for generating high quality 3D faces that entails person-specific and expression-specific geometrical details, recent researches have made efforts to create fine-grained expression bases [36, 56] or predict high-frequency displacement maps [2, 3, 17, 56] to refine expression details. However, image-based 3D face reconstruction algorithms mentioned above may fail to apply well on video data in a frame-by-frame manner, since it may lead to unsteady, unnatural and jittered results, especially when there are frequent expression transitions or mouth movement in daily communication videos. A few works have provided solutions to video-based 3D face reconstruction, by predicting common identity parameters shared among frames [50] and enforcing temporal smoothness of expression parameters between consecutive frames [26, 50]. In [18], a dynamic neural radiance fields model conditioned on expression latent codes is learned to represent 4D facial avatars from an input face video. However, the training methodologies utilized in existing methods ignore the inherent 3D nature of facial dynamics, which is driven by the muscle contraction or relaxation under the skin. Therefore, they still struggle to capture

Code will be available at: https://github.com/kuangcy1998/AU-D3DFace facial expressions dynamics in time, such as the associativity and cooperativity between local expressions.

According to Facial Action Coding System (FACS) [14], which defines a taxonomy of facial movements in a local area as 32 action units (AUs) driven by muscle contraction, all observable facial movements can be expressed by the combinations of multiple AUs. In this paper we propose a framework for dynamic 3D face reconstruction from videos, which can recover AU-interpretable geometrical details. We design a Transformer-based structure, which can autoregressively produce expressive 3D facial mesh given former frames and AU labels. Our major original contribution can be summarized as:

- we propose the first AU-aware dynamic 3D face reconstruction framework that learns facial motion dynamics and spatial correlations from combined geometric and appearance features.
- we combine a Transformer-based Temporal Module and Spatial Module for dynamic 3D AU representation, which can encode the long-term AU occurrence dependency and the history of 3D face motions to autoregressively predict a sequence of 3D facial expressions. It achieves temporally stable and accurate 3D face capturing on videos.
- Our model simultaneously generate dynamic 3D face sequences and 3D AU activation trajectories, and our model achieves SOTA 3D AU recognition accuracy and better reconstruction accuracy in expressions.

2. Related Works

2.1. Dynamic 3D Face Reconstruction

Capturing and predicting a dynamic process of 3D facial motion from RGB videos remains as a highly challenging problem, where substantial efforts in computer vision area have been dedicated to tackle this problem. FML [50] method learns person-specific identity model and appearance model from each input sequence. The Head-to-Head approach [26] trains a model for 3D face tracking and video-based rendering, and the latter model performs face video synthesis in a recurrent way. In [18], they present dynamic neural radiance fields for reconstructing 4D faces in short portrait video sequence of a person. The 3D geometry of faces is encoded as a low-dimensional morphable model that provides explicit control over pose and expressions over time. Learning the dynamic neural radiance fields is the conditioned on the dynamically changing facial expressions and later volume rendering is leveraged to generate the synthesized portrait videos. New applications arise in this area. For instance, generating a 3D talking head from videos has become a emerging research topic, which emphasize on capturing the expression changes, especially mouth movement in a speech video. To generate accurate and natural talking patterns of the mouth in 3D space, Ye et al. [57] proposed a dynamic network that takes video and audio together as a multi-modal input, and regress for the identity and expression coefficients of a 3DMM. In [58], they train a general mapping from the input audio to 3DMM-based facial expression and head pose parameters. Combining them with the face video of a person, this model can generate high-quality 3D talking head and a synthesized video through rendering. In [61], the representations of talking human faces video are modularized into the spaces of speech content, head pose, and identity respectively, which are implicitly learned in a construction-based framework.

2.2. 3D AU Representation & Recognition

Facial AU recognition from RGB images or videos have been extensively studied in computer vision area. In [7], Cao et al. collected a 3D face database – FaceWarehouse, that contains facial scans of 150 individuals with 19 kinds of expressions. A set of 3D blendshapes are constructed, with each representing 3D skin deformation of a facial AU. Similar blendshapes models are proposed in FaceScape [56] and ICT [33]. The major advantage of 3D facial blendshapes over PCA bases lies in the semantic meaning of blendshapes related to AU. Therefore, blendshape models can be used for controllable facial expression generation and animation [38, 49, 51, 53]. Except for 3D AU synthesis, AU detection from 3D data like 3D scans, 3D point clouds, or 3DMMs have been actively studied by researchers. Given a target 3D mesh or scan, one of the most intuitive way is to train classifiers based on the extracted mesh surface features for 3D AU detection [4, 10, 25, 31, 48, 62]. Similarly, for 3D point cloud data, Reale et al. [45] trained a network to directly extract 3D point cloud features and support AU detection. Ariano et al. [1] propose a method of 3D AU detection by using 3DMM coefficients. However, little work has been devoted to connecting the AU characteristics with the task of 3D face reconstruction.

2.3. Transformers in Vision Tasks

The Transformer [52] model, as a strong alternative model to RNNs and CNNs, has shown its promising performance in the area of Natural Language Processing (NLP), due to the attention mechanism. Vision Transformer (ViT) [13] is first introduced to perform image classification task [32]. Since then, transformer-based structure have been applied to various vision tasks such as object detection [41], image segmentation [12], image generation [24, 46], etc. In addition, transformers have been exploited in learning the representation of 3D human bodies for 3D motion reconstruction [35], prediction [39], tracking [23, 65] or synthesis [5, 43]. For face-related tasks, Transformers have

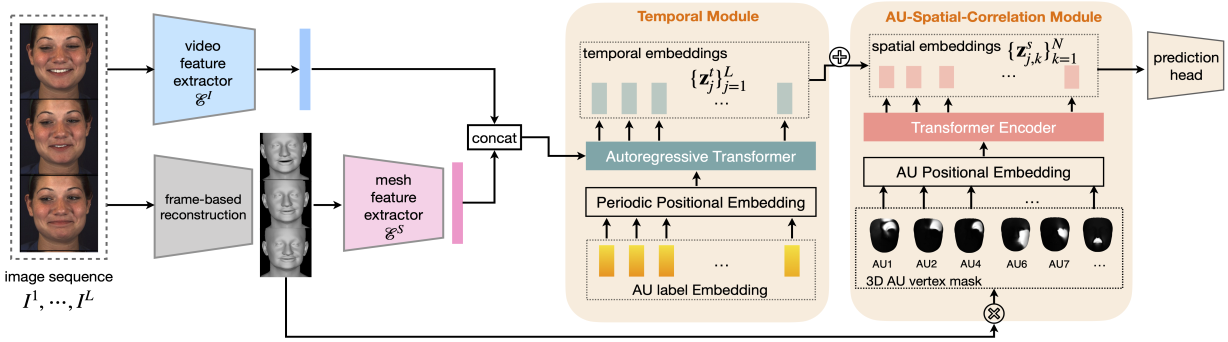
video feature extractor $\mathscr{E}^I$
frame-based reconstruction
mesh feature extractor $\mathscr{E}^S$
image sequence $I^1, \cdots, I^L$
concat
Temporal Module
temporal embeddings $\{\mathbf{z}_j^t\}_{j=1}^L$
Autoregressive Transformer
Periodic Positional Embedding
AU label Embedding
AU-Spatial-Correlation Module
spatial embeddings $\{\mathbf{z}_{j,k}^s\}_{k=1}^N$
Transformer Encoder
AU Positional Embedding
AU1
AU2
AU4
AU6
AU7
3D AU vertex mask
prediction head

Figure 1. Overview of our AU-aware dynamic 3D face reconstruction method: (1) feature encoding module: (Gray) pre-trained frame-based 3D face reconstruction model from DECA [17], frozen during training, generate coarse 3D mesh sequences from images. (Blue) Video feature extractor. (Pink) Mesh feature extractor. (2) The extracted features will be feed into a Temporal Module, which consists of learnable AU label embedding layer, Periodic Positional Embedding layer and an Autoregressive Transformer module. The output will be the temporal embeddings $\boldsymbol{z}_j^t$ for each frame. (3) The coarse mesh sequence are also feed into a AU-Spatial-Correlation Module, which is designed for learning the spatial correlations among AUs from 3D geometry. The output will be spatial embeddings $\boldsymbol{z}_{j,k}^s$ for each AU indexed by k. (4) The temporal and spatial embeddings $\boldsymbol{z}_j^t$ and $\boldsymbol{z}_{j,k}^s$ are combined for 3D reconstruction head and 3D AU classification head.

shown its superiority over other model structures in image-based [22,28,30,55] and video-based [1,60] expression/AU recognition. In spite of the progresses of applying Transformers on 3D human data, such as Mesh Transformer [35] and point-cloud Transformer [15], using Transformer for capturing 3D facial expression dynamics is not fully studied yet. In [16], a Transformer-based seq2seq structure is designed to autoregressively animate a sequence of 3D face meshes from an input audio file. Besides, a lightweight Transformer-based framework is proposed in [31] to perform multi-modal 2D+3D facial expression recognition. In [31], a single 3D facial scan together with a RGB image are feed into the transformer model to fuse multi-modal features. Our proposed method, which differs from [16, 31], aims at recovering dynamic 3D facial expression from videos with an autoregressive Transformer structure, through encoding the context of 3D facial movement.

3. Method

3.1. Method overview

Our model conducts a dynamic refinement for frame-based 3D face reconstruction results on videos. As show in Fig. 1, our framework can be divided into three parts. Given input video frames$\{I_1, \cdots, I_L\}$, we first apply a pre-trained 3D face reconstruction model frame-by-frame to generate the coarse 3D mesh sequences $\boldsymbol{S} = \{S_1, \cdots, S_L\}$. The frame-based reconstruction model does not consider facial expression dynamics in time and cannot present subtle motions in terms of 3D facial AU intensities. To address these problems and produce the refined 3D reconstruction sequences $\hat{\boldsymbol{S}} = \{\hat{S}_1, \cdots, \hat{S}_L\}$ on videos, we design a Transformer-based structure that takes video frames, coarse mesh sequences and AU annotation sequences as input. Since we are targeting at capturing both the temporal dependencies of facial motions and the spatial correlations among different AUs, the designed Transformer module consists of a Temporal model in an autoregressive manner, and a spatial model, which implicitly represents the spatial correlations among AUs from the face geometry. The work flow of our method contains the following steps:
(1) Data preparation: obtaining coarse 3D face meshes with a static frame-based 3D face reconstruction model;
(2) Extracting frame-based appearance features and coarse geometry features with a backbone model;
(3) Predicting 3D temporal embeddings on sequences with Autoregressive Transformer layers; Predicting 3D spatial embeddings with spatial Transformer layers.
(4) Combining the 3D temporal and spatial embeddings for the final prediction head, which contains a classification head for predicting AU occurrence probabilities and a regression head for generating refined 3D mesh sequences.
We introduce each step in detail in Section 3.2 $\sim$ 3.5.

3.2. Feature Extraction

Coarse 3D Face Reconstruction We apply the pre-trained DECA [17] model for static 3D face reconstruction on every single video frame. The DECA encoder (a ResNet-50 CNN) predicts a set of 3D parameters given an image I, including the identity parameter $\boldsymbol{\beta} \in \mathbb{R}^{100}$, expression parameter $\boldsymbol{\psi} \in \mathbb{R}^{50}$, albedo parameter $\boldsymbol{\alpha} \in \mathbb{R}^{50}$, lighting parameter $\boldsymbol{l} \in \mathbb{R}^{27}$, pose parameter $\boldsymbol{\theta} \in \mathbb{R}^{6}$ and camera parameter $\boldsymbol{c} \in \mathbb{R}^{3}$. With a differentiable rendering layer, the reconstructed 3D face can be projected to image plane to generate a synthetic face image $\tilde{I}$. Our model aims at optimizing for the 3D expression part by exploiting the expression dynamics in videos.

Feature encoding For an input sequence of images $\{I_j\}_{j=1}^{L}$, we refer to the Video Swin Transformer [37] as the backbone model, denoted as $\mathcal{E}^{\mathcal{I}}$ for extracting frame-based appearance features $\{f_j^I\}_{j=1}^{L}, f_j^I \in \mathbb{R}^{H \times W \times C_I}$. For the coarse 3D mesh sequences $\{S_j\}_{j=1}^{L}$ generated in Section 3.2, we define a motion encoding layer $\mathcal{E}^{\mathcal{S}}$ to extract only the expression features in 3D space, defined as

$$f_j^S = \mathcal{E}^S(S_j - \bar{S}; \Phi_1) \tag{1}$$

where $\bar{S}$ represents the subject neutral face. Then the temporal feature and spatial feature are concatenated and feed to the Temporal Module: $f_j = \text{concat}(flatten(f_j^I), f_j^S)$.

3.3. Temporal Transformer

A shown in Fig. 1, the concatenated video feature and 3D mesh feature are feed into the Temporal Module, which is for capturing the temporal dependencies of facial motions in terms of AU occurrence. The Temporal Module consist of three parts: a learnable AU-label Embedding layer, a Periodic Positional Embedding layer for injecting frame order information, and an Autoregressive Transformer Module for predicting temporal representations of AUs.

AU-label Embedding Given the AU label $\boldsymbol{y}^{N \times 1}$ for N AUs, we project the AU labels of each frame $\boldsymbol{y}_j, j \in \{1, \cdots, L\}$ to a d-dimension vector $\boldsymbol{h}_j$ through a linear projection function, defined as:

$$\boldsymbol{h}_j = \begin{cases} \boldsymbol{W}^h \cdot \boldsymbol{y}_j + \boldsymbol{b}^h, j = 1, \cdots, L \\ \boldsymbol{b}^h, j = 0 \end{cases} \tag{2}$$

where $j = 0$ represents the begin token and $\boldsymbol{W}^h \in \mathbb{R}^{d \times N}$ and $\boldsymbol{b}^h \in \mathbb{R}^{d \times 1}$ represent the weight matrix and bias.

Periodic Positional Encoding Consider that the AU labels could be quite consistent in a sequence, we refer to the method in [16] to add a Periodic Positional Encoding (PPE) to the AU embedding vectors, indicating the temporal order. The PPE is expressed by the function below.

$$\begin{aligned} PPE_{(j,2i)} &= \sin((j \bmod P)/(10000)^{2j/d}) \\ PPE_{(j,2i+1)} &= \cos((j \bmod P)/(10000)^{2j/d}) \end{aligned} \tag{3}$$

where the i is the dimension index and P is a hyper-parameter defining the period. The AU embedding vector $\boldsymbol{h}_j$ will be added to the PPE before feeding them to the Autoregressive Transformer layer, expressed as:

$$\hat{\boldsymbol{h}}_j = \boldsymbol{h}_j + PPE(j), j = 1, \cdots, L \tag{4}$$

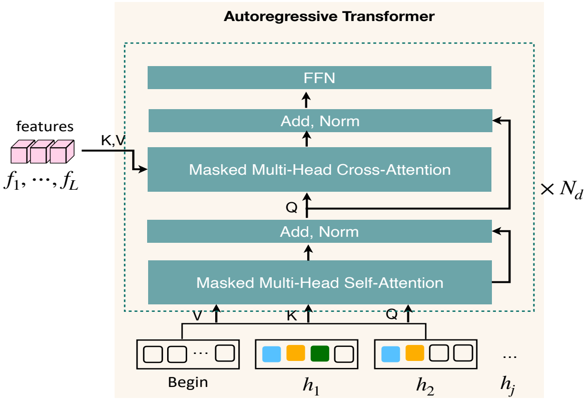
Autoregressive Transformer
FFN
Add, Norm
features
K,V
$f_1, \cdots, f_L$
Masked Multi-Head Cross-Attention
Q
$\times N_d$
Add, Norm
Masked Multi-Head Self-Attention
V
K
Q
Begin
h_1
h_2
h_j

Figure 2. Autoregressive Transformer in Temporal Module.

Autoregressive Transformer To fuse the facial AU dynamics with the encoded mesh and video features in a input sequence of length T, we refer to the transformer decoder architecture used in the GPT models and design an module that autoregressively predict the temporal embeddings for refining the 3D mesh. We formulate the problem as

$$p(\{\boldsymbol{z}_j^t\}_{j=1}^{L} | \{\hat{\boldsymbol{h}}_j, f_j\}_{j=1}^{L}) = \prod_{j=1}^{L} p(\boldsymbol{z}_j^t | \hat{\boldsymbol{h}}_{<j}, f_{<j}) \tag{5}$$

where $\boldsymbol{z}_j^t$ represent the output temporal embeddings for $j\text{−}th$ frame, and we model this distribution with the devised Autoregressive Transformer (denoted as $\mathcal{T}_{\Theta_1}^t$) as shown in Fig. 2. In each layer of $\mathcal{T}^t$, there is a Multi-Head self-attention layer (MHS) and a Multi-Head cross-attention (MHC) layer, inserted with residual connections and layer normalization(LN). The processing of our model $\mathcal{T}^t$ can be written as:

$$\begin{aligned} &\text{self-attn: } \hat{\boldsymbol{h}}_l^{(1)} = \text{LN}(\text{MHS}(Q^{\hat{\boldsymbol{h}}_{l-1}}, K^{\hat{\boldsymbol{h}}_{l-1}}, V^{\hat{\boldsymbol{h}}_{l-1}}) + \hat{\boldsymbol{h}}_{l-1}) \\ &\text{cross-attn: } \hat{\boldsymbol{h}}_l^{(2)} = \text{LN}(\text{MHC}(Q^{\hat{\boldsymbol{h}}_l^{(')}}, K^{\hat{\boldsymbol{h}}_l^{(1)}}, V^{\hat{\boldsymbol{h}}_l^{(1)}}) + \hat{\boldsymbol{h}}_l^{(1)}) \\ &\text{FFN: } \hat{\boldsymbol{h}}_l = \text{FFN}(\hat{\boldsymbol{h}}_l^{(2)}), l = 1, \cdots, N_d \\ &\text{Output:} \boldsymbol{z}_j^t = \hat{\boldsymbol{h}}_{N_d, j} \end{aligned} \tag{6}$$

where l is the layer index and we can concatenate N_d layers in total.

3.4. AU-Spatial-Correlation Module

In addition to the temporal dependencies encoded in the facial motion dynamics, we are also interested in modeling the spatial correlations of AUs in 3D geometry, for refining 3D facial motions based on a single or combinations of AUs. As mentioned in FACS [14], AUs can be positively or negatively correlated, such as *cheek raiser* (AU6) and *lip* *corner raiser* (AU12) are positively correlated. This relationship can be expressed more intuitively in 3D geometry, reflected as the overlapping or opposite vertex movement in a certain facial area. As shown in Fig. 1, we designed a Transformer-based structure as the spatial correlation module, denoted as $\mathcal{T}^S_{\Theta_2}$, that takes 3D local features as input and predict spatial embeddings $\boldsymbol{z}^s_j$ for each frame.

Geometry-based AU tokens The AU Spatial Correlation Module takes geometry-based AU tokens as input, which are generated by applying multiple pre-defined vertex masks $\{\boldsymbol{M}^{au}_k\}^N_{k=1}$ to the coarse 3D mesh $\{S_j\}^L_{j=1}$, where N is the number of AUs. The geometry-based AU-tokens $\boldsymbol{v}_k$ are expressed by

$$\boldsymbol{v}_k = \text{MLP}(\boldsymbol{M}^{au}_k \otimes (S_j - \bar{S})), k = 1, \cdots, N \quad (7)$$

where we use a MLP module to project dense vertex movement to N AU tokens.

Spatial Correlation Module We propose to apply a Transformer Encoder structure, which takes the discriminative AU tokens as input and models AU correlations implicitly with Multi-Head self attention layers. The processing of the spatial module is described below.

$$\begin{aligned}
&\text{self-attn: } \boldsymbol{v}^{(1)}_l = \text{LN}(\text{MHS}(Q^{\boldsymbol{v}_{l-1}}, K^{\boldsymbol{v}_{l-1}}, V^{\boldsymbol{v}_{l-1}}) + \boldsymbol{v}_{l-1})\\
&\text{FFN: } \hat{\boldsymbol{h}}_l = \text{FFN}(\boldsymbol{v}^{(1)}_l), l = 1, \cdots, N_d\\
&\text{Output: } \boldsymbol{z}^s_k = \boldsymbol{v}_{N_d,k}
\end{aligned} \quad (8)$$

3.5. 3D Face Refinement & AU classification

We combine the output temporal embedding $\boldsymbol{z}^t$ and $\boldsymbol{z}^s$ and feed them to subsequent regression head and classification head. The regression head, denoted as a motion decoder $\mathcal{D}$, is used for generating AU-aware 3D facial motions. As mentioned in Section 3.1, the refined 3D reconstruction sequence $\hat{\boldsymbol{S}} = \hat{S}_1, \cdots, \hat{S}_L$ will be produced by adding the predicted AU-based motions to the subject neutral face $\bar{S}$, expressed by

$$\hat{S}_j = \bar{S} + \mathcal{D}(\boldsymbol{z}^t_j + \sum^N_k \boldsymbol{z}^s_{j,k}; \boldsymbol{\theta}_3), j = 1, \cdots, L \quad (9)$$

In the mean time, the classification head is used to predict AU occurrence probability $\{p_k\}^N_{k=1}$ for each frame in the sequence, which is implemented as N linear projection layers.

$$\begin{aligned}
p_{j,k} = &\text{Sigmoid}[(\boldsymbol{W}^z_k)^T(\boldsymbol{z}^t_j + \boldsymbol{z}^s_{j,k})],\\
&k = 1, \cdots, N, j = 1, \cdots, L
\end{aligned} \quad (10)$$

3.6. Training and Inference

Training: During training, we adopt the teacher-forcing scheme for the autoregressive transformer, where it takes ground-truth AU label as input. The loss function for training the full model contains two part: the 3D reconstruction loss build upon weak supervision; the 3D AU classification loss, which is expressed as a weighted multi-label binary cross-entropy loss. Overall, we optimize the following loss function

$$L = \lambda_{img}L_{img} + \lambda_{lmk}L_{lmk} + \lambda_{smooth}L_{smooth} + \lambda_{au}L_{au} \quad (11)$$

with the photometric image loss L_{img}, the projected landmark loss L_{lmk}, the temporal smoothness loss L_{smooth} and the 3D AU classification loss L_{au}.

1. Photometric loss & landmark loss. With the predicted refined mesh $\hat{S}_j$ as mentioned in Eq. 9, we use the pose $\boldsymbol{R}_j$, camera $s, \boldsymbol{t}_j$, texture $\boldsymbol{\delta}_j$ and lighting parameter $\boldsymbol{\gamma}_j$ produced by the coarse reconstruction model to generate rendered facial image $\hat{I}_j$ and projected facial landmarks $\boldsymbol{c}_j$.

$$\begin{aligned}
&X_j = s * \boldsymbol{Pr} * \boldsymbol{R}_j * \hat{S}_j + \boldsymbol{t}_j\\
&\Rightarrow \hat{I}_j, \boldsymbol{c}_j = Renderer(X_j, \boldsymbol{\delta}_j, \boldsymbol{\gamma}_j)
\end{aligned} \quad (12)$$

Then L_{lmk} and L_{img} are calculated to measure the projected landmark error and the image intensity difference, which are expressed as:

$$L_{img} = \frac{1}{L}\sum^L_{j=1}\frac{A_j \odot \|I_j - \hat{I}_j\|_{1,1}}{\|A_j\|_{1,1}}, L_{lmk} = \frac{1}{L}\sum^L_{j=1}\|\boldsymbol{c}_j - \boldsymbol{c}^{gt}_j\|_2 \quad (13)$$

where A_j are pre-computed 2D skin mask for face region and $\boldsymbol{c}_j$ are ground-truth 2D landmark location.

2. Temporal smoothness loss: The temporal smoothness is adopted to generate smooth sequential result, and are expressed as the inter-frame difference between the temporal embedding.

$$L_{smooth} = \sum^{L-1}_{j=1}\|\boldsymbol{z}^t_j - \boldsymbol{z}^t_{j+1}\|_1 \quad (14)$$

3. 3D AU classification loss: we define a loss function based on the cross-entropy between the ground-truth label $y_{j,k}$ and predicted AU occurrence probability $p_{j,k}$:

$$L_{au} = -\frac{1}{L \times N}\sum^L_{j=1}\sum^N_{k=1}(y^{gt}_{j,k}log(p_{j,k}) + (1 - y^{gt}_{j,k})log(1 - p_{j,k})) \quad (15)$$

Inference During inference, we do not have access to the ground-truth AU labels, our model will autoregressively predict the AU probability and refined 3D face for the input sequence.

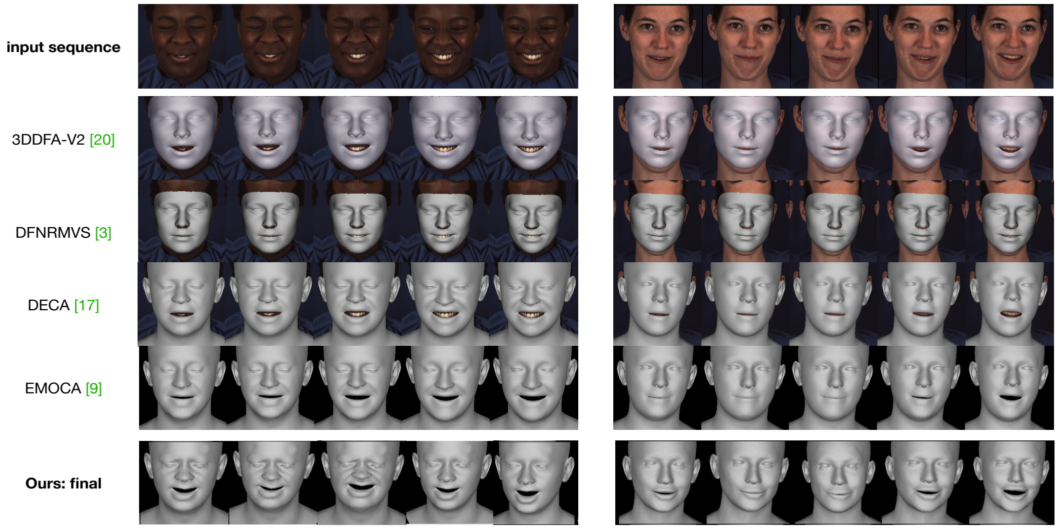

Figure 3. Reconstruction result comparison of our model and state-of-art face reconstruction models on input image sequences (5 frames are displayed): 3DDFA [20], DFNRMVS [3], DECA [17] and EMOCA [9] on BP4D [59] dataset.

4. Experiments

The proposed method is trained ans evaluated on three public benchmark dataset, BP4D [59], DISFA [40], Multiface [54]. The **BP4D** [59] dataset is a spontaneous database containing 328 sequences from 41 subjects performing different facial expressions. Each subject is involved in 8 expression tasks, and their spontaneous facial actions are coded by binary AU labels. Around 140k frames with AU occurrence labels are employed for training & evaluation. The **DISFA** dataset contains recorded videos for 27 subjects and we can extract around 130k frames. Each frame is annotated with AU intensity label ranged from 0 to 5. The annotations for 8 AUs are available. We follow the protocols used in [8,22] and select frames with AU intensity ≥ 2 as positive samples and the rest as negative. On BP4D and DISFA, subject-exclusive three-fold cross-validation experiment protocol is employed for AU detection evaluation. **Multiface** is a large-scale, multi-view dataset containing high-resolution synchronized 2D & 3D videos of facial performances collected from 13 subjects. As the full Multiface is very large, we use a small subset (10 expressions for 3 subjects, 30 videos in total) only for evaluating dynamic 3D reconstruction error.

Implementation details: We uniformly sample the image frames in all benchmark datasets and divide them into short clips of 20 frames in every two seconds, i.e., $L = 20$. So the frame rate is unified for all the data. For the two transformer-based model, we set the number of layers to 3, i.e., $N_d = 3$. The network is optimized using Adam with learning rate lr $= 5e - 5$. The network is trained for 20 epochs. Hyper-parameters used in Eq. 11 are set to $\lambda_{img} = 2, \lambda_{lmk} = 2, \lambda_{smooth} = 1, \lambda_{au} = 5$.

Evaluation protocol: We perform both qualitative and quantitative evaluation of the refined 3D face reconstruction quality. On Multiface, we use the provided 3D mesh sequences as ground-truth and find out the correspondence between the reconstructed mesh vertices to the closest ground-truth mesh surface. The reconstruction accuracy is evaluated in terms of vertex-to-plane mean square error (MSE) and standard deviation (SD). To validate the claim "AU-aware reconstruction", we also evaluate AU classification performance. We employ the F1-score for each AU as the metric, i.e., $F1 = \frac{2 \cdot R \cdot P}{R+P}$, where P is precision and R is recall, and compare with SOTA AU classification methods.

4.1. Qualitative Evaluation

We first show a visual comparison of 3D face reconstruction results on a short video clip (5 consecutive frames as displayed for each clip) in Fig. 3, comparing with existing methods including 3DDFA-v2 [20], DFNRMVS [3]DECA [17], EMOCA [9]. Our proposed method focus on recovering local facial motions caused by AU activation. As shown in Fig. 3, our dynamic reconstruction approach better captures subtle motions related to AUs, such as eye-brow frowning (clip1), nose wrinkling (clip1) and mouth dimpling (clip2). We provide more qualitative evaluation results on sequences in the supplementary material.

4.2. Quantitative Evaluation

We quantitatively evaluate the 3D reconstruction accuracy on Multiface (subset) and AU detection performance on the test fold of BP4D and DISFA. For reconstruction comparison, we compare with DECA [17] and EMOCA [9], since we all use the FLAME [34] mesh topology. For

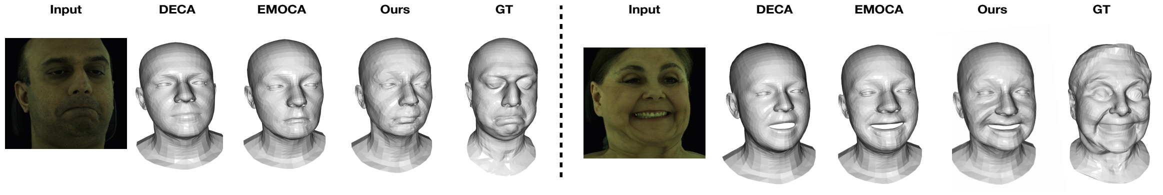

Figure 4. Comparison on Multiface [54] dataset: DECA [17], EMOCA [9] and our model are compared with ground-truth mesh.

Table 1. Reconstruction accuracy evaluation on Multiface.

Methods	Mean (mm)	SD
DECA [17]	1.865	1.405
EMOCA [9]	1.624	1.234
Ours: Spatial	1.587	1.245
Ours: Temporal + Spatial	**1.503**	1.127

AU recognition, we compare our model with different types of AU-detection methods, including geometry-based (DECA [17], EMOCA [9], Kuang et al. [27]), appearance-based (FAU [22], LP-Net [42], SRERL [29]) and combinatory ones (Biomechanics-AU [8]). The performances for Multiface, BP4D and DISFA are provided in Table. 1, Table. 2 and Table. 3. The 3D models DECA [17] and EMOCA [9] cannot be directly applied to AU detection, so we train a subsequent MLP block that takes reconstruction coefficients as input and predict the AU probability.

Comparisons on MultiFace We run our model and DECA [17], EMOCA [9] on 30 videos of Multiface with different expressions. With the predicted mesh sequences, we first apply the Iterative-Closest-Point(ICP) algorithm to find the correspondence between reconstructed and ground-truth meshes. Then we calculate the vertex-to-surface MSE and standard deviation (SD), as shown in Table. 1. With the same mesh topology, our Temporal + Spatial model achieves **uniformly** lower reconstruction error on different regions of the face (as SD is also lower), which indicates that integrating AU knowledge into the model helps to refine local geometric details in a dynamic expression. Compared to EMOCA, which is specially designed for expression recognition, our AU-aware model reduces the 3D error by 7.45%. We visualize the reconstruction result in Fig. 4.

Comparisons on BP4D We first compare the AU detection results for 12 AUs on BP4D dataset. As shown in Table. 2, we marked the highest F1 score for each AU and the average F1 score. Furthermore, we also specifically mark the dynamic models which take sequences as input, with a "*" prefix. For fair comparison, we provide two evaluations of our proposed model: (1) the model with Spatial Module only, which can be considered as a static model without considering the temporal dependency; (2) the dynamic model with both Temporal Module and Spatial Module. Among the listed existing methods, our model is similar to Biomechanics-AU [8] in terms of combining 3D mesh features and images features for dynamic AU prediction. On average, our final dynamic model achieves 1.9% improvement in F1 score compared to Biomechanics-AU [8]. Our model also show its superiority over appearance-based methods (SRERL [29], LP-Net [42], FAU [22]). Although our Spatial Module does not achieve SOTA performance compared to SOTA static model such as FAU [22], our final model is still superior to FAU [22], by further considering the temporal dependencies of AUs. Most importantly, we use the AU detection performance to illustrate that our model successfully integrate AU dynamics and correlations into the 3D face reconstructed process. We select three existing 3D face reconstruction models which focus on capturing accurate facial expressions. Compared to DECA [17], upon which we build our model, we achieve significant AU performance improvement (around 10%). The method proposed by Kuang et al. [27] integrate AU correlations in learning 3D face models, but only apply to single images. In conclusion, our Spatial Module and Temporal Module both contribute to significant AU classification improvement. Our final model achieve SOTA accuracy on AU4, AU10, AU17, AU24.

Comparisons on DISFA We conduct similar experiments on DISFA dataset as BP4D and compare with three different types of models, as shown in Table. 3. Compared to the only dynamic model Biomechanics-AU [8], our final model lifts the average F1 score by 2.9%. Although our average performance is slightly worse than the SOTA model FAU [22], but for most AUs our model generate much better prediction, such as AU6 (13.9% ↑), AU9 (20.2%↑), AU12 (14.1%↑). It shows that our model is promising in applying to some AUs that are hard to directly distinguish from image features.

4.3. Ablation Study

To further prove the effectiveness of each module and the necessity of combining mesh feature with image feature, we conduct ablation study with different combinations of input data and Transformer Modules, as presented in Table. 4. We explore 7 different cases in total on BP4D and DISFA and analyze their performance.

- Video input + Temporal Module $\mathcal{T}^t$: when only taking

Table 2. Comparing the F1-score of AU recognition on BP4D between proposed method and state-of-art methods

Method	AU1	AU2	AU4	AU6	AU7	AU10	AU12	AU14	AU15	AU17	AU23	AU24	**Avg.**
DECA [17] + MLP	54.5	40.1	53.2	54.7	70.8	74.0	71.2	51.8	42.6	66.2	51.0	41.4	56.1
EMOCA [9] + MLP	54.1	51.2	57.4	45.4	73.7	64.2	77.5	55.8	44.3	67.8	53.5	44.7	57.5
Kuang et al. [27]	55.2	57.0	53.7	66.7	77.8	76.4	79.7	59.8	44.1	60.1	**53.6**	46.0	60.8
SRERL [29]	46.9	45.3	55.6	77.1	78.4	83.5	**87.6**	63.9	**52.2**	63.9	47.1	53.3	62.1
LP-Net [42]	43.4	38.0	54.2	77.1	76.7	83.8	87.2	63.3	45.3	60.5	48.1	54.2	61.0
FAU [22]	51.7	49.3	61.0	77.8	79.5	82.9	86.3	67.6	51.9	63.0	43.7	56.3	64.2
*Biomechanics-AU [8]	**57.4**	52.6	64.6	**79.3**	**81.5**	82.7	85.6	**67.9**	47.3	58.0	47.0	44.9	64.1
Ours: Spatial	56.8	**59.5**	64.8	70.2	79.5	76.2	80.4	61.7	44.6	60.6	48.5	58.2	63.4
***Ours:**Temporal + Spatial	54.5	56.2	**67.1**	72.3	81.4	**84.9**	87.5	61.4	43.8	**67.3**	52.5	**64.3**	**66.0**

Table 3. Comparing the F1-score of AU recognition on DISFA between proposed method and state-of-art methods

Method	AU1	AU2	AU4	AU6	AU9	AU12	AU25	AU26	**Avg.**
DECA + MLP	44.7	40.5	55.2	50.5	48.0	66.4	77.0	50.7	54.1
EMOCA + MLP	44.5	41.2	54.8	50.5	48.9	67.8	78.5	51.1	54.7
SRERL	45.7	47.8	59.6	47.1	45.6	73.5	84.3	43.6	55.9
LP-Net	29.9	24.7	72.7	46.8	49.6	72.9	**93.8**	65.0	56.9
FAU	46.1	48.6	72.8	56.7	50.0	72.1	90.8	55.4	**61.5**
*Biomechanics-AU	41.5	44.9	60.3	51.5	50.3	70.4	91.3	**55.3**	58.2
Ours: Spatial	46.8	48.9	45.6	**72.0**	68.8	84.5	61.4	51.3	60.0
***Ours:** Temporal + Spatial	**47.7**	**50.5**	48.2	70.6	**70.2**	**86.2**	62.5	52.7	61.1

Table 4. Ablation study with different combinations of input data and Transformer Modules. Average F1 scores for AU classification on BP4D and DISFA are provided. In the second column, the $\mathcal{T}^t$ and $\mathcal{T}^S$ denote the temporal module and spatial module correspondingly. The last row represents our final model.

Input	Transformer Modules	BP4D	DISFA
Video only	$\mathcal{T}^t$	59.8	56.2
Mesh only	$\mathcal{T}^t$	60.1	54.6
Mesh only	$\mathcal{T}^S$	62.6	56.4
Mesh only	$\mathcal{T}^t + \mathcal{T}^S$	63.1	58.8
Mesh + Video	$\mathcal{T}^t$	61.6	58.9
Mesh + Video	$\mathcal{T}^S$	63.4	60.0
final: Mesh + Video	$\mathcal{T}^t + \mathcal{T}^S$	66.0	61.0

video features, the Spatial Module for AU correlation will not be applicable since it relies on AU tokens generated from 3D meshes. In this case, the network is trained fully depending on appearance features. The 2-nd row in Table. 4 indicates that the AU classification performance on both dataset decrease a lot.

- Mesh input + $\mathcal{T}^t$ or $\mathcal{T}^S$ or $\mathcal{T}^t + \mathcal{T}^S$: when only taking coarse mesh sequences as input, we explore three cases with activating a single Transformer module or both of them. Correspondingly, using $\mathcal{T}^t$ or $\mathcal{T}^S$ will determine the emphasis, on motion dynamics or on AU spatial correlations. The results from row3 to row5 in Table. 4 shows that using Spatial Module for encoding AU correlations helps to generate better AU predictions than using $\mathcal{T}^t$ only. In addition, when we combine the temporal embeddings and spatial embeddings, we can obtain performance improvement over Mesh + $\mathcal{T}^S$ (0.5% ↑ for BP4D, 2.4% ↑ for DISFA).

- Mesh + Video input + $\mathcal{T}^t$ or $\mathcal{T}^S$ or $\mathcal{T}^t + \mathcal{T}^S$: we can draw similar conclusion as the second case. Through combining dynamics modeling and AU spatial correlation modeling, we can get notable better results: (2.6% ↑) on BP4D and (1.0% ↑) on DISFA, comparing with only using $\mathcal{T}^S$.

We can also observe that under the same network structure, it's essential to combine geometrical mesh features and video appearance features to achieve better AU detection performance. Based on the results in {row2, row3, row6} or {row5, row8}, using the coarse input mesh sequences only may fail to provide sufficient geometric representation for each AU but feature combination addresses this problem.

5. Method Limitation

For the Temporal Module used for modeling the facial dynamics, the captured temporal dependency may not be stable and universally applicable due to two reasons. (1) The input to the autoregressive transformer are AU occurrence embeddings. However, the binary AU sequence can only reflect simple dynamics of AU state shift (0 → 1 or 1 → 0). (2) The performance of the dynamic model can be impacted by different fps.

6. Conclusion

In this paper, we presented an AU-aware dynamic 3D face reconstruction approach, for capturing AU-based facial motion dynamics and spatial correlations and refine the reconstructed 3D face geometry. We design a Temporal Module based on autoregressive transformer and a AU-Spatial-Correlation Module, for modeling the temporal dependency and spatial correlations of local facial motions. Experimental results prove the effectiveness of both two modules in terms of improved reconstruction of facial motions and better AU detection performance. In application, our model can be directly applied to simultaneous dynamic 3D face reconstruction and AU detection.

References

[1] Luigi Ariano, Claudio Ferrari, Stefano Berretti, and Alberto Del Bimbo. Action unit detection by learning the deformation coefficients of a 3d morphable model. Sensors, 21(2):589, 2021. 2, 3

[2] Ziqian Bai, Zhaopeng Cui, Xiaoming Liu, and Ping Tan. Riggable 3d face reconstruction via in-network optimization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6216–6225, 2021. 1

[3] Ziqian Bai, Zhaopeng Cui, Jamal Ahmed Rahim, Xiaoming Liu, and Ping Tan. Deep facial non-rigid multi-view stereo. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5850–5860, 2020. 1, 6

[4] Neslihan Bayramoglu, Guoying Zhao, and Matti Pietikäinen. Cs-3dlbp and geometry based person independent 3d facial action unit detection. In 2013 International Conference on Biometrics (ICB), pages 1–6. IEEE, 2013. 2

[5] Uttaran Bhattacharya, Nicholas Rewkowski, Abhishek Banerjee, Pooja Guhan, Aniket Bera, and Dinesh Manocha. Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents. In 2021 IEEE virtual reality and 3D user interfaces (VR), pages 1–10. IEEE, 2021. 2

[6] 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. 1

[7] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: A 3d facial expression database for visual computing. IEEE Transactions on Visualization and Computer Graphics, 20(3):413–425, 2013. 2

[8] Zijun Cui, Chenyi Kuang, Tian Gao, Kartik Talamadupula, and Qiang Ji. Biomechanics-guided facial action unit detection through force modeling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8694–8703, 2023. 6, 7, 8

[9] Radek Daněček, Michael J Black, and Timo Bolkart. Emoca: Emotion driven monocular face capture and animation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 20311–20322, 2022. 6, 7, 8

[10] Antonios Danelakis, Theoharis Theoharis, and Ioannis Pratikakis. Action unit detection in 3 d facial videos with application in facial expression retrieval and recognition. Multimedia Tools and Applications, 77(19):24813–24841, 2018. 2

[11] 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 Computer Vision and Pattern Recognition Workshops, pages 285–295, 2019. 1

[12] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 16321–16330, 2021. 2

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020. 2

[14] Paul Ekman, Wallace V. Friesen, and J. C. Hager. Facial action coding system. A Human Face, Salt Lake City, UT, 2002. 2, 4

[15] Hehe Fan, Yi Yang, and Mohan Kankanhalli. Point 4d transformer networks for spatio-temporal modeling in point cloud videos. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 14204–14213, 2021. 3

[16] 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. 3, 4

[17] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. volume 40, 2021. 1, 3, 6, 7, 8

[18] Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8649–8658, 2021. 1, 2

[19] Thomas Gerig, Andreas Morel-Forster, Clemens Blumer, Bernhard Egger, Marcel Luthi, Sandro Schönborn, and Thomas Vetter. Morphable face models-an open framework. In 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pages 75–82. IEEE, 2018. 1

[20] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3d dense face alignment. In Proceedings of the European Conference on Computer Vision (ECCV), 2020. 1, 6

[21] Guosheng Hu, Fei Yan, Josef Kittler, William Christmas, Chi Ho Chan, Zhenhua Feng, and Patrik Huber. Efficient 3d morphable face model fitting. Pattern Recognition, 67:366–379, 2017. 1

[22] Geethu Miriam Jacob and Bjorn Stenger. Facial action unit detection with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7680–7689, 2021. 3, 6, 7, 8

[23] Jiaxi Jiang, Paul Streli, Huajian Qiu, Andreas Fender, Larissa Laich, Patrick Snape, and Christian Holz. Avatarposer: Articulated full-body pose tracking from sparse motion sensing. In Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V, pages 443–460. Springer, 2022. 2

[24] Yifan Jiang, Shiyu Chang, and Zhangyang Wang. Transgan: Two pure transformers can make one strong gan, and that can scale up. Advances in Neural Information Processing Systems, 34:14745–14758, 2021. 2

[25] Yang Jiao, Yi Niu, Trac D Tran, and Guangming Shi. 2d+ 3d facial expression recognition via discriminative dynamic

range enhancement and multi-scale learning. arXiv preprint arXiv:2011.08333, 2020. 2

[26] Mohammad Rami Koujan, Michail Christos Doukas, Anastasios Roussos, and Stefanos Zafeiriou. Head2head: Video-based neural head synthesis. In 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), pages 16–23. IEEE, 2020. 1, 2

[27] Chenyi Kuang, Zijun Cui, Jeffrey O Kephart, and Qiang Ji. Au-aware 3d face reconstruction through personalized au-specific blendshape learning. In Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIII, pages 1–18. Springer, 2022. 7, 8

[28] Duy Le Hoai, Eunchae Lim, Eunbin Choi, Sieun Kim, Sudarshan Pant, Guee-Sang Lee, Soo-Huyng Kim, and Hyung-Jeong Yang. An attention-based method for multi-label facial action unit detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2454–2459, 2022. 3

[29] Guanbin Li, Xin Zhu, Yirui Zeng, Qing Wang, and Liang Lin. Semantic relationships guided representation learning for facial action unit recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, pages 8594–8601, 2019. 7, 8

[30] Hanting Li, Mingzhe Sui, Feng Zhao, Zhengjun Zha, and Feng Wu. Mvt: mask vision transformer for facial expression recognition in the wild. arXiv preprint arXiv:2106.04520, 2021. 3

[31] Huibin Li, Jian Sun, Zongben Xu, and Liming Chen. Multimodal 2d+ 3d facial expression recognition with deep fusion convolutional neural network. IEEE Transactions on Multimedia, 19(12):2816–2831, 2017. 2, 3

[32] Hongkang Li, Meng Wang, Sijia Liu, and Pin-Yu Chen. A theoretical understanding of shallow vision transformers: Learning, generalization, and sample complexity. In The Eleventh International Conference on Learning Representations, 2023. 2

[33] Ruilong Li, Karl Bladin, Yajie Zhao, Chinmay Chinara, Owen Ingraham, Pengda Xiang, Xinglei Ren, Pratusha Prasad, Bipin Kishore, Jun Xing, and Hao Li. Learning formation of physically-based face attributes, 2020. 2

[34] 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. 6

[35] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 1954–1963, 2021. 2, 3

[36] Jingwang Ling, Zhibo Wang, Ming Lu, Quan Wang, Chen Qian, and Feng Xu. Semantically disentangled variational autoencoder for modeling 3d facial details. IEEE Transactions on Visualization and Computer Graphics, 2022. 1

[37] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 3202–3211, 2022. 4

[38] Wan-Chun Ma, Mathieu Lamarre, Etienne Danvoye, Chongyang Ma, Manny Ko, Javier von der Pahlen, and Cyrus A Wilson. Semantically-aware blendshape rigs from facial performance measurements. In SIGGRAPH ASIA 2016 Technical Briefs, pages 1–4. 2016. 2

[39] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. Weakly-supervised action transition learning for stochastic human motion prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8151–8160, 2022. 2

[40] S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F Cohn. Disfa: A spontaneous facial action intensity database. IEEE Transactions on Affective Computing, 4(2):151–160, 2013. 6

[41] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 8844–8854, 2022. 2

[42] Xuesong Niu, Hu Han, Songfan Yang, Yan Huang, and Shiguang Shan. Local relationship learning with person-specific shape regularization for facial action unit detection. In Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition, pages 11917–11926, 2019. 7, 8

[43] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 10985–10995, 2021. 2

[44] Stylianos Ploumpis, Haoyang Wang, Nick Pears, William AP Smith, and Stefanos Zafeiriou. Combining 3d morphable models: A large scale face-and-head model. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 10934–10943, 2019. 1

[45] Michael J Reale, Benjamin Klinghoffer, Micah Church, Hannah Szmurlo, and Lijun Yin. Facial action unit analysis through 3d point cloud neural networks. In 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), pages 1–8. IEEE, 2019. 2

[46] Xuanchi Ren and Xiaolong Wang. Look outside the room: Synthesizing a consistent long-term 3d scene video from a single image. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3563–3573, 2022. 2

[47] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael Black. Learning to regress 3d face shape and expression from an image without 3d supervision. In Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), June 2019. 1

[48] Arman Savran, BüLent Sankur, and M Taha Bilge. Comparative evaluation of 3d vs. 2d modality for automatic detection of facial action units. Pattern recognition, 45(2):767–782, 2012. 2

[49] Yeongho Seol, John P Lewis, Jaewoo Seo, Byungkuk Choi, Ken Anjyo, and Junyong Noh. Spacetime expression cloning for blendshapes. ACM Transactions on Graphics (TOG), 31(2):1–12, 2012. 2

[50] 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. 1, 2

[51] Diego Thomas and Rin-Ichiro Taniguchi. Augmented blendshapes for real-time simultaneous 3d head modeling and facial motion capture. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3299–3308, 2016. 2

[52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017. 2

[53] Thibaut Weise, Sofien Bouaziz, Hao Li, and Mark Pauly. Realtime performance-based facial animation. ACM transactions on graphics (TOG), 30(4):1–10, 2011. 2

[54] Cheng-hsin Wuu, Ningyuan Zheng, Scott Ardisson, Rohan Bali, Danielle Belko, Eric Brockmeyer, Lucas Evans, Timothy Godisart, Hyowon Ha, Xuhua Huang, et al. Multiface: A dataset for neural face rendering. arXiv preprint arXiv:2207.11243, 2022. 6, 7

[55] Fanglei Xue, Qiangchang Wang, and Guodong Guo. Transfer: Learning relation-aware facial expression representations with transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 3601–3610, 2021. 3

[56] 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. 1, 2

[57] Zipeng Ye, Mengfei Xia, Ran Yi, Juyong Zhang, Yu-Kun Lai, Xuwei Huang, Guoxin Zhang, and Yong-jin Liu. Audio-driven talking face video generation with dynamic convolution kernels. IEEE Transactions on Multimedia, 2022. 2

[58] 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

[59] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. Image and Vision Computing, 32(10):692–706, 2014. 6

[60] Zengqun Zhao and Qingshan Liu. Former-dfer: Dynamic facial expression recognition transformer. In Proceedings of the 29th ACM International Conference on Multimedia, pages 1553–1561, 2021. 3

[61] 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

[62] Kangkang Zhu, Zhengyin Du, Weixin Li, Di Huang, Yunhong Wang, and Liming Chen. Discriminative attention-based convolutional neural network for 3d facial expression recognition. In 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), pages 1–8. IEEE, 2019. 2

[63] Xiangyu Zhu, Xiaoming Liu, Zhen Lei, and Stan Z Li. Face alignment in full pose range: A 3d total solution. IEEE transactions on pattern analysis and machine intelligence, 2017. 1

[64] Xiangyu Zhu, Fan Yang, Di Huang, Chang Yu, Hao Wang, Jianzhu Guo, Zhen Lei, and Stan Z Li. Beyond 3dmm space: Towards fine-grained 3d face reconstruction. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16, pages 343–358. Springer, 2020. 1

[65] Shihao Zou, Yuxuan Mu, Xinxin Zuo, Sen Wang, and Li Cheng. Event-based human pose tracking by spiking spatiotemporal transformer. arXiv preprint arXiv:2303.09681, 2023. 2