AU-aware 3D Face Reconstruction through Personalized AU-specific Blendshape Learning

Chenyi Kuang[1], Zijun Cui[1], Jeffrey O. Kephart[2], and Qiang Ji[1]

[1] Rensselaer Polytechnic Institute
{kuangc2,cuiz3,jiq}@rpi.edu
[2] IBM Thomas J. Watson Research Ctr.
kephart@us.ibm.com

Abstract. 3D face reconstruction and facial action unit (AU) detection have emerged as interesting and challenging tasks in recent years, but are rarely performed in tandem. Image-based 3D face reconstruction, which can represent a dense space of facial motions, is typically accomplished by estimating identity, expression, texture, head pose, and illumination separately via pre-constructed 3D morphable models (3DMMs). Recent 3D reconstruction models can recover high-quality geometric facial details like wrinkles and pores, but are still limited in their ability to recover 3D subtle motions caused by the activation of AUs. We present a multi-stage learning framework that recovers AU-interpretable 3D facial details by learning personalized AU-specific blendshapes from images. Our model explicitly learns 3D expression basis by using AU labels and generic AU relationship prior and then constrains the basis coefficients such that they are semantically mapped to each AU. Our AU-aware 3D reconstruction model generates accurate 3D expressions composed by semantically meaningful AU motion components. Furthermore, the output of the model can be directly applied to generate 3D AU occurrence predictions, which have not been fully explored by prior 3D reconstruction models. We demonstrate the effectiveness of our approach via qualitative and quantitative evaluations.

Keywords: 3DMM, 3D Face Reconstruction, Facial Action Unit

1 Introduction

With the first 3D morphable model (3DMM) proposed by Blanz et al. [7], 3D face modeling and reconstruction have gained sustained attention as a research topic and are frequently used in popular applications like AR/VR, 3D avatar animation, video games and communication. Most existing 3D face reconstruction methods [62,18,42,44] are based on a pre-constructed 3DMM, which divides the space of 3D facial shapes into identity and expression dimensions. The 3DMMs are constructed from a large database of 3D scans. Among various representations, orthogonal PCA basis are widely used such as BFM [22] and FLAME [30] and blendshapes, such as FaceWarehouse [9], FaceScape [53] and Feafa [52]. While the PCA basis usually represents global vertex deformations and does

not have semantic meaning, each blendshape basis represents local deformations related to specific facial movements. Responding to increasing demand for generating high quality 3D faces, more recent works [21,5,4,53] have recovered detailed facial geometry from monocular or multi-view images by dynamically learning a displacement map to add a refinement layer to the 3D geometry. However, these approaches to 3D face reconstruction have not fully considered the inherent 3D nature of facial movement, which is based upon the muscle activation under a local region of the skin.

The Facial Action Coding System (FACS) [20] encodes facial expressions in terms of activation levels of action units (AUs), each of which corresponds to a specific underlying group of facial muscle movements. Thus AU intensities can provide a useful basis for interpreting emotions or other reflections of human internal state such as pain. The definition of AU is widely used in image-based facial behaviour analysis tasks such as AU detection. Many computational methods [14,38,26,54] have been developed for directly inferring AU activation from appearance features. Except for image-based AU detection, combining AU with 3D faces has been explored in terms of learning 3D AU classifiers from 3D scans in BP4D dataset [56] and synthesizing facial images from a specified set of AU activation through a generative model [47,34].

Little work, however, has been devoted to constructing the FACS-based correspondence between 3DMM basis and AUs. Such line of work is important for two reasons. First, semantically mapping the 3DMM basis to each AU can help reconstruct 3D subtle motions caused by AU activation, which in turn reflects the underlying muscle activation. Second, spatial relationships among AUs derived from general anatomical knowledge [33,26,57,58] can be incorporated into the 3D geometric basis generating process. Achieving this type of coherency between AU and 3D face models benefits both synergistically: incorporating AU information in 3D geometry helps capture more accurate facial motions, while incorporating geometric information helps detect challenging AUs.

Based on the considerations above, we propose a multi-stage training framework that generates a finer-grained blendshape basis that is specific to each input subject and each AU. We explicitly apply AU labels and a pre-learned Bayesian Network that captures generic inter-relationships among AUs to constrain the 3D model coefficients during training. Our main original contributions include:

- We propose a deep learning framework that explicitly learns personalized and AU-explainable 3D blendshapes for accurate 3D face reconstruction.
- We perform a multi-stage training process and utilize AU labels and a generic AU relationship prior to constrain the AU specific blendshape learning process, using a mixture of AU-labeled and unlabeled images.
- Our model simultaneously generates an AU basis and a realistic 3D face reconstruction with improved capture of subtle motions. We achieve state-of-art 3D reconstruction results on BU3DFE [55] and the Now Challenge [42] dataset. Moreover, our model directly performs 3D AU detection because the 3D coefficients are readily mapped to AU activation probabilities.

2 Related Works

Personalized 3D Face Models. 3DMMS were first proposed [7] to model in three dimensions human variations in facial shape, expression, skin color, etc. Since then, 3DMMs have been extended and used widely for 3D face representation or reconstruction tasks. They are usually constructed from a large 3D database and have separate bases for identity, expression, texture, etc. The Basel Face Model(BFM) [40], FLAME [30] and LSFM [8] are popular 3DMMs for 3D reconstruction tasks. More recently, FaceScape [53] and ICT-Face [29] have been proposed as "high-resolution" 3D face models that provide meticulous geometric facial details. Some researchers have attempted to dynamically generate or update accurate 3DMMs directly from images or videos. Tewari et al.[48] learned a 3D facial shape and appearance basis from multi-view images of the same person. Later, Tewari et al. extended their work to learn a more complete 3DMM including shape, expression and albedo from video data [49]. Chaudhuri et al. [11] proposed an alternative that learns a personalized 3D face model by performing on-the-fly updating on the shape and expression blendshapes.

3DMM-based Face Reconstruction. Other researchers have explored using a pre-constructed 3DMM to perform 3D face reconstruction from monocular images. 3DMM parameters representing camera, head pose, identity, expression, texture and illumination are estimated via regression. 3DMM-based face reconstruction models have achieved great performance improvements in head pose estimation [61,51,1], 3D face alignment [61,23,1] and facial detail recovery [5,4,62,21]. Moreover, Chang et al. [10] utilize 3DMM to reconstruct 3D facial expressions and apply the expression coefficients for expression recognition. Feng et al. [21] propose to learn an animatable detailed 3D face model that can generate expression-dependent geometric details such as wrinkles. However, using a 3D face reconstruction model to capture subtle local facial motions caused by activation of AUs has not been explored previously.

3D AU Modeling and Detection. 3DMMs are closely related to 3D expression/AU modeling and synthesis. Liu et al. [35] applied 3DMMs to train an adversarial network to synthesize AU images with given AU intensities. Song et al. [47] conducted an unsupervised training scheme to regress 3D AU parameters and generate game-like AU images through differentiable rendering. Li et al. [29] constructs a non-linear 3DMM with expression shapes closely related to FACS units and can be used to produce high-quality 3D expression animation.

Except for 3D AU synthesis, AU detection from 3D data have been actively studied by researchers and 3D scans, 3D point clouds, or 3DMMs can be used. Given a target 3D mesh or scan, classifiers can be trained based on the extracted mesh surface features for 3D AU detection [28,25,60,43,6,17]. Similarly, for 3D point cloud data, Reale et al. [41] trained a network to directly extract 3D point cloud features and support AU detection. Tulyakov et al. [50] learned a pose invariant face representation from the point cloud for more robust AU detection. Ariano et al. [3] propose a method of 3D AU detection by using 3DMM coefficients where they first remove the identity component from their SLC-3DMM and then train a classifier on 3D meshes for AU detection.

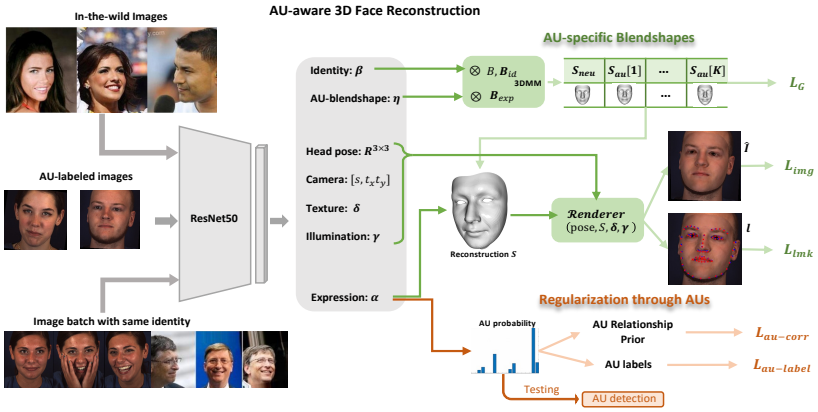

Fig. 1: Overview of our model structure and pipeline: (1) (grey) Basic module for regressing all 3D reconstruction parameters (2) (green) Pipeline for 3D face reconstruction, including subject neutral face generation, AU-specific blendshape construction and differentiable rendering process (3) (orange) Module for integrating AU prior knowledge and applying AU labels as constraints on $\boldsymbol{\alpha}$. During training the output AU probabilities are used to compute two AU regularization losses; during testing the output can be directly used for 3D AU detection.

Our model differs from the above methods in that we construct person-specific and AU-specific 3D models for AU-interpretable face reconstruction without requiring any ground-truth 3D data for an input subject.

3 Proposed Method

In general 3D face reconstruction, the shape and expression spaces are spanned by pre-constructed bases. A target 3D face $\boldsymbol{S}$ can be represented by:

$$\boldsymbol{S} = \bar{B} + \boldsymbol{B}_{id}\boldsymbol{\beta} + \boldsymbol{B}_{exp}\boldsymbol{\alpha} \tag{1}$$

where $\bar{B}$ is the 3DMM mean shape, $\boldsymbol{B}_{id}, \boldsymbol{B}_{exp}$ are shape bases and expression bases; $\boldsymbol{\beta}, \boldsymbol{\alpha}$ are the vectors of shape and expression coefficients respectively.

The texture of S can also be represented by a linear model:

$$\boldsymbol{T} = \bar{\boldsymbol{T}} + \boldsymbol{B}_{tex}\boldsymbol{\delta} \tag{2}$$

where $\boldsymbol{B}_{tex}$ is a texture basis (usually PCA) and $\boldsymbol{\delta}$ is a texture coefficient vector.

Our method, illustrated in Fig. 1, significantly extends the general 3DMM of Eq. 1. It contains three modules: (1) the basic module for regressing all 3D reconstruction parameters, including pose and camera parameters, identity coefficients $\boldsymbol{\beta}$, AU-blendshape coefficients $\boldsymbol{\eta}$, expression coefficients $\boldsymbol{\alpha}$, texture coefficients $\boldsymbol{\delta}$ and illumination parameters $\boldsymbol{\gamma}$; (2) the module for constructing AU-specific blendshapes; (3) the module for incorporating AU regularization terms and AU detection. The remainder of this section discusses each of these in turn.

3.1 AU-specific 3D Model Learning

We describe the construction of an identity-consistent subject-neutral face and AU-specific blendshapes.

Identity-consistent parameters learning. With the regressed identity coefficients $\boldsymbol{\beta}$, the neutral blendshape for a subject is calculated by:

$$S_{neu} = \bar{B} + \boldsymbol{B}_{id}\boldsymbol{\beta} \tag{3}$$

Our model focuses on analysing 3D facial motions in sequence data like BP4D [56]. We want to reduce the identity bias during reconstruction for multiple frames of the same subject and therefore we can focus more on the non-rigid facial expressions. For the purpose of better estimating a subject-neutral face, we employ the identity-consistent constraint on the learning of identity parameter $\boldsymbol{\beta}$. We sample a small batch of F images $I_1, \cdots, I_F$ for the same subject according to the identity labels and build F Siamese structures with shared network weights. For multi-frame input $[I_1, \cdots, I_F]$, from the output of Siamese structures, we extract identity coefficients $[\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_F]$ and enforce the similarity between every pair of identity coefficients by the following identity-consistency loss:

$$L_{id} = \frac{1}{2F(F-1)} \sum_{i,j=1}^{F} \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_2 \tag{4}$$

AU-specific Blendshape Learning. A linear 3DMM usually contains an identity basis $\boldsymbol{B}_{id}$ and expression basis $\boldsymbol{B}_{exp}$, as expressed in Eq. 1. However, to adapt the 3D reconstruction model to an AU detection task, we want to construct an AU-specific blendshape for each of K blendshapes. For this purpose, we introduce an additional set of person-specific coefficients $\boldsymbol{\eta} = [\boldsymbol{\eta}_1, \cdots, \boldsymbol{\eta}_K]$ to compose AU-specific blendshapes $\boldsymbol{B}_{au} = [[B_{au}[1], \cdots, B_{au}[K]]$.

$$B_{au}[k] = \sum_{m=1}^{M} \eta_{k,m} W_k \odot \boldsymbol{B}_{exp}[m], \forall k \in 1, ..., K \tag{5}$$

where M is the number of expression bases and $B_{au}[k]$ is the subject blendshape for AU_k, and W_k is the pre-generated mask constraining local deformation in certain face regions for each AU. Each $\boldsymbol{\eta}_k$ has the dimension of the number of expression bases and $\eta_{k,m}$ is the m_{th} factor of $\boldsymbol{\eta}_k$. For each input image, our model generates K subject-AU-specific blendshapes by predicting the AU coefficients $\boldsymbol{\eta}$. By linearly combining the AU-specific blendshapes with the expression coefficients $\boldsymbol{\alpha}$, a 3D face can be expressed as:

$$S = S_{neu} + \boldsymbol{B}_{au}\boldsymbol{\alpha} \tag{6}$$

Combining Eq. 5 and Eq. 6, we find that $\boldsymbol{\alpha}$ and $\boldsymbol{\eta}$ are coupled together to compose a 3D face shape and thus there may exist multiple solutions of $\boldsymbol{\alpha}$ and $\boldsymbol{\eta}$, making it impossible to properly match blendshapes with AUs. To jointly learn $\boldsymbol{\eta}$ and $\boldsymbol{\alpha}$, we constrain $\boldsymbol{\alpha}$ by directly mapping it to AU occurrence probability.

3.2 AU-aware 3D Face Reconstruction

As depicted in Fig. 1, the expression coefficients $\boldsymbol{\alpha}$ are the key factor to adapt the AU information to 3D geometry. In this section, we will demonstrate the process of applying two AU-related regularization terms as constraints of learning $\boldsymbol{\alpha}$.

AU Activation Probability. Before introducing AU-related regularization terms, we design mapping functions from 3D expression parameters $\boldsymbol{\alpha}$ to AU occurrence probabilities. The intuition underlying the design of the mapping function is that $\boldsymbol{\alpha}$ represents the deformation intensity in 3D space of each blendshape component, which is closer to the real-world AU activation defined in FACS [20] compared with 2D image features. Therefore, it is natural to apply a threshold on each dimension of $\boldsymbol{\alpha} = [\alpha_1, \cdots, \alpha_K]^T$ to determine whether the corresponding AU is activated. We denote the activation status of AU_i as z_i with $z_i = \{0, 1\}$, and denote the probability of its occurrence as p_i with $p_i = p(z_i = 1)$. The mapping function $f : \alpha_i \rightarrow p_i$ then becomes:

$$p_i = p(z_i = 1) = \sigma(\frac{\alpha_i - \tau}{\epsilon}) \tag{7}$$

where $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function. τ is a threshold on expression intensity and ϵ is the temperature factor. Both τ and ϵ are pre-defined hyper-parameters. With the mapping in Eq. 7, AU-related regularization terms based on AU occurrence probability are then applied to optimize $\boldsymbol{\alpha}$ by means described in detail below.

Regularization Through AUs. To learn semantically meaningful $\boldsymbol{\alpha}$, we derive constraints from both AU occurrence labels and generic prior knowledge on AU relationships, whereby the former applies to AU-annotated images and the latter applies to any dataset with various subjects. These constraints are incorporated into the proposed AU-specific blendshape learning as regularization terms.

A. Regularization with AU labels For images with available AU occurrence labels, we define a loss function based on the cross-entropy between the ground-truth label and predicted occurrence probability p_i:

$$L_{au-label} = -\frac{1}{|C|}\sum_{i \in C}(z_i^{gt} log(p_i) + (1 - z_i^{gt}) log(1 - p_i)) \tag{8}$$

where z_i^{gt} is the ground-truth label for AU_i and C is the set of annotated AUs.
B. Regularization with generic AU relationships. AU labels are not available for all training data. To further constrain the learning of $\boldsymbol{B}_{au}$ and $\boldsymbol{\alpha}$, we consider generic AU relationships defined by FACS [20]. AUs can be positively or negatively correlated, depending on the underlying muscle anatomy. We summarize the most commonly considered AU relationships in Table. 1. These correlations are generic and are applicable across different subjects. AU correlations can be represented as probability inequality constraints [57]. For a positive correlated

Table 1: AU Correlation

AU relations	AU pairs
positive correlated	(1,2), (4,7) , (4,9) (7,9),(6,12), (9,17) (15,17), (15,24), (17,24), (23,24)
negative correlated	(2,6), (2,7), (12,15), (12,17)

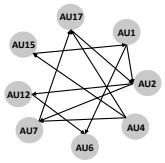

Fig. 2: Pre-learned BN structure

AU pair (i, j), given the occurrence of AU j, the probability of the occurrence of AU i is larger than the probability of its absence,

$$p(z_i = 1|z_j = 1) > p(z_i = 0|z_j = 1) \tag{9}$$

Similarly, we can derive probability constraints for negative correlations [57].

A Bayesian Network (BN) can then be learned from generic probability constraints by following [16]. The learned BN captures the joint probability of all possible configurations of 8 AUs. Its structure is visualized in Fig. 2.

Instead of employing the joint probability of all possible AUs [16], we focus on local probabilities of pairs of AUs. By employing local pairwise probabilities, we regularize α_i and α_j to be similar (or distinct) if AU_i and AU_j are positively (or negatively) correlated. Specifically, we obtain the pairwise probability of AU pair (i, j), i.e., $\boldsymbol{p}_{ij}^{prior}(z_i, z_j)$, from the learned Bayesian Network by marginalizing out the remaining nodes. The pre-generated pairwise probabilities for different AU pairs, i.e., $\boldsymbol{p}_{ij}^{prior}(z_i, z_j)$, encode generic AU relationship knowledge and apply to all subjects. Let C_2 represent the set of AU pairs (i, j) of interest[3]. The pre-generated pairwise probabilities $\{\boldsymbol{p}_{ij}^{prior}(z_i, z_j)\}_{(i,j)\in C_2}$ are then integrated into the learning of $\boldsymbol{\alpha}$ through the proposed regularization term as follows:

$$L_{au-corr} = \frac{1}{|C_2|}\sum_{(i,j)\in C_2}(KL(p_i(z_i) \parallel \mathbb{E}_{p_j(z_j)}p^{prior}(z_i|z_j)) + KL(p_j(z_j) \parallel \mathbb{E}_{p_i(z_i)}p^{prior}(z_j|z_i))) \tag{10}$$

In Eq. 10, $p_i(z_i)$ and $p_j(z_j)$ are calculated with Eq. 7, $p^{prior}(z_j|z_j)$ are computed based on pairwise prior from the pre-learned BN: $p^{prior}(z_j|z_j) = \frac{p^{prior}(z_i,z_j)}{\sum_{z_i=0,1} p^{prior}(z_i,z_j)}$. Both $p_i(z_i)$ and $\mathbb{E}_{p_j(z_j)}p^{prior}(z_i|z_j)$ are Bernoulli Distribution. $KL(\cdot||\cdot)$ represents Kullback-Leibler divergence. By applying $L_{au-corr}$, AU prior relationships are leveraged to better learn AU-adaptive expression coefficients $\boldsymbol{\alpha}$.

With the AU regularization terms above, our personalized AU basis are consistent with AUs. Reciprocally, AU-specific blendshapes support convenient 3D AU detection. One can apply Eq. 7 to the expression coefficients $\boldsymbol{\alpha}$ during testing to obtain the AU activation probabilities, and regard the AU_i as activated only if $p_i > 0.5$. The impact on AU detection performance is evaluated in section. 4.

[3] We select a set of eight AU pairs mentioned in Table. 1 that are available in both the learned BN and the AU indices with $C_2 := \{(1, 2), (4, 7), (6, 12), (15, 17), (2, 6), (2, 7), (12, 15), (12, 17)\}$

3D Face Reconstruction The output of our model are 3D reconstruction parameters $\boldsymbol{\rho} = [s, \boldsymbol{R}, \boldsymbol{t}_{2d}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\eta}, \boldsymbol{\delta}, \boldsymbol{\gamma}]$, where s is the scaling factor; $\boldsymbol{R}$ is the rotation matrix for head pose pitch, yaw and roll, $\boldsymbol{t}_{2d}$ is the 2d translation vector, $\boldsymbol{\alpha}$ are the predicted blendshape coefficients; $\boldsymbol{\delta}$ are the texture coefficients and $\boldsymbol{\gamma}$ is the illumination parameter. A weak perspective projection of the reconstructed 3D face can be expressed by

$$\boldsymbol{X} = s * \boldsymbol{Pr} * \boldsymbol{R} * (\bar{B} + \boldsymbol{B}_{id}\boldsymbol{\beta} + \boldsymbol{B}_{au}\boldsymbol{\alpha}) + \boldsymbol{t} \tag{11}$$

where $\boldsymbol{Pr} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$ is the weak perspective projection matrix. From projected vertices, we can select 68 facial landmarks $\boldsymbol{l} \in \mathbb{R}^{68\times 2}$ and compute the projected landmark loss with ground-truth landmarks $\boldsymbol{l}^{gt}$ (detected by landmarks detectors) for 3D alignment.

For the illumination model, we follow the typical assumption that the face is a Lambertian surface and the illumination is modeled by Spherical Harmonics(SH) [37]. The color $\boldsymbol{c}$ of a vertex $\boldsymbol{x}_i$ with its surface normal $\boldsymbol{n}_i$ and texture $\boldsymbol{t}_i$ can be formulated as $\boldsymbol{c}(\boldsymbol{n}_i, \boldsymbol{t}_i|\gamma) = \boldsymbol{t}_i \sum_{k=1}^{K^2} \gamma_k \cdot \boldsymbol{H}_k(\boldsymbol{n_i})$, where $\boldsymbol{H}_k : \mathbb{R}^3 \to \mathbb{R}$ is SH basis functions, γ_k are the corresponding SH coefficients and $K = 3, \boldsymbol{\gamma} \in \mathbb{R}^{K^2}$.

With a differentiable renderer module, we generate a synthetic image $\hat{I}$ using the projected 3D vertices $\boldsymbol{X}$ along with the texture and illumination parameters, i.e., $\hat{I} = \mathcal{R}(\boldsymbol{X}, \boldsymbol{\delta}, \boldsymbol{\gamma})$. By computing the pixel-level differences between the original image I and $\hat{I}$, the loss is back propagated to update $\boldsymbol{\rho}$.

3.3 Training loss

We perform a three-stage training process, each stage progressively adding more terms to the loss function as described in Algorithm 1. For the first stage (training the 3D reconstruction), the model training is self-supervised with the identity-consistency constraint. The model takes image batches as input. At the second stage, we train our model only on AU-annotated data. For the last stage, we further fine-tune our model with AU priors introduced in Section 3.2.

The total training loss function for the final stage is:

$$\begin{aligned} L = &\lambda_{img} L_{img} + \lambda_{lmk} L_{lmk} + \lambda_{id} L_{id} + \lambda_G L_G + L_{sp} \\ &+ \lambda_{au-label} L_{au-label} + \lambda_{au-corr} L_{au-corr} \end{aligned} \tag{12}$$

The loss term L_{id} is defined in Eq. 4, the AU regularization terms $L_{au-label}, L_{au-corr}$ are defined in Eq. 8 and Eq. 10. $L_{img}, L_{lmk}, L_{id}, L_G, L_{sp}$ are described below.

Photometric loss. The main component of the training loss is the pixel-level loss between the synthetic image $\hat{I}$ and original image I, formulated as:

$$L_{img} = \frac{\sum_m A_m \|I - \hat{I}\|_2}{\sum_m A_m} \tag{13}$$

where A_m is pre-generated facial skin mask so that we only calculate the pixel difference in face region.

Algorithm 1 Multi-Stage Training Process

Stage1: Identity-aware baseline model training
Input: $B^0_{neu}, \boldsymbol{B}^0_{au}$,
$\{[I^{s_i}_1, \cdots, I^{s_i}_F], [\boldsymbol{l}^{s_i}_1, \cdots, \boldsymbol{l}^{s_i}_F]^{gt} [A^{s_i}_1, \cdots, A^{s_i}_F]\}^{N_2}_{i=1}$
Training Loss: $L_1 = \lambda_{img} L_{img} + \lambda_{lmk} L_{lmk} + \lambda_G L_G + \lambda_{id} L_{id} + \lambda_{sp} L_{sp}$
Stage2: AU-adaptive training with AU labels
Input: $\{I_i, \boldsymbol{l}^{gt}_i, A_i, \boldsymbol{z}^{gt}_i\}^{N_1}_{i=1}, B^0_{neu}, \boldsymbol{B}^0_{au}, \boldsymbol{p}^{prior}_{ij}$
Training Loss: $L_2 = \lambda_{img} L_{img} + \lambda_{lmk} L_{lmk} + \lambda_G L_G + \lambda_{au-label} L_{au-label} + \lambda_{sp} L_{sp}$
Stage3: AU-adaptive training with AU prior
Input: $\{I_i, \boldsymbol{l}^{gt}_i, A_i\}^{N_1}_{i=1}, B^0_{neu}, \boldsymbol{B}^0_{au}, \boldsymbol{p}^{prior}_{ij}$
Training Loss: $L_3 = \lambda_{img} L_{img} + \lambda_{lmk} L_{lmk} + \lambda_G L_G + \lambda_{sp} L_{sp} + \lambda_{au-corr} L_{au-corr}$

Projected landmark loss. We select 68 landmark vertices on the mesh model and define the landmark loss term as the distance between the projected 3D landmarks $\hat{\boldsymbol{l}}$ and the pre-generated results $\boldsymbol{l}$ from landmark detectors.

$$L_{lmk} = \|\boldsymbol{l} - \boldsymbol{l}^{gt}\|_2 \tag{14}$$

Deformation gradient loss. Except for the soft constraints of AU labels and AU prior knowledge on expression coefficients $\boldsymbol{\alpha}$, we further apply a regularization term on the generated $\boldsymbol{B}_{au}$ to ensure they are locally deformed and semantically mapped to each AU. Inspired by the work of Chaudhur et al. [11] and Li et al. [27], we pre-define a template neutral face $B^0_{neu}$the same number of AU-blendshape template $\boldsymbol{B}^0_{au}$ by performing Non-Rigis ICP [2] process from BFM template to ICT-Face [29] expression template and impose regularization term of deformation gradient similarity, formulated as below:

$$L_G = \sum_{k=1}^{K} \|\boldsymbol{G}_{(B_{au}[k] \to S_{neu})} - \boldsymbol{G}_{(B^0_{au}[k] \to B^0_{neu})}\|_F \tag{15}$$

where $\boldsymbol{G}_{(B_{au}[k] \to S_{neu})}$, is the deformation gradient between the k_{th} AU-blendshape and neutral shape for a specific subject following the calculation defined in [12]. The loss L_G is utilized to enforce that the learned $\boldsymbol{B}_{au}$ are not exaggerated and have similar local deformations as the template ICT blendshapes so that $\boldsymbol{B}_{au}$ can semantically match with each AU. Without this loss, $\boldsymbol{B}_{au}$ will be deformed in a free-form manner and lose the interpretability as AU-blendshapes.

Coefficients sparsity loss. To avoid generating implausible faces we apply a regularization term to constrain small 3D coefficients, including $\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\delta}$, which is denoted as the sparsity loss L_{sp} and it's based on the L_2 norm of the coefficients.

$$L_{sp} = \lambda_{sp,1} \|\boldsymbol{\beta}\|_2 + \lambda_{sp,2} \|\boldsymbol{\alpha}\|_2 + \lambda_{sp,3} \|\boldsymbol{\delta}\|_2 \tag{16}$$

The training process is described in Algorithm 1.

4 Experiments

Datasets. To fully assess the proposed model under different environments, we evaluate both face reconstruction performance and AU detection performance on the following benchmark datasets:

- **CelebA** [36] contains more than 200k celebrity images with annotations of 40 attributes. We utilize the full CelebA images and the corresponding landmark and identity annotation for training. The landmarks are used to crop and initially align the images; identity attributes are used to generate image triplets for the identity consistency training.
- **BP4D** [56] 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 encoded by binary AU labels and AU intensity labels. Around 140k frames with AU occurrence labels are employed for evaluation. Subject-exclusive three-fold cross-validation experiment protocol is employed for AU detection evaluation.
- **BU3DFE** [55] is a 3D facial expression database with 3D scans available for 2500 face shapes of neutral face and six expressions. We use BU3DFE for the evaluation of 3D reconstruction performance.
- **Now Challenge** [42] provides around 2k 2D multi-view images of 100 subjects categorized by neutral, expressional, occluded and selfie images. We employee Now dataset only for 3D reconstruction evaluation with provided ground-truth 3D scans.

Implementation Details. We use the BFM [22] 3DMM for face reconstruction. In Eq. 5, the deformation mask is generated by using expression shapes provided by the ICT-Face [29] model. We use NICP [2] tools to project each ICT-face vertex to the closest triangle of the BFM mesh. We first generate a binary weight map by thresholding the vertex deformation for each blendshape and then perform Gaussian smoothing on the boundary to generate W_k (visualizations are available in the supplementary material). We use CelebA and BP4D as our training data. At the first stage, we sampled small image batches from both databases with $F = 3$ in Eq. 4. In Eq. 7, we set the hyper-parameter $\tau = 0.15$ and $\epsilon = 0.05$ by grid-search. We assign $\epsilon = 0.01$, which is used to amplify the difference between $\boldsymbol{\alpha}$ and τ.

For BP4D, we generate multiple triplet samples for each subject and each sequence. We train the **baseline** model for five epochs. At the second stage, we split BP4D into 3 folds and select two folds for AU-supervised training. The trained model is denoted as **baseline+AU label** model. For the last stage, we perform AU-adaptive training on both datasets by employing AU prior regularization term. The final model is denoted as **baseline+AU label+AU prior** model. We use CelebA and two folds of BP4D images for training. We train 170k iterations with batch size of 30.

4.1 Evaluation of 3D Reconstruction Error

To validate the reconstruction efficiency of AU-specific blendshapes, we perform quantitative and qualitative comparisons of the reconstruction results against state-of-the-art weakly-supervised 3D face reconstruction methods.

Table 2: 3D Reconstruction Error on BU3DFE [55]

Methods	Mean (mm)	SD
Chaudhuri et al. [11]	1.61	0.31
Shang et al. [44]	1.55	0.32
Bai et al. [4]	1.21	**0.25**
FML [48]	1.78	0.45
Ours-final	**1.11**	0.28

Table 3: 3D Reconstruction Error on **Now** [42] validation set

Methods	Mean (mm)	SD
Shang et al. [44]	1.87	2.63
Dib et al. [19]	1.57	1.31
RingNet [42]	1.53	1.31
Deep3dFaceRecon-pytorch [18]	1.41	1.21
DECA [21]	1.38	**1.18**
Ours-final	**1.33**	1.21

Table 4: Ablation Study on 3D Reconstruction

Stages with different losses (`ResNet50`)	BU3DFE [55]		Now [42]	
	Mean (mm)	SD	Mean (mm)	SD
backbone	1.62	0.68	1.90	1.45
backbone+L_{id}	1.58	0.50	1.84	1.56
backbone+L_G	1.37	0.35	1.52	1.36
backbone+$L_{id}+L_G$ (**baseline**)	1.32	0.32	1.47	1.38
backbone+$L_{id}+L_G+L_{au-label}$	1.18	0.30	1.39	**1.20**
Ours-final: *backbone*+$L_{id}+L_G+L_{au-label}+L_{au-corr}$	**1.11**	**0.28**	**1.33**	1.21

Quantitative evaluation. We perform monocular face reconstruction evaluation on the BU3DFE and **Now** validation datasets. On the BU3DFE dataset, we follow the evaluation procedure of [48] and [5] to pre-generate a dense-correspondence map between each ground-truth scan and reconstructed 3D face, applying the Iterative-Closest-Point(ICP) algorithm using Open3d [59]. On the **Now** dataset, we compute the correspondence by following the RingNet evaluation procedure [42]. With the correspondence between each scan vertex and the BFM mesh surface, the reconstruction accuracy is evaluated in terms of vertex-to-plane mean square error (MSE) in units of mm and standard deviation (SD).

A. Comparison to State-of-the-art Methods. We compare our model against four state-of-the-art methods: Chaudhuri et. al [11], Shang et al. [44], Bai et al. [4] and FML [48] on both the BU3DFE and **Now** datasets. Results are shown in Table 2 and Table 3, respectively. On both benchmark datasets, our final model outperforms existing state-of-the-art methods. On BU3DFE, our model achieves an MSE of 1.11 MSE, which is 0.67 less than the MSE achieved by FML [48]. On the **Now** validation dataset, our final model's MSE is also better than the state-of-the-art methods [42,18,44,19]. We provide visualizations examples of error maps on BU3DFE and **Now** in the supplementary materials.

B. Ablation Study. To further study the effectiveness of each component in section 3.3, we report the performance of our model trained on the same training data but with different loss terms on both the BU3DFE and **Now** validation datasets. Results are shown in Table 4. In the following, the "backbone" refers to the ResNet50 structure trained with basic loss terms including L_{img}, L_{lmk}, L_{sp}. We refer to the model trained with $L_{id}+L_G$ as identity-aware **baseline** model; model trained with $L_{id}+L_G+L_{au-label}$ is referred to as **baseline+AU label** model; our final model, i.e., the model trained with $L_{id}+L_G+L_{au-label}+L_{au-corr}$, is referred to as **baseline+AU label+AU prior** model.

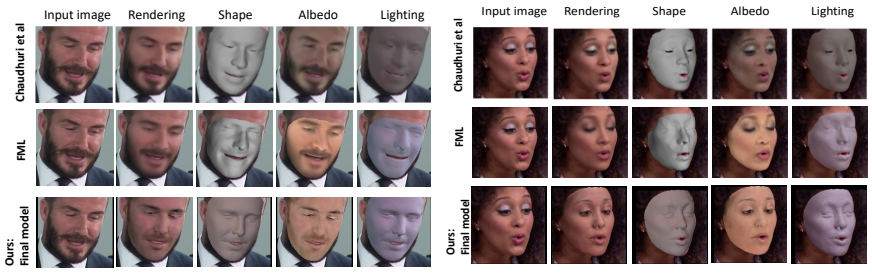

Fig. 3: Reconstruction result visualization on VoxCeleb2[13] using **ours: baseline+AU label+AU prior**, Chaudhuri et al.[11] and FML [48]. Our model can produce accurate facial expressions.

The first four rows of Table 4 reveal that, prior to adding any AU regularization terms, adding L_G provides the most significant performance improvement. This demonstrates the effectiveness of constraining locally deformed blendshapes, even before they are made AU-specific. Advancing to the fifth and sixth rows of Table 4, where the regularization terms for AU labels and AU prior relationships are added, we observe that AU-based regularization provides another significant gain in 3D accuracy over the **baseline** model for both datasets. Furthermore, even at intermediate stages, our models can sometimes outperform state-of-the-art methods. For example, on BU3DFE, our **baseline+AU label** model achieves MSE = 1.18, which is better than MSE = 1.21 achieved by [4]. Since the available ground-truth scans in **Now** validation data do not have significant facial expressions, the degree to which AU regularization improves the final model is somewhat less than for BU3DFE. Even without AU regularization, our **baseline+AU label** model outperforms nearly all of the prior methods.

Qualitative evaluation. The PCA basis used in BFM [22] has no semantic meaning, and usually contains global deformations. In contrast, our learned AU-blendshapes have the advantage that they are capable of representing subtle local facial motions closely related to AUs, and moreover they are adapted to the known AU relationship priors described in Section 3.2. This advantage is illustrated qualitatively in Fig. 3, where we compare the 3D reconstruction result on VoxCeleb2[13] images with the models of FML [48] and Chaudhur et al. [11].It is evident from the examples given in Fig. 3 that our model can produce more reliable 3D facial motion details, especially in eye-brow motion and mouth motion. In Fig. 4, we perform a more comprehensive comparison of our model trained using different loss functions with state-of-art reconstruction methods. By self-comparison, our model **(C)** and **(D)** capture 3D expressions more accurately than our baseline model **(B)**. In Fig. 5, we provide testing examples on **BU3DFE** and **Now**. For more qualitative results, please reference our supplementary materials.

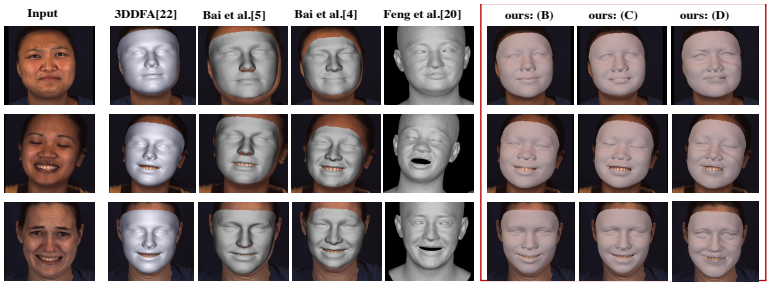

Fig. 4: Reconstruction result comparison of our model and state-of-art face reconstruction models: 3DDFA [23], Bai et al. [5], Bai et al. [4], Feng et al. [21] on BP4D [56] dataset. For our model, **(B)** is **baseline** model, **(C)** is **baseline+AU label** model and **(D)** is **baseline+AU label + AU prior** model.

(a) BU3DFE (b) Now challenge

Fig. 5: Reconstruction visualization on BU3DFE samples and Now Challenge samples using our **baseline+AU label+AU prior** model.

Table 5: Comparison to state-of-the-art methods on BP4D (F1 score)

Method	AU1	AU2	AU4	AU6	AU7	AU10	AU12	AU14	AU15	AU17	AU23	AU24	**Avg.**
DRML[32]	36.4	[41.8]	43.0	55.0	67.0	66.3	65.8	54.1	33.2	48.0	31.7	30.0	48.3
ROI[31]	36.2	31.6	43.4	77.1	73.7	85.0	87.0	62.6	45.7	58.0	38.3	37.4	56.4
JAA-Net[45]	47.2	44.0	54.9	77.5	74.6	84.0	86.9	61.9	43.6	60.3	42.7	41.9	60.0
DSIN[15]	51.7	40.4	56.0	76.1	73.5	79.9	85.4	62.7	37.3	62.9	38.8	41.6	58.9
LP-Net[39]	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
ARL [46]	45.8	39.8	55.1	75.7	77.2	82.3	86.8	58.8	47.6	62.1	47.4	55.4	61.1
Jacob et al.[24]	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**
Ours-final	**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

4.2 Evaluation of 3D AU Detection Performance

Comparison to State-of-the-art Methods. To prove that the learned 3DMM is AU-interpretable with the proposed AU regularization terms, we perform AU detection directly using our 3D model and compare to state-of-the-art image-frame-based methods accordingly, including EAC-Net [32], ROI [31], JAA-Net [45], DSIN [15] and LP-Net [39], ARL [46], Jacob et al. [24]. We stress that our model performs AU detection as an outcome of 3D face reconstruction, which is a totally different mechanism from these SOTA methods, and yet our proposed method still achieves comparable AU detection performance for within-dataset evaluation on BP4D. During testing, given a set of predicted expression parameters $\boldsymbol{\alpha}$, we apply Eq. 7 to obtain activation probabilities and regard AU_i as active if $p_i > 0.5$. Results are shown in Table. 5.

According to Table 5, the overall average F1-score of our 3D model achieves comparable performance compared to most of the frame-based methods. More

Table 6: Ablation Study on AU Detection

Models with different losses (`ReseNet50`)	**Avg. (F1)**
backbone	30.3
backbone+L_{id}	30.4
backbone+L_G	51.0
backbone+ $L_{id} + L_G$ (**baseline**)	56.0
backbone+$L_{id} + L_G + L_{au-label}$	60.0
Ours-final: *backbone*+$L_{id} + L_G + L_{au-label} + L_{au-corr}$	**60.8**

importantly, our model performs significantly better performance on three challenging AUs: AU1, AU2 and AU23 and very close performance with [24] on AU7 and AU17. For AUs with distinguishable vertex deformations (like AU1 and AU2) that can be more easily identified from the overall geometry, our model can achieve good performance. For AUs with highly correlated blendshapes, i.e. the vertex deformations are overlapped and similar, our model is more susceptible to misclassification than image-based methods.

Ablation Study. To better understand whether the effect of each loss component on AU detection performance is consistent with 3D reconstruction, we employ the same model nomenclature as introduced in the previous section and report the average AU detection F1-score. During the training of the **baseline** model, no AU labels are used; only generic AU-blendshapes are used as weak constraints for AU modeling. Comparing the **baseline+AU label** F1-score to that of **baseline**, it is clear that AU label regularization helps significantly. Adding the AU relationship priors helps as well, albeit to a lesser extent.

5 Conclusion

We have proposed a novel framework for learning subject-dependent AU blendshapes by directly applying AU-related regularization terms on a 3D face model. With a learned AU-specific basis, our model is able to generate accurate 3D face reconstruction, especially for subtle motions in the eye and mouth regions, and can be directly utilized for AU detection. Experimental results demonstrate that our model achieves state-of-the-art 3D reconstruction accuracy and generates comparable AU detection results through integrating AU information during the model learning. Most importantly, quantitative evaluation of the two tasks shows that incorporating AU information in 3D geometry helps recover more realistic and explainable facial motions and 3D basis provide a new perspective of detecting challenging AUs, indicating a great potential for using our model in conjunction with image-based methods in a complementary fashion to create an AU detector that combines the best of both.

Acknowledgment The work described in this paper is supported in part by the U.S. National Science Foundation award CNS 1629856.

References

1. Albiero, V., Chen, X., Yin, X., Pang, G., Hassner, T.: img2pose: Face alignment and detection via 6dof, face pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7617–7627 (2021)
2. Amberg, B., Romdhani, S., Vetter, T.: Optimal step nonrigid icp algorithms for surface registration. In: 2007 IEEE conference on computer vision and pattern recognition. pp. 1–8. IEEE (2007)
3. Ariano, L., Ferrari, C., Berretti, S., Del Bimbo, A.: Action unit detection by learning the deformation coefficients of a 3d morphable model. Sensors **21**(2), 589 (2021)
4. Bai, Z., Cui, Z., Liu, X., Tan, P.: Riggable 3d face reconstruction via in-network optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6216–6225 (2021)
5. Bai, Z., Cui, Z., Rahim, J.A., Liu, X., Tan, P.: Deep facial non-rigid multi-view stereo. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5850–5860 (2020)
6. Bayramoglu, N., Zhao, G., Pietikäinen, M.: Cs-3dlbp and geometry based person independent 3d facial action unit detection. In: 2013 International Conference on Biometrics (ICB). pp. 1–6. IEEE (2013)
7. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: Proceedings of the 26th annual conference on Computer graphics and interactive techniques. pp. 187–194 (1999)
8. Booth, J., Roussos, A., Zafeiriou, S., Ponniah, A., Dunaway, D.: A 3d morphable model learnt from 10,000 faces. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5543–5552 (2016)
9. Cao, C., Weng, Y., Zhou, S., Tong, Y., Zhou, K.: Facewarehouse: A 3d facial expression database for visual computing. IEEE Transactions on Visualization and Computer Graphics **20**(3), 413–425 (2013)
10. Chang, F.J., Tran, A.T., Hassner, T., Masi, I., Nevatia, R., Medioni, G.: Expnet: Landmark-free, deep, 3d facial expressions. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). pp. 122–129. IEEE (2018)
11. Chaudhuri, B., Vesdapunt, N., Shapiro, L., Wang, B.: Personalized face modeling for improved face reconstruction and motion retargeting. In: European Conference on Computer Vision. pp. 142–160. Springer (2020)
12. Chu, W.S., la Torre, F.D., Cohn, J.F.: Selective transfermachine for personalized facial action unit detection. In: CVPR (2013)
13. Chung, J.S., Nagrani, A., Zisserman, A.: Voxceleb2: Deep speaker recognition. arXiv preprint arXiv:1806.05622 (2018)
14. Corneanu, C., Madadi, M., Escalera, S.: Deep structure inference network for facial action unit recognition. In: ECCV (2019)
15. Corneanu, C., Madadi, M., Escalera, S.: Deep structure inference network for facial action unit recognition. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 298–313 (2018)
16. Cui, Z., Song, T., Wang, Y., Ji, Q.: Knowledge augmented deep neural networks for joint facial expression and action unit recognition. Advances in Neural Information Processing Systems **33** (2020)
17. Danelakis, A., Theoharis, T., Pratikakis, I.: 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)

18. Deng, Y., Yang, J., Xu, S., Chen, D., Jia, Y., Tong, X.: Accurate 3D face reconstruction with weakly-supervised learning: From single image to image set. In: Computer Vision and Pattern Recognition Workshops. pp. 285–295 (2019)
19. Dib, A., Thebault, C., Ahn, J., Gosselin, P., Theobalt, C., Chevallier, L.: Towards high fidelity monocular face reconstruction with rich reflectance using self-supervised learning and ray tracing. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2021)
20. Ekman, P., Friesen, W.V., Hager, J.C.: Facial action coding system. A Human Face, Salt Lake City, UT (2002)
21. Feng, Y., Feng, H., Black, M.J., Bolkart, T.: Learning an animatable detailed 3D face model from in-the-wild images. vol. 40 (2021), https://doi.org/10.1145/3450626.3459936
22. Gerig, T., Morel-Forster, A., Blumer, C., Egger, B., Luthi, M., Schönborn, S., Vetter, T.: Morphable face models-an open framework. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). pp. 75–82. IEEE (2018)
23. Guo, J., Zhu, X., Yang, Y., Yang, F., Lei, Z., Li, S.Z.: Towards fast, accurate and stable 3d dense face alignment. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020)
24. Jacob, G.M., Stenger, B.: Facial action unit detection with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7680–7689 (2021)
25. Jiao, Y., Niu, Y., Tran, T.D., Shi, G.: 2d+ 3d facial expression recognition via discriminative dynamic range enhancement and multi-scale learning. arXiv preprint arXiv:2011.08333 (2020)
26. Li, G., Zhu, X., Zeng, Y., Wang, Q., Lin, L.: Semantic relationships guided representation learning for facial action unit recognition. In: AAAI (2019)
27. Li, H., Weise, T., Pauly, M.: Example-based facial rigging. Acm transactions on graphics (tog) **29**(4), 1–6 (2010)
28. Li, H., Sun, J., Xu, Z., Chen, L.: Multimodal 2d+ 3d facial expression recognition with deep fusion convolutional neural network. IEEE Transactions on Multimedia **19**(12), 2816–2831 (2017)
29. Li, R., Bladin, K., Zhao, Y., Chinara, C., Ingraham, O., Xiang, P., Ren, X., Prasad, P., Kishore, B., Xing, J., Li, H.: Learning formation of physically-based face attributes (2020)
30. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4d scans. ACM Trans. Graph. **36**(6), 194–1 (2017)
31. Li, W., Abtahi, F., Zhu, Z.: Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1841–1850 (2017)
32. Li, W., Abtahi, F., Zhu, Z., Yin, L.: Eac-net: Deep nets with enhancing and cropping for facial action unit detection. IEEE transactions on pattern analysis and machine intelligence **40**(11), 2583–2596 (2018)
33. Li, Y., Chen, J., Zhao, Y., Ji, Q.: Data-free prior model for facial action unit recognition. IEEE Transactions on affective computing **4**(2), 127–141 (2013)
34. Liu, Z., Song, G., Cai, J., Cham, T.J., Zhang, J.: Conditional adversarial synthesis of 3d facial action units. Neurocomputing **355**, 200–208 (2019)
35. Liu, Z., Song, G., Cai, J., Cham, T.J., Zhang, J.: Conditional adversarial synthesis of 3d facial action units. Neurocomputing **355**, 200–208 (2019)

36. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision (ICCV) (December 2015)
37. Müller, C.: Spherical harmonics, vol. 17. Springer (2006)
38. Niu, X., Han, H., Yang, S., Shan, S.: Local relationship learning with person-specific shape regularization for facial action unit detection. In: CVPR (2019)
39. Niu, X., Han, H., Yang, S., Huang, Y., Shan, S.: Local relationship learning with person-specific shape regularization for facial action unit detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition workshops (2019)
40. Paysan, P., Knothe, R., Amberg, B., Romdhani, S., Vetter, T.: A 3d face model for pose and illumination invariant face recognition. In: 2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance. pp. 296–301. Ieee (2009)
41. Reale, M.J., Klinghoffer, B., Church, M., Szmurlo, H., Yin, L.: Facial action unit analysis through 3d point cloud neural networks. In: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019). pp. 1–8. IEEE (2019)
42. Sanyal, S., Bolkart, T., Feng, H., Black, M.: 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) (Jun 2019)
43. Savran, A., Sankur, B., Bilge, M.T.: Comparative evaluation of 3d vs. 2d modality for automatic detection of facial action units. Pattern recognition **45**(2), 767–782 (2012)
44. Shang, J., Shen, T., Li, S., Zhou, L., Zhen, M., Fang, T., Quan, L.: 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 16. pp. 53–70. Springer (2020)
45. Shao, Z., Liu, Z., Cai, J., Ma, L.: Deep adaptive attention for joint facial action unit detection and face alignment. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 705–720 (2018)
46. Shao, Z., Liu, Z., Cai, J., Wu, Y., Ma, L.: Facial action unit detection using attention and relation learning. IEEE transactions on affective computing (2019)
47. Song, X., Shi, T., Feng, Z., Song, M., Lin, J., Lin, C., Fan, C., Yuan, Y.: Unsupervised learning facial parameter regressor for action unit intensity estimation via differentiable renderer. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 2842–2851 (2020)
48. Tewari, A., Bernard, F., Garrido, P., Bharaj, G., Elgharib, M., Seidel, H.P., Pérez, P., Zollhofer, M., Theobalt, C.: Fml: Face model learning from videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10812–10822 (2019)
49. Tewari, A., Seidel, H.P., Elgharib, M., Theobalt, C., et al.: Learning complete 3d morphable face models from images and videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3361–3371 (2021)
50. Tulyakov, S., Vieriu, R.L., Sangineto, E., Sebe, N.: Facecept3d: real time 3d face tracking and analysis. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 28–33 (2015)
51. Wu, C.Y., Xu, Q., Neumann, U.: Synergy between 3dmm and 3d landmarks for accurate 3d facial geometry. arXiv preprint arXiv:2110.09772 (2021)

52. Yan, Y., Lu, K., Xue, J., Gao, P., Lyu, J.: Feafa: A well-annotated dataset for facial expression analysis and 3d facial animation. In: 2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). pp. 96–101. IEEE (2019)
53. Yang, H., Zhu, H., Wang, Y., Huang, M., Shen, Q., Yang, R., Cao, X.: 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. pp. 601–610 (2020)
54. Yang, H., Wang, T., Yin, L.: Adaptive multimodal fusion for facial action units recognition. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 2982–2990 (2020)
55. Yin, L., Wei, X., Sun, Y., Wang, J., Rosato, M.J.: A 3d facial expression database for facial behavior research. In: 7th international conference on automatic face and gesture recognition (FGR06). pp. 211–216. IEEE (2006)
56. Zhang, X., Yin, L., Cohn, J.F., Canavan, S., Reale, M., Horowitz, A., Liu, P., Girard, J.M.: Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. Image and Vision Computing **32**(10), 692–706 (2014)
57. Zhang, Y., Dong, W., Hu, B., Ji, Q.: Classifier learning with prior probabilities for facial action unit recognition. In: CVPR (2018)
58. Zhang, Y., Dong, W., Hu, B., Ji, Q.: Weakly-supervised deep convolutional neural network learning for facial action unit intensity estimation. In: CVPR (2018)
59. Zhou, Q.Y., Park, J., Koltun, V.: Open3d: A modern library for 3d data processing. arXiv preprint arXiv:1801.09847 (2018)
60. Zhu, K., Du, Z., Li, W., Huang, D., Wang, Y., Chen, L.: Discriminative attention-based convolutional neural network for 3d facial expression recognition. In: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019). pp. 1–8. IEEE (2019)
61. Zhu, X., Liu, X., Lei, Z., Li, S.Z.: Face alignment in full pose range: A 3d total solution. IEEE transactions on pattern analysis and machine intelligence (2017)
62. Zhu, X., Yang, F., Huang, D., Yu, C., Wang, H., Guo, J., Lei, Z., Li, S.Z.: 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. pp. 343–358. Springer (2020)