Accurate 3D Face Reconstruction with Facial Component Tokens

Tianke Zhang[1,2*] Xuangeng Chu[2] Yunfei Liu[2] Lijian Lin[2] Zhendong Yang[1,2]
Zhengzhuo Xu[1,2] Chengkun Cao[2] Fei Yu[3] Changyin Zhou[3] Chun Yuan[1†] Yu Li[2†]
[1]Tsinghua Shenzhen International Graduate School
[2]International Digital Economy Academy (IDEA) [3]Vistring Inc.

Figure 1: **Exemplar 3D face reconstruction results of our method named TokenFace** (first row: inputs, second row: results). Our approach achieves faithful reconstruction results for challenging cases, including varying sizes, ages, poses, races, and partial occlusions. Additionally, our method enables accurate transferring a target expression to the persons (third row).

Abstract

Accurately reconstructing 3D faces from monocular images and videos is crucial for various applications, such as digital avatar creation. However, the current deep learning-based methods face significant challenges in achieving accurate reconstruction with disentangled facial parameters and ensuring temporal stability in single-frame methods for 3D face tracking on video data. In this paper, we propose ***TokenFace****, a transformer-based monocular 3D face reconstruction model. TokenFace uses separate tokens for different facial components to capture information about different facial parameters and employs temporal transformers to capture temporal information from video data. This design can naturally disentangle different facial components and is flexible to both 2D and 3D training data. Trained on hybrid 2D and 3D data, our model shows its power in accurately reconstructing faces from images and producing stable results for video data. Experimental results on popular benchmarks NoW and Stirling demonstrate that TokenFace achieves state-of-the-art performance, outperforming existing methods on all metrics by a large margin.*

1. Introduction

The analysis and reconstruction of human faces from images are critical research topics in computer vision due to their vast range of applications. A vital technology is the creation of a detailed 3D model that accurately captures both the geometry and appearance of the face from visual data. However, creating such a model is challenging when working with monocular input where there is no access to 3D information from multiple views or sensors.

In this work, we focus on the task of 3D face reconstruction using a 3D deformable face model where the problem can be expressed as estimating the parameters of the 3D face model. While optimization-based approaches have been proposed to solve this model-fitting problem, recent advances in deep learning have facilitated that neural network-based methods can predict these parameters from training data. However, unlike 2D tasks that are easy for annotation, the lack of 3D ground truth data makes 3D face reconstruction challenging. As a result, many existing methods [17, 14, 50] only supervise the 2D rendering results during training, leading to sub-optimal performance in 3D space. For instance, DECA [17] and Deep3dFaceRecon [14] leverage self-supervised training on 2D images, while MICA [50] uses 2D-3D data pairs but only reconstructs the face shape

*This work was done while Tianke Zhang was interning at IDEA.
†Corresponding authors.

without expression, pose, and other information. Dense Landmark [43], on the other hand, requires a large amount of annotated data for training and relies on a dense landmark for 3D face fitting.

Most of the previous deep learning-based methods use Convolutional Neural Networks (CNNs) to regress the parameters together as one vector from the input image. However, these approaches suffer from entanglement between different facial parameters, hindering the ability to improve performance.In addition to this, there are some methods [10, 4] that have tried transformer models for face reconstruction. However, they either rely on conditional GANs or follow the original transformer structure, making the reconstruction results not highly reflective of the original. In this work, we propose ***TokenFace***, a transformer-based 3D face reconstruction model that utilizes six tokens, including shape, expression, jaw pose, camera, texture, and lighting, to effectively distinguish different image features and decouple various parameters. This helps to improve the disentanglement of individual components for more accurate reconstruction. Our model is trained with large-scale hybrid datasets consisting of both 3D scanned data and 2D in-the-wild face images with a multi-stage training pipeline. During the training phase, we can flexibly train our model based on the data type, using rendered 2D images for the 2D image dataset and mesh vertex supervision for datasets with 3D ground truth. To capture temporal information in video datasets, we introduce a temporal transformer between adjacent frames. Our loss functions are designed to address specific data types and focus on dimensions of vertices, identity, image consistency, and other factors. Our proposed method outperforms previous methods and demonstrates excellent performance in 3D face reconstruction, as evidenced by its performance in NoW Benchmark and Stirling Benchmark, with a minimum of 10 % improvement in accuracy. Example results of our method are illustrated in Fig. 1. Our model also exhibits stable performance in video reconstruction using our temporal modeling. Our methods enable many practical applications like facial expression transfer between different avatars as shown in the last row in Fig. 1.

Our contribution can be summarized as follows.

- We present a framework for 3D face reconstruction from monocular images based on transformers. Our approach uses separate tokens to improve the disentanglement of individual components for more accurate reconstruction.
- We train our network on large-scale hybrid dataset of both 2D and 3D data containing a large variety of faces and expressions using our training pipeline.
- Our framework can be naturally extended to 3D face reconstruction in videos with straightforward temporal modeling to improve temporal stability.
- Our model surpasses other methods by achieving notably lower reconstruction errors on benchmarks. For example, we achieve mean errors of 0.95 on NoW benchmark (previous best is 1.11) and 0.95 on Stirling benchmark (previous best is 1.16).

2. Related Work

3D Face Morphable Model. The 3D face morphable model (3DMM) [5, 32, 6, 7, 28, 45, 3, 27] has been studied for a long time in 3D face research field, which usually defines a linear model to represent the geometric structure and texture of a human face. Recently, more and more 3D face datasets were proposed, which provide more identities and expressions. These data make it possible to construct 3DMM models with better generalization ability [6] and better expression deform-ability [41, 7, 28, 45, 3]. Also, some 3DMM methods [27, 45, 3] also obtained higher reconstruction accuracy with the development of data capturing system. Besides the increasing quality of data, many works [41, 7, 45, 31, 1, 38, 27, 20] try to improve the 3DMM models from other aspects. These works [7, 45, 41] design bi-linear or multi-linear models that decompose the identity and expression of faces. Nonlinear models have also been used to improve the accuracy of facial deformation. Neumann et al. [31] decomposes the captured face mesh sequences into sparse and localized deformation components. Furthermore, generative adversarial networks (GAN) were also used to build the non-linear 3DMMs [1, 38, 27, 20].

Monocular Face Reconstruction Based on 3DMM. 3DMM-based reconstruction methods play a crucial role in the field of 3D reconstruction of monocular images. Along with the proposal of the FLAME model [28], many reconstruction methods [18, 49, 14, 21, 17, 50] based on FLAME achieve promising results of monocular face reconstruction. These methods generally follow a self-supervised or weakly supervised training framework and the 3D face reconstruction task is reduced to a model-fitting problem with the help of 3DMM. Among them, the early methods [34, 32, 37] try to regress the parameters of 3DMM using the facial landmarks. The current deep models [9, 45, 17, 3] usually predict the parameters directly from the input image. Furthermore, to achieve a more accurate reconstruction of facial details which are not parameterized by 3DMM, some works [9, 23, 45, 17, 3] use a two-stage framework, which firstly predicts a rough face mesh and then refined the facial details through depth and displacement map.

Transformer for Face Analysis. Following the success of transformers in natural language processing (NLP) tasks [40, 15], vision transformers [16, 29] have also achieved the state-of-the-art performance in many computer vision tasks. ViT [16] splits an image into patches and flattens the patches as a token sequence. Zhong et al. [48] modifies ViT and shows competitive performance on face recognition. Based on the

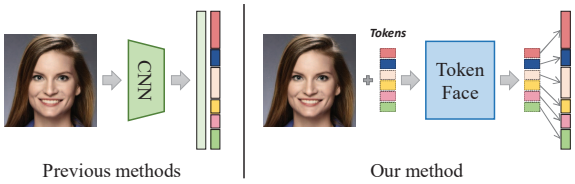

Figure 2: **A comparison of existing methods and ours**. Previous works typically rely on a CNN to predict a single 1D vector and segment it into different facial model parameters (colored boxes). Instead, we use separate tokens to encode independent facial components and aggregate this information with image data to obtain updated tokens, which are then used to predict the face parameters.

large-scale pre-training and transfer ability of transformers, FaRL [47] achieves superior performance on facial analysis tasks including face parsing and face alignment.

3. Method

3.1. Building TokenFace

Traditional methods for 3D face reconstruction typically rely on Convolutional Neural Networks (CNNs) to extract facial features and recover 3D information. These methods take a face image as input and regress a 1D vector, which is later segmented into different facial model parameters such as shape, expression, and pose, as shown in Fig. 2 (left). However, these approaches interleave the different facial components from the beginning, making it difficult to disentangle their effects and hindering the performance in accuracy. This is because CNNs can only extract these features from the last layer using global pooling and linear layers. As a result, the coupling between pose and expression can lead to inaccurate 3D face reconstruction.

To improve the disentanglement between different facial components in 3D face reconstruction, we propose a novel network structure based on the Vision Transformer (ViT) architecture [16]. Recent studies have shown that ViT's feature extraction ability often outperforms that of CNN networks when trained on large-scale data. We directly replace the ResNet50 backbone used in [17] with ViT-Base [16] and use a global token, *i.e.* a vector, as the task token to predict the facial parameters, similar to the behavior of the CNN-based model. As shown in Table 1, changing the backbone does improve reconstruction accuracy, but only marginally (the test setup is the same as the ones in our ablation study whose details are presented later in Sec. 5). Therefore, simply replacing the backbone is insufficient to achieve significant improvements in results.

To adapt ViT to 3D face reconstruction task, we introduce six tokens to represent shape, expression, jaw pose, camera pose, texture, and lighting, respectively. These tokens are combined with image tokens to create a comprehensive input,

Backbone	Reconstruction Error		
	Median	Mean	Std
ResNet50 [22]	1.10	1.41	1.19
ViT-Base [16]	1.08	1.35	1.15

Table 1: **Results of naively changing backbone from CNN to ViT.** We test the same training and testing scheme using both ResNet50 and ViT-Base. The model with ViT-Base achieves slightly better performance.

which enables greater independence between different facial component parameters, reducing mutual influence. Fig. 2 (right) illustrates our method, while Fig. 3 shows the main structure of our TokenFace. We first divide the input image into patches, flatten them, and add position embeddings to build image tokens. Then, we append the six learnable facial component tokens to the image token and feed them into the transformer blocks together. Specifically, the six tokens are denoted as: $\boldsymbol{\beta}$ for shape, $\boldsymbol{\psi}$ for expression, $\boldsymbol{\theta}$ for jaw pose, $\boldsymbol{C}$ for the camera's affine matrix, $\boldsymbol{\alpha}$ for texture, and $\boldsymbol{\iota}$ for light. We use six FLAME headers corresponding to our FLAME tokens to estimate the FLAME parameters after encoding. The resulting output comprises 300-dimensional shape parameters, 100-dimensional expression parameters, 3-dimensional jaw pose parameters, 7-dimensional camera parameters, 50-dimensional texture parameters, and 27-dimensional light parameters. The camera parameters $\boldsymbol{C}$ consist of **s**cale (1-dim), **r**otation (3-dim), and **t**ranslation (3-dim). Finally, we reconstruct the 3D face mesh based on the FLAME model and our predicted parameters.

There are two more interesting advantages of using individual face component tokens. First, we can generate a metrical face in 3D using shape, expression, and jaw pose and set other tokens to zero. With additional camera pose, texture, and lighting, the face can be projected into the 2D image plane. This is particularly useful when training on a large-scale hybrid dataset of both 2D and 3D data. This hybrid dataset can contain a larger variety of faces and facial expressions, allowing for training a better 3D face reconstruction model. The second advantage of using separate tokens for each facial component is the ability to create a simple temporal modeling. This approach can be used to build a temporal transformer on top of our TokenFace model, which can aggregate temporal information from different frames at token level and generate more stable results for video prediction.

3.2. Hybrid Training Strategy

To take full advantage of publicly available 2D and 3D datasets, we design a hybrid training strategy that enables model learning on different types of data. Specifically, 3D data provide precise FLAME meshes for full supervision. Meanwhile, 2D images provide a large amount of photo-

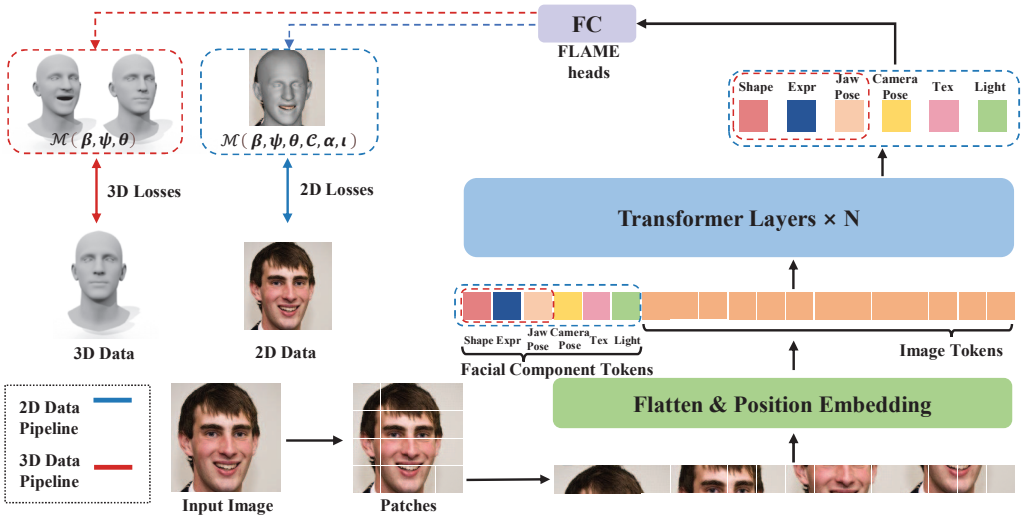

Figure 3: **Illustration of our pipeline**. We partition input image into patches, flatten them, and add position embedding to build the image tokens. The learnable facial component tokens, including shape, expression, jaw pose, camera pose, texture, and lighting, are appended to the image token sequence and fed to the transformer blocks together. The output facial component tokens at the end of the transformer are converted to FLAME parameters using FLAME heads (FC layers). 3D face mesh is recovered from the FLAME model and our predicted parameters. For 2D and 3D data, different losses are used.

metric information (*e.g.*, texture, lighting, *etc.*) for self-supervision. Our main motivation is to simultaneously learn these two types of supervision for a better 3D face reconstruction. Mathematically, the overall training target is to minimize

$$\mathcal{L}_{all} = \lambda_{3D} \cdot \mathcal{L}_{3D} + \lambda_{2D} \cdot \mathcal{L}_{2D}, \tag{1}$$

where $\lambda_{3D}, \lambda_{2D}$ are two hyper-parameters to balance loss terms. $\mathcal{L}_{2D}$ and $\mathcal{L}_{3D}$ denote the loss for 2D data and 3D data, respectively.

3.2.1 Full Supervision for 3D Data

For the 3D dataset, we can fully supervise the vertex of mesh directly. Therefore, the 3D-related loss $\mathcal{L}_{3D}$ is defined as

$$\mathcal{L}_{3D} = \lambda_{mesh}\mathcal{L}_{mesh} + \lambda_{vc}\mathcal{L}_{vc}, \tag{2}$$

where $\mathcal{L}_{mesh}$ is mesh loss, $\mathcal{L}_{vc}$ is vertex consistency loss.

Mesh Loss. For the data with non-neural expressions, we reconstruct the face based on the estimated shape $\boldsymbol{\beta}$, expression $\boldsymbol{\psi}$ and jaw pose $\boldsymbol{\theta}$ parameters, while for the data only with neural expressions, we reconstruct the face using the shape parameters $\boldsymbol{\beta}$. The $\mathcal{L}_{mesh}$ is defined as

$$\mathcal{L}_{mesh} = w|\mathcal{V}^i - \mathcal{V}^i_{gt}|_1, \tag{3}$$

where w is the region-dependent weight of FLAME vertices. $\mathcal{V}^i_{gt}$ is the ground truth 3D vertices. $\mathcal{V}$ is extracted from face mesh $\mathcal{M}$, which is

$$\mathcal{M} = \text{FLAME}(\boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{\theta}). \tag{4}$$

Vertex Consistency Loss. In order to further disentangle the shape and expression parameters, we train our model on the meshes with the same identity (*i.e.*, which should share the same shape parameter) and different expressions. Similar to DECA [17], we adopt vertex consistency loss $\mathcal{L}_{vc}$ to constrain the objectiveness. $\mathcal{L}_{vc}$ is defined as

$$\begin{aligned} \mathcal{L}_{vc} &= \sum^{M} w|\mathcal{V}^{b\to a} - \mathcal{V}^b_{gt}|_1, \\ \mathcal{M}^{b\to a} &= \text{FLAME}(\boldsymbol{\beta}_a, \boldsymbol{\psi}_b, \boldsymbol{\theta}_b), \end{aligned} \tag{5}$$

where $\mathcal{V}^{b\to a}$ is the generated vertices with exchanging shape $\boldsymbol{\beta}_b$ to $\boldsymbol{\beta}_a$. M denotes the number of samples from the same identity. Besides, these shapes, expressions, and jaw poses are from different images with the same identity.

3.2.2 Self-Supervision for 2D Data

For the 2D Dataset, we need to perform self-supervised training at the image level. The 2D-related loss $\mathcal{L}_{2D}$ is defined as

$$\begin{aligned} \mathcal{L}_{2D} = \lambda_{eyes}\mathcal{L}_{eyes} + \lambda_{lips}\mathcal{L}_{lips} + \lambda_{sc}\mathcal{L}_{sc} + \lambda_{id}\mathcal{L}_{id} \\ +\omega_{lmk}\mathcal{L}_{lmk} + \omega_{photo}\mathcal{L}_{photo} + \mathcal{L}_{reg}, \end{aligned} \tag{6}$$

which includes 7 loss terms. Here we introduce each of the loss terms in $\mathcal{L}_{2D}$ for learning on 2D data.

Landmark Loss. We calculate the $L1$ distance between the projected landmarks of the mesh and the pre-processed ground-truth landmarks. Note that the non-visible key points will not be involved in landmark loss.

Eyelids & Lips Loss. Inspired by DECA [17], we also introduce eyes & lips loss function to add additional constraints on the lip and eye region. It could make the model perform better in scenes such as video. :

$$\begin{aligned} \mathcal{L}_{eyes} &= \sum_{(i,j)\in E} \|\mathbf{k}_i - \mathbf{k}_j - s\Pi(V_i - V_j)\|_1, \\ \mathcal{L}_{lips} &= \sum_{(i,j)\in L} \|\mathbf{k}_i - \mathbf{k}_j - s\Pi(V_i - V_j)\|_1, \end{aligned} \tag{7}$$

where E and L are the set of upper/lower eyelid and lips landmarks pairs respectively. $\mathbf{k}_i$ and $\mathbf{k}_j$ mean 2D keypoints, which are detected by face landmark detector *. V_i and V_j denote the 3D face landmarks in the reconstructed mesh. s is the scale parameter of the camera parameters, and $s\Pi$ means projecting the 3D points to the image coordinate.

Image Photometric Loss. As a two-dimensional visual supervision, we also add supervision to ensure the consistency between the rendered image $\mathcal{I}_r$ and input image $\mathcal{I}$:

$$\mathcal{L}_{photo} = \|m_{\mathcal{I}} \cdot (\mathcal{I} - \mathcal{I}_r)\|_1, \tag{8}$$

where $\mathcal{I}_r$ = FLAME-Render$(\boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{\theta}, \boldsymbol{C}, \boldsymbol{\alpha}.\boldsymbol{\iota})$ and $m_{\mathcal{I}}$ is the mask of the face.

Shape Consistency Loss. As with the 2D-3D datasets, on 2D datasets, we also need to keep the consistency of the shape in different images of the same identity, i.e. the shape parameter in FLAME. In contrast to the above, we define $\mathcal{L}_{sc}$ with the following equation:

$$\begin{aligned} \mathcal{L}_{sc} &= \sum^{M} m_{\mathcal{I}} \|\mathcal{I}^{b\to a} - \mathcal{I}^{b}_{gt}\|_1, \\ \mathcal{I}^{b\to a} &= \text{FLAME-Render}(\boldsymbol{\beta}_a, \boldsymbol{\psi}_b, \boldsymbol{\theta}_b, \boldsymbol{C}_b, \boldsymbol{\alpha}_b.\boldsymbol{\iota}_b), \end{aligned} \tag{9}$$

where $m_{\mathcal{I}}$ is the face mask of the image $\mathcal{I}$, and $\mathcal{I}^{b\to a}$ is the rendered image with exchanging shape $\boldsymbol{\beta}_b$ to $\boldsymbol{\beta}_a$. M denotes the number of images from the same identity.

Identity Loss. To better reconstruct the details of the input image, we optimize our model with a perceptual loss based on an advanced face recognition model [13]:

$$\mathcal{L}_{id} = 1 - \frac{f(\mathcal{I}) f(\mathcal{I}_r)}{\|f(\mathcal{I})\|_2 \cdot \|f(\mathcal{I}_r)\|_2}, \tag{10}$$

where $f(\cdot)$ is the feature extractor.

Regularization. To avoid over-fitting of the facial parameters, we add a regularization loss to the 3DMM coefficients of the regression:

$$\mathcal{L}_{reg} = \omega_{\boldsymbol{\alpha}} \cdot \|\boldsymbol{\alpha}\|^2 + \omega_{\boldsymbol{\beta}} \cdot \|\boldsymbol{\beta}\|^2 + \omega_{\boldsymbol{\psi}} \cdot \|\boldsymbol{\psi}\|^2, \tag{11}$$

*http://dlib.net/

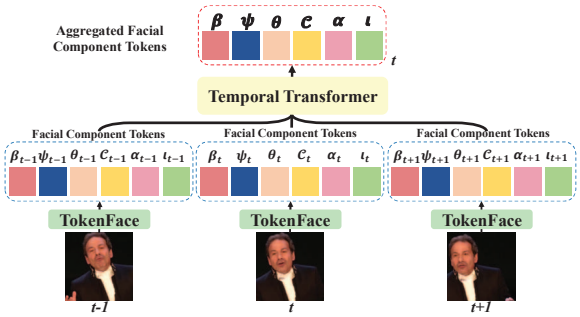

Figure 4: **Temporal model of our method**. We obtain facial component tokens for each frame independently and append them together and send to an additional temporal transformer to predict the aggregated facial component tokens for the middle frame.

where $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\psi}$ are estimated flame parameters, and $\omega_{\boldsymbol{\alpha}}, \omega_{\boldsymbol{\beta}}, \omega_{\boldsymbol{\psi}}$ are hyper-parameters to balance the loss scale.

Pose-Aware Loss Function. In order to solve the problem of inaccurate landmark estimation in large head pose, we propose the Pose-Aware Loss (PAL) function to adaptively balance the weights of $\mathcal{L}_{lmk}$ and $\mathcal{L}_{photo}$ in Eq.(6). In detail, We first pre-process the training data to obtain the face orientation, where x and z represent the pitch and yaw angle of the face, respectively. Then we judge the degree of the large pose of a human face by the maximum value of azimuth and elevation angles, *i.e.* $\xi = max(\zeta_1, \zeta_2) - \pi/4$, where ζ_1 and ζ_2 denotes the azimuth and the elevation respectively. Next, we take ξ as input and adjust the values of ω_{lmk} and ω_{photo} using a designed linear function:

$$\omega_{lmk} = \begin{cases} 1.6, & \xi \le 0, \\ a\xi + b. & \xi > 0. \end{cases} \tag{12}$$

$$\omega_{photo} = \begin{cases} 2.2, & \xi \le 0, \\ c\xi + d. & \xi > 0. \end{cases} \tag{13}$$

3.3. Temporal Supervision for Video Data

In addition to monocular image reconstruction, we aim for our model to perform well on videos. However, existing methods often require additional smoothing processing after frame-by-frame reconstruction to reduce jitters in videos. Instead of relying on additional processing, we propose an end-to-end method that achieves superior results on videos. By inputting three adjacent frames $t-1$, t, and $t+1$ into a three-layer temporal Transformer, we obtain a set of estimated FLAME tokens. These tokens are then used to reconstruct the face of frame t, improving the continuity of facial motion between adjacent frames. The corresponding framework is illustrated in Fig. 4.

4. Experiment

In this section, we present the results of our method and compare it with other approaches. Due to page limitations, we provided additional content including more results and video demos in our *supplementary materials*.

4.1. Datasets

3D Datasets. In order to reconstruct accurate metric 3D faces, we collect the currently available open-source 3D scan datasets (FLAME topology) for supervised learning of estimating shape parameters from RGB images. As in MICA [50], we use neutral face data (RGB images and the corresponding neutral face shape parameters) in unified metric space from Stirling [19], Florence [2], LYHM [12], FaceWarehouse [7], and FRGC [33]. In addition, we include 3D face datasets with expressions from FaceWarehouse [7] and WCPA [24].

2D Datasets. Besides 3D data, we use common 2D face datasets (images only) in our self-supervised training, including FFHQ [25], FaceScape [45], BUPT-Balanced [42], and VoxCeleb2 [11], CelebA [30], VGGFace2 [8]. We adopt AFLW2000 [46] as the validation set.

4.2. Implementation Details

Training Details. We implement our method with PyTorch[14]. We use Adam [26] as the optimizer, in which the learning rate is set to 1×10^{-4}. We resize the input image to the size of 224×224. The model is trained in 10 epochs with batch size 8. We use FaRL [47] as the pre-trained weight to initialize the transformer backbone. The facial component tokens are all initialized to 0. In our experiments, we set hyper-parameters $\lambda_{3D} = 0.6$ and $\lambda_{2D} = 0.4$ in Eq. (1). We set $\lambda_{mesh} = 2.0$ and $\lambda_{vc} = 1.2$ in Eq. (2). In Eq. (6), we set $\lambda_{eyes} = 0.8$, $\lambda_{lips} = 1.0$. In Eq. (11), we set $\omega_{\boldsymbol{\alpha}} = \omega_{\boldsymbol{\beta}} = \omega_{\boldsymbol{\psi}} = 1\times10^{-4}$. In Eq. (12) and Eq. (13), we empirically set $a = -0.76$, $b = 1.6$, $c = 0.38$, and $d = 2.2$.

Evaluation Dataset. In the field of monocular image 3D reconstruction, we usually use the median, mean and standard deviation of the distance of each point to characterize the performance of the reconstruction method, after aligning the predicted mesh reconstructed from the specified image with the meshes of the real scan. Among them, the two commonly used benchmarks are NoW [35] and Stirling [19].

4.3. Comparison with Existing Methods

Quantitative Results. We conducted a quantitative evaluation of our method over the NoW and Stirling benchmarks. The results on the NoW benchmark are presented in Table 2, while the results on the Stirling benchmark are shown in Table 3. It is worth noting that we excluded the Stirling data from our training in Table 3 and train a new model from the beginning for a fair comparison. Our method achieves the top performance with the least errors across all metrics on both benchmarks, as demonstrated in the tables. Notably, our method outperforms the previous best method (MICA) by a remarkable margin.

Qualitative Results. In Fig. 5, we present a visual comparison of 3D face reconstruction results from different representative methods that are publicly available. Our method demonstrates robustness in accurately reconstructing faces across different shapes, races, and ages, as shown in the figure. Notably, our method achieves the best mesh recovery of face shapes and accurately captures expressions.

4.4. Face Tracking Results

We conducted a comparison on video data to validate that our method can produce faithful 3D face reconstruction and temporal coherent results with our temporal module. Specifically, we select a sample video clip and compare the photo-metrical reconstruction errors with the original video frames. To ensure consistency in the comparison, we use FLAME texture for mapping to compare the consistency of the face at the mesh level and the coherence of movement. Fig. 6 presents the error curves of video reconstruction for different approaches. Our method outperforms the other methods in terms of the accuracy and coherence of the reconstructed video.

5. Ablation Study

To ensure a fair comparison of model performance in our ablation experiments, we employ the validation dataset of the NoW dataset [35] as quantitative test data to evaluate the reconstruction performance. We use the 3D face reconstruction error as the evaluation metric.

Ablation Study on the Effect of Separate Tokens. Previous 3D face reconstruction models typically use a single, long code and split it into different component parameters, resulting in coupling between different parameters and decreased reconstruction accuracy. In our study, we investigate the effectiveness of using separate tokens for different facial parameters in our TokenFace model. We perform an ablation study on the number of tokens by merging closely related tokens, and present the results in Table 4. Our experiment reveals that merging different parameter types into the same token increases the coupling effect between parameters and leads to poorer reconstruction. Conversely, increasing the number of tokens improves decoupling and reconstruction accuracy. Our findings demonstrate the strong decoupling effect of TokenFace's separate tokens for different FLAME parameters, resulting in greater independence and specificity in expressing different components.

Ablation Study of Dataset Types for Training. To take advantage of the flexibility of our proposed network structure

Method	Non-Metrical			Metrical		
	Median↓	Mean↓	Std↓	Median↓	Mean↓	Std↓
3DMM-CNN [39]	1.84	2.33	2.05	3.91	4.84	4.02
FLAME 2020 template [28]	1.21	1.53	1.31	1.49	1.92	1.68
PRNet [18]	1.5	1.98	1.88	-	-	-
Deep3DFaceRecon [Tensorflow] [14]	1.23	1.54	1.29	2.26	2.90	2.51
RingNet [35]	1.21	1.53	1.31	1.50	1.98	1.77
Deep3DFaceRecon [PyTorch] [14]	1.11	1.41	1.21	1.62	2.21	2.08
MGCNet [36]	1.31	1.87	2.63	1.70	2.47	3.02
3DDFA-v2 [21]	1.23	1.57	1.39	1.53	2.06	1.95
SynergyNet [44]	1.27	1.59	1.31	2.28	2.86	2.39
DECA [17]	1.09	1.38	1.18	1.35	1.80	1.64
Dense Landmark [43]	1.02	1.28	1.08	1.36	1.73	1.47
MICA [50]	0.90	1.11	0.92	1.08	1.37	1.17
TokenFace (Ours)	**0.76**	**0.95**	**0.82**	**0.97**	**1.24**	**1.07**

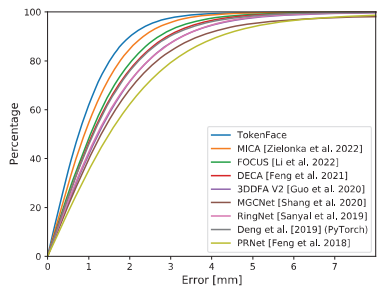

Table 2: **Quantitative comparisons on NoW benchmark** [35]. The metric is the 3D face reconstruction error. Best results are highlighted in bold.

Figure 5: **Visual comparison of 3D face reconstruction quality of ours and some other representative methods**. From top to bottom are input image, Deep3DFaceRecon [14], 3DDFAv2 [21], DECA [17], and TokenFace (Ours).

Method	Median↓	Mean↓	Std↓
FLAME 2020 template [28]	1.22	1.55	1.35
RingNet [35]	1.15	1.46	1.27
Deep3DFaceRecon [TensorFlow] [14]	1.13	1.43	1.25
Deep3DFaceRecon [Pytorch] [14]	0.99	1.27	1.15
3DDFA-v2 [21]	1.20	1.55	1.45
DECA [17]	1.03	1.32	1.18
MICA [50]	0.92	1.16	1.04
TokenFace (Ours)	**0.88**	**0.95**	**0.96**

Table 3: **Quantitative results on Stirling benchmark** [19].

in selecting different types of tokens, we conducted mixed training using a hybrid dataset that contains 2D and 3D data. To investigate the impact of dataset type on performance, we conduct ablation experiments using 2D-only, 3D-only, and mixed datasets for training. In the case of 3D-only datasets, we use a loss function at the vertex level, with camera pose, texture, and light values set to zero in the six tokens. Results presented in Table 5 show that the models trained using only 3D data outperformed those trained using only 2D data on the NoW validation set due to the higher accuracy of ground truths at the vertex level. Moreover, models trained with mixed datasets outperform those trained with only 3D datasets, attributed to the increased quantity of data and the multiple types of supervision provided by the mixed dataset.

Ablation Study on Weighting between 2D and 3D Data. Since our training dataset consists of both 2D and 3D data, we want to investigate the impact of adjusting the weight between them. To this end, we conduct experiments with different weightings while keeping other conditions fixed and the results are shown in Table 6. As shown, the weight balance between 2D and 3D data does have an effect on the

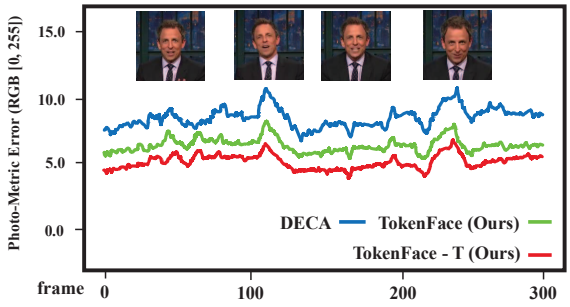

Figure 6: **Visualization of face tracking error on a video clip**. As there is no ground truth 3D face mesh, we calculate the photometric error. TokenFace–T denotes TokenFace with the temporal module.

#Tokens	Merging Detail	Reconstruction Error		
		Median	Mean	Std
1	[S, E, J, C, T, L]	1.08	1.41	1.19
3	[S, E, J, C], T, L	1.05	1.31	1.10
5	[S, E], J, C, T, L	0.96	1.15	1.02
6	S, E, J, C, T, L	**0.79**	**0.99**	**0.85**

Table 4: **Ablation study on the number of tokens.** Reducing the number of tokens by merging leads to a drop in reconstruction performance. The six tokens, denoted as **S** (shape), **E** (expression), **J** (jaw pose), **C** (camera pose), **T** (texture), and **L** (lighting), derived from the specific parameters of the FLAME model, are crucial for the model's ability to decouple different facial parameters and achieve accurate reconstructions. [·] represents the merging operation.

2D	3D	Reconstruction Error					
		Median ↓	Δ	Mean ↓	Δ	Std. ↓	Δ
✓		1.03	+0.24	1.31	+0.32	1.11	+0.26
	✓	0.85	+0.06	1.08	+0.09	0.89	+0.04
✓	✓	**0.79**	-	**0.99**	-	**0.85**	-

Table 5: **Ablation study on using different types of datasets.** We train our model on three types of datasets: 2D data only, 3D data only, and hybrid datasets of 2D and 3D. The model trained on mixed datasets of 2D and 3D outperforms those trained on either 2D or 3D data alone.

final results, and careful selection of the weight is required. We found that using a weight of (0.4, 0.6) result in the highest performance. This is our setting in reporting the main results.

Ablation Study on Adaptive Loss Function. To improve the effectiveness of large pose face reconstruction during training on 2D data, we introduce a pose-aware loss function in Sec. 3.2.2. To demonstrate the effectiveness of this loss function, we conduct an ablation study by comparing the models trained with and without the proposed loss, and then test the model on large pose examples. As shown in Fig. 7,

ω_{2D}	ω_{3D}	Reconstruction Error		
		Median	Mean	Std
0.3	0.7	0.83	1.04	0.89
0.4	**0.6**	**0.79**	**0.99**	**0.85**
0.5	0.5	0.81	1.02	0.87
0.6	0.4	0.88	1.10	0.93

Table 6: **Ablation study on different balance weights on 2D and 3D data**. Based on this table, we select the best weighting parameter $(0.4, 0.6)$ in our experiments.

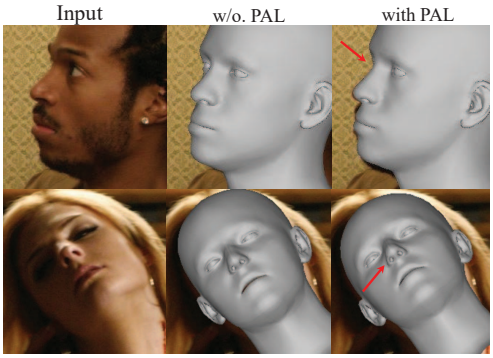

Figure 7: **Visualization of Effects of using Adaptive Weights**. We test two large pose cases with side face and head tilt. It can be seen that the model with adaptive loss function weights has more accurate alignment for large pose face reconstruction.

despite achieving similar reconstruction effects for shape and expression, the model trained with adaptive loss function weights produces more accurate reconstructions in pose for faces with large poses, such as side faces and head tilts.

6. Conclusion

In this paper, we introduce TokenFace, a transformer-based method for reconstructing 3D faces from monocular images. By using six independent facial component tokens to estimate six parameters that represent the reconstructed face, we are able to disentangle the different parameters of FLAME. Additionally, we design a temporal transformer that captures temporal information in videos, resulting in significantly improved accuracy and continuity in video face reconstruction. Our TokenFace achieves state-of-the-art performance in the challenging NoW Benchmark [35] and Stirling Benchmark [19], with a large improvement over previous methods. Furthermore, our method demonstrates stable and accurate performance in video face tracking with our temporal modeling.

Acknowledgement. This work was supported by the National Key R&D Program of China(2022YFB4701400 / 4701402), the SZSTC Grant(JCYJ20190809172201639, WDZC20200820200655001), the Shenzhen Key Laboratory(ZDSYS20210623092001004).

References

[1] Timur Bagautdinov, Chenglei Wu, Jason Saragih, Pascal Fua, and Yaser Sheikh. Modeling facial geometry using compositional vaes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3877–3886, 2018. 2

[2] Andrew D. Bagdanov, Alberto Del Bimbo, and Iacopo Masi. The florence 2d/3d hybrid face dataset. In *Proceedings of the 2011 Joint ACM Workshop on Human Gesture and Behavior Understanding*, J-HGBU '11, page 79–80, New York, NY, USA, 2011. ACM. 6

[3] Linchao Bao, Xiangkai Lin, Yajing Chen, Haoxian Zhang, Sheng Wang, Xuefei Zhe, Di Kang, Haozhi Huang, Xinwei Jiang, Jue Wang, et al. High-fidelity 3d digital human head creation from rgb-d selfies. *ACM Transactions on Graphics (TOG)*, 41(1):1–21, 2021. 2

[4] Shubhajit Basak, Peter Corcoran, Rachel McDonnell, and Michael Schukat. 3d face-model reconstruction from a single image: A feature aggregation approach using hierarchical transformer with weak supervision. *Neural Networks*, 156:108–122, 2022. 2

[5] 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. 2

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

[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, 6

[8] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018. 6

[9] Anpei Chen, Zhang Chen, Guli Zhang, Kenny Mitchell, and Jingyi Yu. Photo-realistic facial details synthesis from single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9429–9439, 2019. 2

[10] Zhuo Chen, Yuesong Wang, Tao Guan, Luoyuan Xu, and Wenkai Liu. Transformer-based 3d face reconstruction with end-to-end shape-preserved domain transfer. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12):8383–8393, 2022. 2

[11] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *Cornell University - arXiv*, 2018. 6

[12] Hang Dai, Nick Pears, William Smith, and Christian Duncan. Statistical modeling of craniofacial shape and texture. *International Journal of Computer Vision*, 128:547–571, 2020. 6

[13] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 5

[14] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1, 2, 6, 7

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[16] 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, 3

[17] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (ToG)*, 40(4):1–13, 2021. 1, 2, 3, 4, 5, 7

[18] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *Proceedings of the European conference on computer vision (ECCV)*, pages 534–551, 2018. 2, 7

[19] Zhen-Hua Feng, Patrik Huber, Josef Kittler, Peter Hancock, Xiao-Jun Wu, Qijun Zhao, Paul Koppen, and Matthias Rätsch. Evaluation of dense 3d reconstruction from 2d face images in the wild. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 780–786. IEEE, 2018. 6, 7, 8

[20] Leonardo Galteri, Claudio Ferrari, Giuseppe Lisanti, Stefano Berretti, and Alberto Del Bimbo. Deep 3d morphable model refinement via progressive growing of conditional generative adversarial networks. *Computer Vision and Image Understanding*, 185:31–42, 2019. 2

[21] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3d dense face alignment. In *European Conference on Computer Vision*, pages 152–168. Springer, 2020. 2, 7

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

[23] Loc Huynh, Weikai Chen, Shunsuke Saito, Jun Xing, Koki Nagano, Andrew Jones, Paul Debevec, and Hao Li. Mesoscopic facial geometry inference using deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8407–8416, 2018. 2

[24] Yueying Kao, Bowen Pan, Miao Xu, Jiangjing Lyu, Xiangyu Zhu, Yuanzhang Chang, Xiaobo Li, Zhen Lei, and Zixiong Qin. Single-image 3d face reconstruction under perspective projection. *arXiv preprint arXiv:2205.04126*, 2022. 6

[25] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In

Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 4401–4410, 2019. 6

[26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[27] Ruilong Li, Karl Bladin, Yajie Zhao, Chinmay Chinara, Owen Ingraham, Pengda Xiang, Xinglei Ren, Pratusha Prasad, Bipin Kishore, Jun Xing, et al. Learning formation of physically-based face attributes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3410–3419, 2020. 2

[28] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Transactions on Graphics*, 2017. 2, 7

[29] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 2

[30] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018):11, 2018. 6

[31] Thomas Neumann, Kiran Varanasi, Stephan Wenger, Markus Wacker, Marcus Magnor, and Christian Theobalt. Sparse localized deformation components. *ACM Transactions on Graphics (TOG)*, 32(6):1–10, 2013. 2

[32] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *2009 sixth IEEE international conference on advanced video and signal based surveillance*, pages 296–301. Ieee, 2009. 2

[33] P Jonathon Phillips, Patrick J Flynn, Todd Scruggs, Kevin W Bowyer, Jin Chang, Kevin Hoffman, Joe Marques, Jaesik Min, and William Worek. Overview of the face recognition grand challenge. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 947–954. IEEE, 2005. 6

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

[35] 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)*, pages 7763–7772, June 2019. 6, 7, 8

[36] Jiaxiang Shang, Tianwei Shen, Shiwei Li, Lei Zhou, Mingmin Zhen, Tian Fang, and Long Quan. Self-supervised monocular 3d face reconstruction by occlusion-aware multi-view geometry consistency. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV*, pages 53–70. Springer, 2020. 7

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

[38] Luan Tran, Feng Liu, and Xiaoming Liu. Towards high-fidelity nonlinear 3d face morphable model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1126–1135, 2019. 2

[39] Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gérard Medioni. Regressing robust and discriminative 3d morphable models with a very deep neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5163–5172, 2017. 7

[40] 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

[41] Daniel Vlasic, Matthew Brand, Hanspeter Pfister, and Jovan Popovic. Face transfer with multilinear models. In *ACM SIGGRAPH 2006 Courses*, pages 24–es. 2006. 2

[42] Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. *International Conference on Computer Vision*, 2019. 6

[43] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Matthew Johnson, Jingjing Shen, Nikola Milosavljević, Daniel Wilde, Stephan Garbin, Toby Sharp, Ivan Stojiljković, et al. 3d face reconstruction with dense landmarks. In *European Conference on Computer Vision*, pages 160–177. Springer, 2022. 2, 7

[44] Cho-Ying Wu, Qiangeng Xu, and Ulrich Neumann. Synergy between 3dmm and 3d landmarks for accurate 3d facial geometry. In *2021 International Conference on 3D Vision (3DV)*, pages 453–463. IEEE, 2021. 7

[45] 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. 2, 6

[46] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Towards large-pose face frontalization in the wild. In *In Proceeding of International Conference on Computer Vision*, Venice, Italy, October 2017. 6

[47] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18697–18709, 2022. 3, 6

[48] Yaoyao Zhong and Weihong Deng. Face transformer for recognition. *arXiv preprint arXiv:2103.14803*, 2021. 2

[49] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 146–155, 2016. 2

[50] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. In *European Conference on Computer Vision (ECCV)*. Springer International Publishing, Oct. 2022. 1, 2, 6, 7