# Towards Accurate Facial Motion Retargeting with Identity-Consistent and Expression-Exclusive Constraints

**Langyuan Mo[1,2], Haokun Li[1], Chaoyang Zou[3], Yubing Zhang[3], Ming Yang[3], Yihong Yang[4], Mingkui Tan[1,5*]**

[1] School of Software Engineering, South China University of Technology, [2] Pazhou Laboratory, [3] CVTE Research, [4] MINIEYE, [5] Key Laboratory of Big Data and Intelligent Robot (South China University of Technology), Ministry of Education
{selymo, selihaokun}@mail.scut.edu.cn, {zhangyubing, yangming}@cvte.com,
zouchaoyang2021@gmail.com, vivid2vivian@hotmail.com, mingkuitan@scut.edu.cn

## Abstract

We address the problem of facial motion retargeting that aims to transfer facial motion from a 2D face image to 3D characters. Existing methods often formulate this problem as a 3D face reconstruction problem, which estimates the face attributes such as face identity and expression from face images. However, due to the lack of ground-truth labels for both identity and expression, most 3D-face reconstruction-based methods fail to capture the facial identity and expression accurately. As a result, these methods may not achieve promising performance. To address this, we propose an identity-consistent constraint to learn accurate identities by encouraging consistent identity prediction across multiple frames. Based on a more accurate identity, we are able to obtain a more accurate facial expression. Moreover, we further propose an expression-exclusive constraint to improve performance by avoiding the co-occurrence of contradictory expression units (e.g., "brow lower" vs. "brow raise"). Extensive experiments on facial motion retargeting and 3D face reconstruction tasks demonstrate the superiority of the proposed method over existing methods. Our code and supplementary materials are available at https://github.com/deepmo24/CPEM.

## 1 Introduction

Facial motion retargeting, which aims to transfer facial motion (*i.e.*, facial expression and head pose) from monocular RGB images to 3D targets, is a key technology for many applications, such as virtual actors in movies and games, and avatars in virtual reality and teleconferencing (Zollhöfer et al. 2018; Egger et al. 2020; Shi et al. 2020; Peihao et al. 2020). Different from the face reenactment task (Ha et al. 2020; Yao et al. 2021) that transfers facial motion from 2D images to 2D images, the task of facial motion retargeting aims to transfer facial motion from 2D images to 3D characters. However, this task is very challenging since it requires capturing accurate facial expressions from only 2D images. More critically, it is difficult to train facial expression extraction models due to the insufficiency of annotated data in real-world applications.

Most existing methods (Tuan Tran et al. 2017; Genova et al. 2018; Deng et al. 2019; Chaudhuri et al. 2020; Shang et al. 2020) attempt to learn facial expressions by solving a 3D face reconstruction problem based on the 3D Morphable Model (3DMM)(Blanz and Vetter 1999). Specifically, they predict the coefficients of the face shape (*i.e.*, face identity and expression) and combine these coefficients with the face model to reconstruct a 3D face for training. In this way, they can obtain the facial expression derived from the reconstructed 3D face. However, these methods may fail to capture facial expressions accurately for the following reasons.

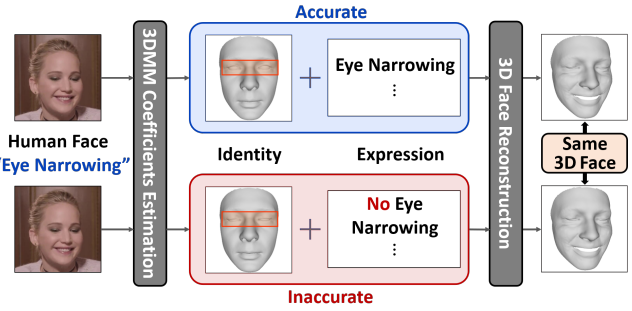


Figure 1: Limitations of existing 3D face reconstruction-based methods for facial expression estimation. Since there exists no independent supervision of identity and expression, existing methods may produce an incorrect identity (*i.e.*, small eye) with an incorrect expression (*i.e.*, no eye narrowing) in order to reconstruct a 3D face shape correctly.

First, reconstructing accurate identity is necessary for capturing accurate expressions, which, however, is very difficult due to the lack of ground-truth labels for the face identity and expression. Specifically, the 3D face mainly consists of two parts: identity and expression. Most existing 3D face reconstruction methods mainly focus on whether the reconstructed face shape is accurate or not. However, in practice, inaccurate identity and expression can also generate a satisfied face shape (see Figure 1), which can not satisfy the requirements of the retargeting task, i.e., accurate expression. Thus, besides obtaining an accurate 3D face shape, how to predict an accurate identity is also vital for the desire of accurate expression. To this end, some methods attempt to enforce identity consistency across multiple images of the same person. FML (Tewari et al. 2019) and Personalized (Chaudhuri et al. 2020) used average pooling to aggregate the identity features from multiple images. RingNet (Sanyal et al.

*Corresponding author.

2019) enforced the distance between the same identities to be smaller than that of different identities by a margin. However, their constraints are too slack to enforce the consistency of predicted identities and thus result in poor estimation performance of expressions. Therefore, how to promote consistent and accurate identity prediction is still an open question.

Second, contradictory facial expression units (*e.g.*, "brow lower" vs. "brow raise") exist that are not expected to appear simultaneously in one face. However, due to the lack of direct supervision on these expression units, it is hard to avoid the co-occurrence of contradictory expression units, which creates difficulty in learning accurate facial expressions. Existing methods (Chaudhuri, Vesdapunt, and Wang 2019; Chaudhuri et al. 2020) only employ an $l_1$ loss to enforce sparse expression coefficients, and thus cannot effectively avoid the co-occurrence of contradictory expression units.

To address the above issues, we propose a Consistent Parameter Estimation Model (CPEM) that incorporates two well-designed constraints for accurate expression extraction. First, considering that a good identity is an essential condition for precisely capturing expression, we propose an identity-consistent constraint. Specifically, we explicitly enforce all predicted identity coefficients of the same person across multiple frames to approximate the average predicted identity, which helps to learn a consistent and accurate identity that is robust to different expressions. Therefore, accurate identity estimation promotes accurate expression estimation. Furthermore, we propose an expression-exclusive constraint to suppress the co-occurrence of contradictory expression units. Specifically, we first define a contradictory expression pair set from the expression attributes with prior knowledge. Then, to suppress the contradictory expression units, we explicitly deactivate the undesired expression unit to zero according to a carefully designed suppressing rule. This further facilitates a more accurate expression estimation.

We summarize the contributions of this paper as follows.

- To better learn accurate expressions, we turn our problem into a problem of learning accurate identity. To this end, we propose a simple yet effective identity-consistent constraint for learning consistent and accurate identities across multiple frames from the same person, which significantly promotes expression estimation performance.
- To avoid the co-occurrence of contradictory expression units, we propose an expression-exclusive constraint to suppress contradictory expression units from appearing together to achieve better expression estimation.
- Extensive experiments on facial motion retargeting and 3D face reconstruction benchmarks show our method brings significant improvement in reconstructing accurate expressions compared with state-of-the-art methods.

## 2 Related Work

**3D Face Reconstruction.** Recent 3D face reconstruction methods mostly use deep neural networks to estimate the 3DMM coefficients. Some methods (Richardson, Sela, and Kimmel 2016; Dou, Shah, and Kakadiaris 2017; Guo et al. 2018) use real 3D scans to generate synthetic rendered images as supervision. Others (Feng et al. 2018; Yi et al. 2019; Cao et al. 2019; Lang et al. 2019; Guo et al. 2020) propose different network architectures and use 3DMM coefficients or 3D face labels from a fitted 3D face dataset (Zhu et al. 2016). To overcome the limitation of the lacking realistic 3D face data, Genova et al. (2018) trained a regression network using only unlabeled photographs with a differentiable renderer. To exploit complementary information from different images, Deng et al. (2019) performed multi-image face reconstruction by shape aggregation. RingNet (Sanyal et al. 2019) enforced the shape consistency of multiple images by requiring the distance between matched pairs to be smaller than unmatched pairs by a margin. Tewari et al. (2019, 2021) proposed a multi-frame video-based framework to learn a face model from data and then perform 3D face reconstruction. Moreover, multi-view methods (Wu et al. 2019; Shang et al. 2020) exploit multi-view consistency to improve the 3D face reconstruction performance, especially under large pose situations. However, all these methods mainly focus on the final reconstructed 3D face but ignore the accurate estimation of expressions. In contrast, we focus on capturing accurate facial expressions for effective facial motion retargeting.

**Face Tracking and Retargeting.** Early optimization-based methods (Weise et al. 2011; Bouaziz, Wang, and Pauly 2013; Li et al. 2013) usually optimize the tracking parameters of the face model and adaptively correct the expression blendshapes using depth scans. Afterwards, some methods (Cao et al. 2013a; Cao, Hou, and Zhou 2014; Cao et al. 2015) learned 3D facial shape regressors and optimized the parameters of the expression blendshapes only with 2D video frames, which required either calibration for each user or a specifically designed expression model. Recently, learning-based methods (Chaudhuri, Vesdapunt, and Wang 2019; Chaudhuri et al. 2020) trained a deep neural network to estimate the parameters of expression blendshapes and conduct facial motion retargeting. Chaudhuri et al. (Chaudhuri, Vesdapunt, and Wang 2019) proposed a multi-task framework to jointly learn to predict the face bounding box and the 3DMM parameters. To recover the facial expression details, Chaudhuri et al. (2020) proposed to jointly learn a personalized face blendshape model and estimate the tracking parameters in a multi-frame framework. To better learn the blendshape model, they aggregated the identity features from multiple images by average pooling. However, these methods may not learn a consistent and accurate identity and neglect the co-occurrence of the contradictory expression units, which results in inaccurate expression estimation.

## 3 Preliminaries

In this paper, we aim to learn accurate facial expressions from input images for facial motion retargeting. To achieve this, we resort to 3D face reconstruction which estimates face identity, expression and so on from images. For convenience, we introduce the 3D face model, illumination model and camera model used for 3D face reconstruction in the following.

**3D Face Model.** We use the linear 3DMM representation model as our 3D face model in the paper, in which the face shape $\mathbf{S} \in \mathbb{R}^{V\times 3}$and the face texture $\mathbf{T} \in \mathbb{R}^{V\times 3}$ of the 3D

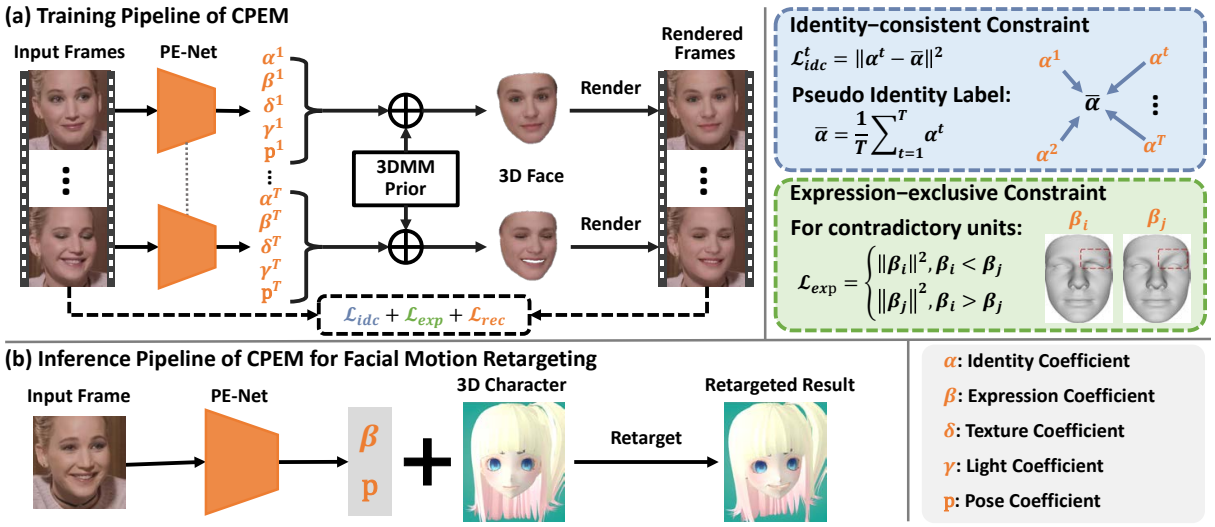


Figure 2: An overview of our framework. We use the Parameter Estimation Network (PE-Net) to estimate the 3DMM coefficients from multiple video frames of a person in the training. To promote accurate expression estimation, we use novel identity-consistent constraint $\mathcal{L}_{idc}$ and expression-exclusive constraint $\mathcal{L}_{exp}$ as well as other 3D face reconstrcution losses $\mathcal{L}_{rec}$ to train our model. During the inference, we only take the expression and head pose coefficients for facial motion retargeting.

face are represented by two affine models:

$$\mathbf{S} = \overline{\mathbf{S}} + \boldsymbol{\alpha}\mathbf{B}_{id}^{\top} + \boldsymbol{\beta}\mathbf{D}_{exp}^{\top}, \quad \mathbf{T} = \overline{\mathbf{T}} + \boldsymbol{\delta}\mathbf{B}_{tex}^{\top}, \qquad (1)$$

where $\overline{\mathbf{S}}$ and $\overline{\mathbf{T}}$ are the average face shape and texture respectively, and $V$ is the number of vertices of the 3D face. $\mathbf{B}_{id}$ and $\mathbf{B}_{tex}$ denote the PCA bases of identity and texture respectively, while $\mathbf{D}_{exp}$ denotes the expression model. $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ are the corresponding coefficient vectors of the face model. We use the popular 2009 Basel Face Model (BFM) (Paysan et al. 2009) for $\overline{\mathbf{S}}, \overline{\mathbf{T}}, \mathbf{B}_{id}$ and $\mathbf{B}_{tex}$ and exclude the ear and neck region. Moreover, we use the delta blendshapes (*i.e.*, displacements from the rest pose) taken from the FaceWarehouse database (Cao et al. 2013b) as the expression model $\mathbf{D}_{exp}$, since these blendshapes have a clear semantic meaning for describing facial expressions. The model contains 46 expression units as described in Facial Action Coding System (Friesen and Ekman 1978), in which each expression unit has a value from 0 to 1, representing the expression intensity from weak to strong. Note that the delta blendshapes have been transferred to the topology of the BFM model using deformation transfer (Sumner et al. 2004). As a result, we have $\boldsymbol{\alpha} \in \mathbb{R}^{80}$, $\boldsymbol{\beta} \in \mathbb{R}^{46}$ and $\boldsymbol{\delta} \in \mathbb{R}^{80}$.

**Illumination Model.** To model the scene illumination, we assume a Lambertian surface for the 3D face and approximate the scene illumination with Spherical Harmonics (SH) (Ramamoorthi and Hanrahan 2001). Specifically, using the face texture and surface normal as input, the scene light can be calculated via the SH basis functions with the corresponding SH coefficient $\boldsymbol{\gamma}$. We choose the first three bands of SH basis functions following (Deng et al. 2019), such that $\boldsymbol{\gamma} \in \mathbb{R}^9$.

**Camera Model.** To project the reconstructed 3D face into the 2D image plane, we use the perspective camera model with an empirically-selected focal length. Therefore, the pose $\mathbf{p}$ of the 3D face is represented by an Euler rotation $\mathbf{r} \in SO(3)$ and translation $\boldsymbol{\tau} \in \mathbb{R}^3$.

Last, we concatenate all the required 3DMM coefficients into a single vector $\mathbf{x} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\gamma}, \mathbf{p})$ that is used to reconstruct the 3D face and render it back to the image plane with the differentiable renderer (Lassner and Zollhofer 2021). In particular, the expression coefficient $\boldsymbol{\beta}$ and head pose coefficient $\mathbf{p}$ are used for facial motion retargeting.

## 4 Consistent Parameter Estimation Model

We focus on solving the problem that captures the facial expressions accurately from only 2D images for facial motion retargeting. To this end, we use the 3D face reconstruction framework which inputs a face image and outputs the reconstructed 3D face shape with a combination of face identity and expression. However, existing methods often fail to estimate the face identity and expression accurately due to the lack of ground-truth labels for both of them. In this paper, considering that reconstructing an accurate identity is necessary for estimating accurate expression, we propose an identity-consistent constraint to explicitly enforce a consistent identity coefficient prediction across multiple frames. In this way, we are able to learn a consistent and accurate identity for the same person, which will in turn improve accurate expression estimation. To extract more accurate expressions, we further propose an expression-exclusive constraint to regularize our model to avoid predicting contradictory expression coefficients simultaneously. An overview of our framework is shown in Figure 2.

Formally, given $T$ frames of the same person as inputs, we first use the parameter estimation network to estimate $T$ groups of 3DMM coefficients. Then, we combine the predicted coefficients with the 3DMM prior model to reconstruct the 3D faces that are then rendered back to the image plane with the differentiable renderer. Finally, we train our parame-

ter estimation network via the proposed identity-consistent and expression-exclusive constraints as well as several losses for 3D face reconstruction in a self-supervised manner. The loss function is given as

$$\mathcal{L} = \sum_{t=1}^{T} \mathcal{L}_{rec}^{t} + \lambda_{idc}\mathcal{L}_{idc}^{t} + \lambda_{exp}\mathcal{L}_{exp}^{t}, \tag{2}$$

where $\mathcal{L}_{rec}^{t}$, $\mathcal{L}_{idc}^{t}$, and $\mathcal{L}_{exp}^{t}$ are the self-supervised losses for 3D face reconstruction and the proposed identity-consistent and expression-exclusive losses, respectively. $\lambda_{idc}$ and $\lambda_{exp}$ are the hyper-parameters of the corresponding losses. For simplicity, we omit the superscript $t$ in the following sections.

During the inference, our model takes one image as input each time and predicts the 3DMM coefficients, where the expression coefficient $\boldsymbol{\beta}$ and head pose coefficient $\mathbf{p}$ are used for facial motion retargeting to any 3D target with expression blendshapes consistent with ours, as shown in Figure 2.

### 4.1 Identity-consistent Constraint

Due to the lack of ground-truth labels for the face identity and expression, it is hard to train a 3D face reconstruction model that can estimate the face identity and expression accurately. To learn more accurate facial expressions for the retargeting purpose, we turn our problem to learning accurate identity. Motivated by the fact that the identity of the same person across multiple frames should be consistent, we seek to enforce consistent prediction of identity coefficients across multiple frames of a person. However, it is non-trivial for the model to learn consistent identity representation for the same person without ground-truth identity labels. To address this, we propose to use the average of the predicted identity coefficients for $T$ frames as the pseudo identity label $\overline{\boldsymbol{\alpha}}$ to supervise the identity coefficient output of each frame. [1]

Specifically, we enforce all output identity coefficients to approximate the pseudo identity label during training. Formally, our identity-consistent constraint in the $t$-th frame is defined as the mean square error between the predicted identity coefficient $\boldsymbol{\alpha}^t$ and the pseudo identity label $\overline{\boldsymbol{\alpha}}$:

$$\mathcal{L}_{idc}^{t} = \left\| \boldsymbol{\alpha}^t - \overline{\boldsymbol{\alpha}} \right\|^2. \tag{3}$$

Note that the identity label is constantly updated during training but as a fixed label to supervise the identity outputs of the parameter estimation network in the backpropagation of each batch. In this way, our model gradually learns consistent and accurate identity estimation in the self-supervised training, which in turn improves accurate expression estimation.

Intuitively, the average identity coefficient from multiple frames of the same person is more accurate than the specific identity coefficient of each frame, since it reduces the variation caused by different expressions from each frame. Therefore, through explicitly approximating the predicted identities to the average predicted identity, our model fully exploits the identity information from multiple frames to estimate consistent and accurate identity from different images of a person even though the expressions are various.

[1] Some discussion on different choices of the pseudo identity label is put in the supplementary materials.

### 4.2 Expression-exclusive Constraint

Our expression model is made up of 46 expression units (*i.e.*, blendshapes), in which several expression units are contradictory such as "brow lower" and "brow raise". However, in the training process, there is no direct supervision of the expression coefficients to avoid the co-occurrence of these contradictory expression units, which brings difficulties in learning accurate expressions. To address this, we propose an expression-exclusive constraint to guide the model to suppress those expression units that should not appear. Specifically, we first define a contradictory expression pair set as $\mathcal{O}$ from the expression blendshape model with prior knowledge, where each pair $(i, j) \in \mathcal{O}$ is the subscript of $\boldsymbol{\beta}$. Here, $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_j$ are the expression coefficients of a contradictory expression unit pair [2]. Assuming that for two contradictory expression units, the unit with the larger value is dominant and should be kept, we instantiate the expression-exclusive constraint as the expression-exclusive loss as follows:

$$\mathcal{L}_{exp} = \sum_{(i,j)\in\mathcal{O}} \left\| \mathbb{1}\left\{\boldsymbol{\beta}_i > \boldsymbol{\beta}_j\right\} \cdot \boldsymbol{\beta}_j \right\|^2 + \left\| \mathbb{1}\left\{\boldsymbol{\beta}_j > \boldsymbol{\beta}_i\right\} \cdot \boldsymbol{\beta}_i \right\|^2, \tag{4}$$

where $\mathbb{1}\{\cdot\}$ is the indicator function, in which $\mathbb{1}\{a\} = 1$ if $a$ is true and $\mathbb{1}\{a\} = 0$ if $a$ is false.

In this way, our CPEM learns to deactivate the expression units that should not appear and thus promote a more accurate prediction of the expression coefficients. Essentially, we take a "winner-take-all" strategy for the contradictory expression units, which is reasonable if the model has good expression prediction performance. In practice, we can easily meet this requirement by adding the expression-exclusive loss after training the model for enough iterations.

### 4.3 Multi-frame Loss for 3D Face Reconstruction

We train the proposed CPEM in a self-supervised manner without using 3D supervision. In addition to the two constraints we proposed in the above sections, we also include several loss functions for 3D face reconstruction. They are defined as follows:

$$\begin{aligned} \mathcal{L}_{rec} =& \lambda_{pho}\mathcal{L}_{pho} + \lambda_{per}\mathcal{L}_{per} + \lambda_{lm}\mathcal{L}_{lm} \\ &+ \lambda_{reg}\mathcal{L}_{reg} + \lambda_{sp}\mathcal{L}_{sp}, \end{aligned} \tag{5}$$

where $\mathcal{L}_{pho}, \mathcal{L}_{per}, \mathcal{L}_{lm}, (\mathcal{L}_{reg}, \mathcal{L}_{sp})$ are photometric loss, perception loss, landmark loss and regularization loss, respectively. Following previous approaches (Deng et al. 2019; Shang et al. 2020), the hyper-parameters $\lambda_{pho}$, $\lambda_{per}$, $\lambda_{lm}$, $\lambda_{reg}$ and $\lambda_{sp}$ are set to 1.9, 0.2, 0.1, 1e-4 and 0.1, respectively, in all our experiments.

**Photometric Loss.** We use the $l_{2,1}$ loss (Thies et al. 2016) to compute the photometric discrepancy between the $t$-th input frame $\mathbf{I}$ and the rendered frame $\hat{\mathbf{I}}$. The loss is given by

$$\mathcal{L}_{pho} = \frac{\sum_{i\in\mathcal{M}} A_i \cdot \left\| \mathbf{I}_i - \hat{\mathbf{I}}_i \right\|_2}{\sum_{i\in\mathcal{M}} A_i}, \tag{6}$$

[2] The detailed information of the defined contradictory expression pair set is described in the supplementary materials.

where $i$ denotes the pixel index, and $\mathcal{M}$ is the reprojected face region generated by the differentiable renderer. $A$ is a face mask with a value of 1 in the face skin region, and a value of 0 elsewhere obtained by an existing face segmentation method (Yu et al. 2018), which can reduce the error brought by occlusion such as eyeglasses.

**Perception Loss.** To produce more realistic face shapes, we also use a pretrained face recognition network (Cao et al. 2018) to employ a perception loss during training (Deng et al. 2019; Genova et al. 2018). Specifically, we extract the deep features of the input image $\mathbf{I}$ and the rendered image $\hat{\mathbf{I}}$, and compute the cosine distance to measure the similarity between the two features. The loss is defined as

$$\mathcal{L}_{per} = 1 - \frac{\langle f(\mathbf{I}), f(\hat{\mathbf{I}}) \rangle}{\|f(\mathbf{I})\| \cdot \|f(\hat{\mathbf{I}})\|}, \tag{7}$$

where $f(\cdot)$ denotes the deep features extracted from the face recognition network and $\langle \cdot, \cdot \rangle$ denotes vector inner product.

**Landmark Loss.** The landmark loss measures the difference between ground-truth 2D facial landmarks $\mathbf{Q}$ and the corresponding landmarks $\hat{\mathbf{Q}}$ in the reconstructed 3D face, where $\hat{\mathbf{Q}}$ are already projected into the image plane by the learned camera model. The loss is defined as

$$\mathcal{L}_{lm} = \sum_{i=1}^{n} \omega_i \cdot \left\| \mathbf{Q}_i - \hat{\mathbf{Q}}_i \right\|^2, \tag{8}$$

where $n$ denotes the number of landmarks, $i$ denotes the landmark index and $\omega_i$ is the landmark weight, which we set to 1 for the face contour and 10 for the inner face region to reduce the impact of contour landmarks. To acquire more accurate facial landmarks as ground truth, we use a combination of landmarks detected from a 3D face alignment method (Bulat and Tzimiropoulos 2017) and a 2D face alignment method (King 2009) (see the supplementary material).

**Regularization Loss.** To prevent face shape and texture degeneration, we add a commonly-used regularization loss on the estimated 3DMM coefficients to enforce a prior distribution towards the mean face. The loss is given as

$$\mathcal{L}_{reg} = \lambda_\alpha \|\boldsymbol{\alpha}\|^2 + \lambda_\delta \|\boldsymbol{\delta}\|^2, \tag{9}$$

Following Deng et al. (2019), the hyper-parameters are empirically set to $\lambda_\alpha = 1.0$ and $\lambda_\delta = 1.7e-3$. Besides, we also impose a $l_1$ loss $\mathcal{L}_{sp}$ to enforce sparse expression coefficients following Chaudhuri et al. (2020).

**Differences with FML and RingNet.** FML (Tewari et al. 2019) used the average pooling (Avgpool) strategy to aggregate the identities across multiple frames, while RingNet (Sanyal et al. 2019) enforced the distance between the same identities to be smaller than that of different identities by a margin. However, these two strategies often failed to predict consistent identity for the same person, as shown in Figure 5. In contrast, our proposed identity-consistent constraint explicitly enforces consistent identity prediction during training, which is able to learn more consistent and accurate identity, thereby facilitating more accurate expression estimation. Furthermore, we propose an expression-exclusive constraint to extract more accurate expressions.

# 5 Experiments

## 5.1 Experimental Settings

**Implementation Details.** We implement our method based on PyTorch (Paszke et al. 2019) and use the differentiable renderer from Pytorch3d (Lassner and Zollhofer 2021). We use an Adam optimizer (Kingma and Ba 2015) with a learning rate of 1e-4. We train our model for 300K iterations with a batch size of 8 and an input size of 224×224, and only use the expression-exclusive loss in the last 100K iterations. We use ResNet50 (He et al. 2016) as the backbone of the parameter estimation network. [3] We change the output dimension of the last fully connected layer to output the 3DMM coefficients and use the sigmoid function on the expression branch. By default, we set $T = 4$, $\lambda_{idc} = 1000$, and $\lambda_{exp} = 10$.

**Datasets.** We train our model on three publicly available datasets: VoxCeleb2 (Joon Son et al. 2018), 300W-LP (Zhu et al. 2016) and FEAFA (Yan et al. 2019). VoxCeleb2 has more than 140K videos of about 6K identities in the training set, and about 5K videos in the testing set. 300W-LP contains synthesized large-pose face images from 300W (Sagonas et al. 2013). We consider the set of images of the same person with different poses as a video. FEAFA is a facial expression dataset containing 123 facial videos of 122 identities with about 100K frames. To measure the expression accuracy, we collect an expression test set from FEAFA. More details about this test set are put in the supplementary.

**Baselines.** We compare our method with MS-SFN (Chaudhuri et al. 2019), Personalized (Chaudhuri et al. 2020) and RingNet (Sanyal et al. 2019) on the facial motion retargeting task. We reimplement the first two methods since the code and models are unavailable. For a fair comparison, we use the same training datasets and keep the same settings as ours in these methods. We also compare our method with several state-of-the-art 3D face reconstruction methods on 3D face reconstruction and 2D face alignment tasks. Specifically, we compare with the following baseline methods: 3DDFA (Zhu et al. 2016), PRNet (Feng et al. 2018), RingNet (Sanyal et al. 2019), Deng et al. (2019), MS-SFN (Chaudhuri et al. 2019) and 3DDFA-V2 (Guo et al. 2020).

## 5.2 Qualitative Results

**Facial Motion Retargeting.** We evaluate the effectiveness of our method on FEAFA (Yan et al. 2019) test set as shown in Figure 3. Benefited from the proposed identity-consistent and expression-exclusive constraints, our method achieves the most accurate facial motion retargeting results compared with baseline methods. Specifically, the facial motion retargeted by MS-SFN (Chaudhuri et al. 2019) yields both inaccurate head pose and facial expressions. Through enforcing the identity consistency with different strategies, RingNet (Sanyal et al. 2019) and Personalized (Chaudhuri et al. 2020) estimate a little more accurate expressions. However, they still predict inconsistent identity thus the retargeting results are worse than ours (*e.g.*, smaller lip corner pull, smaller nose wrinkle, and smaller mouth open).

[3] More results with a light-weight MobileNet-V2 (Sandler et al. 2018) backbone are put in the supplementary materials.

| Method | Eye Close | Brow Lower | Brow Raise | Mouth Open | Lip Suck | Lip R/D | Kiss | Nose Wrinkle | Lip Corner Pull | Lip Corner Stretch | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|
| MS-SFN | 0.204 | 0.585 | 0.358 | 0.430 | 0.396 | 0.381 | 0.687 | 0.909 | 0.509 | 0.502 | 0.496 |
| RingNet | 0.276 | **0.214** | 0.303 | 0.234 | 0.592 | 0.260 | 0.441 | 0.767 | 0.423 | 0.390 | 0.390 |
| Personalized | 0.148 | 0.552 | 0.269 | 0.159 | 0.596 | 0.265 | 0.445 | 0.805 | 0.413 | 0.323 | 0.397 |
| Ours (w/o $\mathcal{L}_{idc} + \mathcal{L}_{exp}$) | 0.336 | 0.273 | 0.492 | 0.325 | 0.829 | 0.334 | 0.573 | 0.896 | 0.467 | 0.430 | 0.495 |
| Ours (w/ $\mathcal{L}_{idc}$) | 0.137 | 0.371 | **0.077** | 0.158 | **0.385** | **0.231** | 0.236 | 0.484 | 0.319 | 0.293 | 0.269 |
| Ours (w/ $\mathcal{L}_{idc} + \mathcal{L}_{exp}$) | **0.127** | 0.341 | 0.108 | **0.104** | 0.445 | 0.241 | **0.201** | **0.379** | **0.227** | **0.271** | **0.244** |

Table 1: Comparisons of facial motion retargeting accuracy (measured by Mean Absolute Error) on FEAFA test set with different methods. The lower error means the method performs better for capturing expressions.

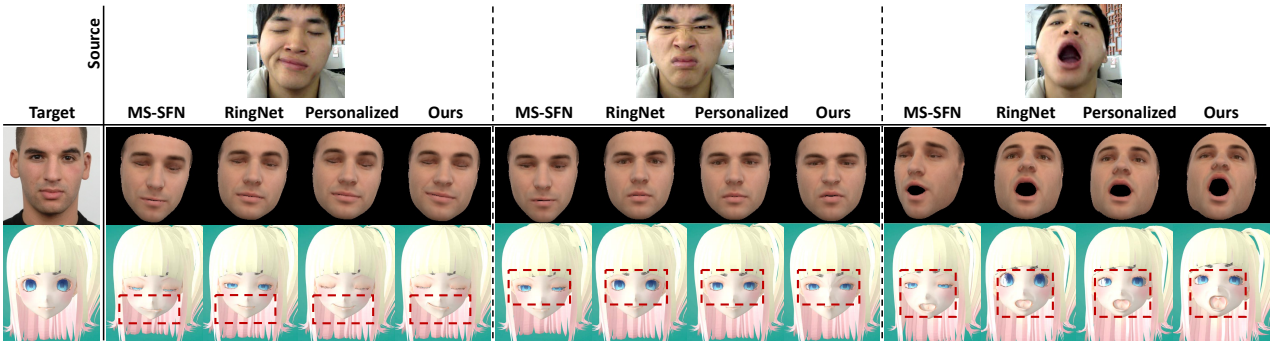


Figure 3: Facial motion retargeting comparison with different methods. We show the source person with different expressions in the first row. The second and third rows present the retargeting results(*i.e.*, transferring facial expression and head pose) to a target human and a target 3D character, respectively. We highlight some retargeting details using the box in red dash line.

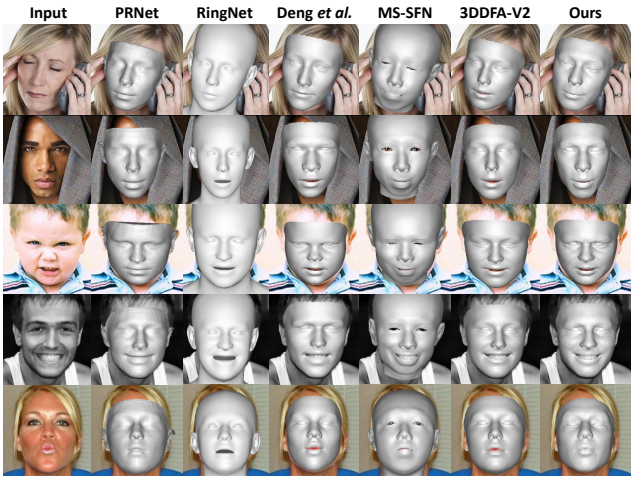


Figure 4: Visual comparisons on 3D face reconstruction with state-of-the-art methods.

**3D Face Reconstruction.** We show visual comparison results of reconstructing 3D face geometry on AFLW2000-3D dataset (Zhu et al. 2016) in Figure 4. Specifically, we evaluate the proposed CPEM on several images in the conditions of occlusion, extreme light and large expressions. The results show that our method reconstructs comparable 3D face geometry compared with other methods in most cases. Moreover, when dealing with large expressions (*e.g.*, the first and last rows in Figure 4), our CPEM even obtains better 3D face reconstruction performance than other baseline methods.

### 5.3 Quantitative Results

**Facial Motion Retargeting.** We quantitatively evaluate the facial motion retargeting performance regarding the Mean Absolute Error (MAE) of expression units on the collected FEAFA expression test set. As shown in Table 1, our method achieves the lowest MAE and outperforms existing methods by a large margin. This means our method estimates more accurate expressions by learning consistent and accurate identities, as well as avoiding the co-occurrence of contradictory expression units. In contrast, MS-SFN (Chaudhuri et al. 2019), RingNet (Sanyal et al. 2019) and Personalized (Chaudhuri et al. 2020) obtain much higher MAEs because they predict inconsistent and inaccurate identities, which hampers accurate expression estimation.

**2D Face Alignment.** We evaluate the face tracking performance of our method on the AFLW2000-3D dataset (Zhu et al. 2016) using the Normalized Mean Error (NME) as the evaluation metric. Specifically, the NME metric is defined as the average Euclidean distance between the 68 predicted and ground truth 2D landmarks with the bounding box size as the normalization factor. The results in Table 2 show that our method achieves the best results in the situation of small and medium yaw angles, while performs slightly worse than 3DDFA-v2 (Guo et al. 2020) at large yaw angles. However, 3DDFA-v2 needs additional 3D supervised information, while our model is only trained with 2D supervision.

| Method | AFLW2000-3D(68pts) | | | |
|---|---|---|---|---|
| | [0,30] | [30,60] | [60,90] | Mean |
| 3DDFA | 3.10 | 4.17 | 5.49 | 4.25 |
| PRNet | 2.76 | 3.56 | 4.67 | 3.66 |
| Deng *et al.* | 3.36 | 5.12 | 7.88 | 5.45 |
| MS-SFN | 3.43 | 4.20 | 6.27 | 4.63 |
| 3DDFA-v2 | 2.89 | 3.57 | **4.49** | 3.65 |
| CPEM(ours) | **2.68** | **3.48** | 4.75 | **3.64** |

Table 2: Comparisons of NME(%) for 68 landmarks on AFLW2000-3D dataset (with 3 groups based on yaw angles).

| Method | MICC Florence | | | FW |
|---|---|---|---|---|
| | Cooperative | Indoor | Outdoor | |
| RingNet | 2.09±0.48 | 2.13±0.46 | 2.10±0.47 | 2.47±0.32 |
| Deng *et al.* | 2.98±1.00 | 2.22±0.55 | 2.06±0.48 | 2.15±0.32 |
| MS-SFN | 2.37±0.59 | 2.52±0.66 | 2.99±0.90 | 2.20±0.45 |
| 3DDFA-v2 | **1.94±0.52** | 2.10±0.50 | **1.98±0.49** | 2.26±0.43 |
| CPEM(ours) | 2.08±0.59 | **2.09±0.54** | 2.02±0.54 | **2.03±0.33** |

Table 3: Geometric reconstruction error(mm) on MICC Florence and Facewarehouse(FW) dataset. Bold number for the best result and underline number for the second-best result.

**3D Face Reconstruction.** We quantitatively evaluate the geometric reconstruction capability of our method on the MICC Florence dataset (Bagdanov et al. 2011) and the FaceWarehouse dataset (Cao et al. 2013b). Following Genova et al. (2018), we calculate the point-to-plane root mean square error with the average shape for each video in different scenarios, and average the results. Moreover, we use 9 identities in FaceWarehouse dataset to calculate the point-to-point root mean square error following Deng et al. (2019). The results in Table 3 show that our method achieves promising results on both datasets. Specifically, our method achieves the best results in both indoor scenarios of the MICC Florence and FaceWarehouse datasets due to its superiority in capturing accurate expressions.

## 5.4 Ablation Study

To evaluate the effectiveness of the proposed two constraints (*i.e.*, identity-consistent and expression-exclusive constraints) and the sensitivity of hyper-parameters, we conduct a series of ablation studies on FEAFA expression test set.

**Effectiveness of Proposed Constraints.** To investigate the effectiveness of the proposed identity-consistent constraint $\mathcal{L}_{idc}$ and expression-exclusive constraint $\mathcal{L}_{exp}$, we compare the quantitative facial motion retargeting results of the models optimized with and without these two losses. In Table 1, both proposed constraints contribute to promising performance. Specifically, the identity-consistent constraint creates a large improvement in facial motion retargeting performance, which demonstrates the effectiveness of identity-consistent constraint. Combining these two constraints achieves slight improvement on average compared to using $\mathcal{L}_{idc}$ alone. Therefore, $\mathcal{L}_{exp}$ further helps to capture accurate expressions.

We further evaluate the prediction stability of identity coefficients on the test set of Voxceleb2. Specifically, we calculate

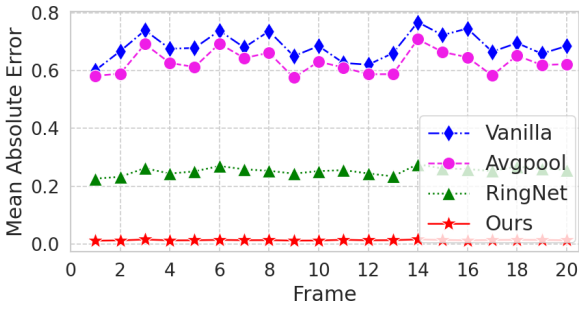


Figure 5: Comparisons of the identity consistency in terms of MAE of the identity coefficients between each subsequent frame and the first frame on the test set of Voxceleb2.

| Parameter | $\lambda_{idc}$ | | | | $\lambda_{exp}$ | | |
|---|---|---|---|---|---|---|---|
| | 1 | 100 | 1000 | 10000 | 1 | 10 | 100 |
| Avg | 0.491 | 0.286 | **0.244** | 0.279 | 0.260 | **0.244** | 0.255 |

Table 4: Effect of the hyper-parameters $\lambda_{idc}$ and $\lambda_{exp}$ on the facial motion retargeting in terms of MAE.

MAE of identity coefficients between each subsequent frame and the first frame on a video, and average the results of 40 videos. From Figure 5, the average pooling strategy (Chaudhuri et al. 2020) and RingNet (Sanyal et al. 2019) fail to predict stable identities for the same person. Moreover, without the identity-consistent constraint (Vanilla), a high MAE shows that the model usually predicts inconsistent identities for the same person. In contrast, we obtain a very low MAE with the proposed identity-consistent constraint, resulting in more accurate expression estimation (see Table 1).

**Sensitivity of Hyper-parameters.** In this section, We evaluate the sensitivity of two hyper-parameters, *i.e.*, $\lambda_{idc}$ and $\lambda_{exp}$. As shown in Table 4, our model achieves the best performance when setting $\lambda_{idc}$=1000 and $\lambda_{exp}$=10. When the hyper-parameter of $\lambda_{idc}$ is as small as 1, the facial motion retargeting accuracy is greatly impaired compared to the best result (0.244 $vs.$ 0.491). As $\lambda_{idc}$ increases from 1 to 1000, the performance is stably improved, which further demonstrates the effectiveness of the proposed identity-consistent constraint. For $\lambda_{exp}$, the performance is less sensitive to the hyper-parameter of $\lambda_{exp}$ than that of $\lambda_{idc}$.

# 6 Conclusion

In this paper, we have proposed an effective approach to accurately capture facial expressions to improve facial motion retargeting performance. To accurately capture facial expressions, we proposed a simple yet effective identity-consistent constraint to explicitly enforce a consistent identity prediction. Moreover, we proposed an expression-exclusive constraint to avoid the co-occurrence of contradictory expression units, which further improves the expression estimation performance. Extensive experiments on facial motion retargeting and 3D face reconstruction benchmarks demonstrate the superiority of the proposed method in estimating accurate expressions over previous state-of-the-art approaches.

## Acknowledgments

This work was partially supported by National Natural Science Foundation of China (NSFC) 62072190, Ministry of Science and Technology Foundation Project 2020AAA0106900, Key-Area Research and Development Program of Guangdong Province (2019B010155001), Program for Guangdong Introducing Innovative and Enterpreneurial Teams 2017ZT07X183.

## References

Bagdanov, A. D.; Del Bimbo, A.; and Masi, I. 2011. The florence 2d/3d hybrid face dataset. In *Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding*, 79–80.

Blanz, V.; and Vetter, T. 1999. A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, 187–194.

Bouaziz, S.; Wang, Y.; and Pauly, M. 2013. Online modeling for realtime facial animation. *ACM Transactions on Graphics*, 32(4): 1–10.

Bulat, A.; and Tzimiropoulos, G. 2017. How far are we from solving the 2D & 3D Face Alignment problem? (and a dataset of 230,000 3D facial landmarks). In *Proceedings of the IEEE International Conference on Computer Vision*.

Cao, C.; Bradley, D.; Zhou, K.; and Beeler, T. 2015. Real-time high-fidelity facial performance capture. *ACM Transactions on Graphics*, 34(4): 1–9.

Cao, C.; Hou, Q.; and Zhou, K. 2014. Displaced dynamic expression regression for real-time facial tracking and animation. *ACM Transactions on Graphics*, 33(4): 1–10.

Cao, C.; Weng, Y.; Lin, S.; and Zhou, K. 2013a. 3D shape regression for real-time facial animation. *ACM Transactions on Graphics*, 32(4): 1–10.

Cao, C.; Weng, Y.; Zhou, S.; Tong, Y.; and Zhou, K. 2013b. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3): 413–425.

Cao, J.; Mo, L.; Zhang, Y.; Jia, K.; Shen, C.; and Tan, M. 2019. Multi-marginal wasserstein gan. *Advances in Neural Information Processing Systems*, 32: 1776–1786.

Cao, Q.; Shen, L.; Xie, W.; Parkhi, O. M.; and Zisserman, A. 2018. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition*, 67–74. IEEE.

Chaudhuri, B.; Vesdapunt, N.; Shapiro, L.; and Wang, B. 2020. Personalized Face Modeling for Improved Face Reconstruction and Motion Retargeting. In *Proceedings of the European Conference on Computer Vision*.

Chaudhuri, B.; Vesdapunt, N.; and Wang, B. 2019. Joint face detection and facial motion retargeting for multiple faces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9719–9728.

Deng, Y.; Yang, J.; Xu, S.; Chen, D.; Jia, Y.; and Tong, X. 2019. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 0–0.

Dou, P.; Shah, S. K.; and Kakadiaris, I. A. 2017. End-to-end 3D face reconstruction with deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5908–5917.

Egger, B.; Smith, W. A.; Tewari, A.; Wuhrer, S.; Zollhoefer, M.; Beeler, T.; Bernard, F.; Bolkart, T.; Kortylewski, A.; Romdhani, S.; et al. 2020. 3d morphable face models—past, present, and future. *ACM Transactions on Graphics (TOG)*, 39(5): 1–38.

Feng, Y.; Wu, F.; Shao, X.; Wang, Y.; and Zhou, X. 2018. Joint 3d face reconstruction and dense alignment with position map regression network. In *Proceedings of the European Conference on Computer Vision*, 534–551.

Friesen, E.; and Ekman, P. 1978. Facial action coding system: a technique for the measurement of facial movement. *Palo Alto*, 3(2): 5.

Genova, K.; Cole, F.; Maschinot, A.; Sarna, A.; Vlasic, D.; and Freeman, W. T. 2018. Unsupervised training for 3d morphable model regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8377–8386.

Guo, J.; Zhu, X.; Yang, Y.; Yang, F.; Lei, Z.; and Li, S. Z. 2020. Towards Fast, Accurate and Stable 3D Dense Face Alignment. In *Proceedings of the European Conference on Computer Vision*.

Guo, Y.; Cai, J.; Jiang, B.; Zheng, J.; et al. 2018. Cnn-based real-time dense face reconstruction with inverse-rendered photo-realistic face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(6): 1294–1307.

Ha, S.; Kersner, M.; Kim, B.; Seo, S.; and Kim, D. 2020. Marionette: Few-shot face reenactment preserving identity of unseen targets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 10893–10900.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Joon Son, C.; Arsha, N.; and Andrew, Z. 2018. VoxCeleb2: Deep Speaker Recognition. In *Proceedings of the Interspeech 2018*, 1086–1090.

King, D. E. 2009. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10: 1755–1758.

Kingma, D. P.; and Ba, J. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Lang, Y.; Liang, W.; Wang, Y.; and Yu, L.-F. 2019. 3d face synthesis driven by personality impression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 1707–1714.

Lassner, C.; and Zollhofer, M. 2021. Pulsar: Efficient Sphere-Based Neural Rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1440–1449.

Li, H.; Yu, J.; Ye, Y.; and Bregler, C. 2013. Realtime facial animation with on-the-fly correctives. *ACM Transactions on Graphics*, 32(4): 42–42.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Paysan, P.; Knothe, R.; Amberg, B.; Romdhani, S.; and Vetter, T. 2009. A 3D face model for pose and illumination invariant face recognition. In *IEEE International Conference on Advanced Video and Signal Based Surveillance*, 296–301. IEEE.

Peihao, C.; Yang, Z.; Mingkui, T.; Hongdong, X.; Deng, H.; and Chuang, G. 2020. Generating Visually Aligned Sound from Videos. *IEEE Transactions on Image Processing*.

Ramamoorthi, R.; and Hanrahan, P. 2001. A signal-processing framework for inverse rendering. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, 117–128.

Richardson, E.; Sela, M.; and Kimmel, R. 2016. 3D face reconstruction by learning from synthetic data. In *International Conference on 3D Vision*, 460–469. IEEE.

Sagonas, C.; Tzimiropoulos, G.; Zafeiriou, S.; and Pantic, M. 2013. 300 Faces in-the-Wild Challenge: The First Facial Landmark Localization Challenge. In *IEEE International Conference on Computer Vision Workshops*, 397–403. IEEE Computer Society.

Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; and Chen, L.-C. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4510–4520.

Sanyal, S.; Bolkart, T.; Feng, H.; and Black, M. J. 2019. Learning to regress 3D face shape and expression from an image without 3D supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7763–7772.

Shang, J.; Shen, T.; Li, S.; Zhou, L.; Zhen, M.; Fang, T.; and Quan, L. 2020. Self-Supervised Monocular 3D Face Reconstruction by Occlusion-Aware Multi-view Geometry Consistency. In *Proceedings of the European Conference on Computer Vision*, 53–70.

Shi, T.; Zuo, Z.; Yuan, Y.; and Fan, C. 2020. Fast and Robust Face-to-Parameter Translation for Game Character Auto-Creation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 1733–1740.

Sumner, R. W.; and Popović, J. 2004. Deformation transfer for triangle meshes. *ACM Transactions on Graphics*, 23(3): 399–405.

Tewari, A.; Bernard, F.; Garrido, P.; Bharaj, G.; Elgharib, M.; Seidel, H.-P.; Pérez, P.; Zollhofer, M.; and Theobalt, C. 2019. Fml: Face model learning from videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 10812–10822.

Tewari, A.; Seidel, H.-P.; Elgharib, M.; Theobalt, C.; et al. 2021. Learning Complete 3D Morphable Face Models from Images and Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3361–3371.

Thies, J.; Zollhofer, M.; Stamminger, M.; Theobalt, C.; and Nießner, M. 2016. Face2face: Real-time face capture and reenactment of RGB videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2387–2395.

Tuan Tran, A.; Hassner, T.; Masi, I.; and Medioni, G. 2017. 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*, 5163–5172.

Weise, T.; Bouaziz, S.; Li, H.; and Pauly, M. 2011. Realtime performance-based facial animation. *ACM Transactions on Graphics*, 30(4): 1–10.

Wu, F.; Bao, L.; Chen, Y.; Ling, Y.; Song, Y.; Li, S.; Ngan, K. N.; and Liu, W. 2019. Mvf-net: Multi-view 3d face morphable model regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 959–968.

Yan, Y.; Lu, K.; Xue, J.; Gao, P.; and Lyu, J. 2019. Feafa: A well-annotated dataset for facial expression analysis and 3d facial animation. In *IEEE International Conference on Multimedia & Expo Workshops*, 96–101. IEEE.

Yao, G.; Yuan, Y.; Shao, T.; Li, S.; Liu, S.; Liu, Y.; Wang, M.; and Zhou, K. 2021. One-shot Face Reenactment Using Appearance Adaptive Normalization. volume 35, 3172–3180.

Yi, H.; Li, C.; Cao, Q.; Shen, X.; Li, S.; Wang, G.; and Tai, Y.-W. 2019. Mmface: A multi-metric regression network for unconstrained face reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7663–7672.

Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; and Sang, N. 2018. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European Conference on Computer Vision*, 325–341.

Zhu, X.; Lei, Z.; Liu, X.; Shi, H.; and Li, S. Z. 2016. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 146–155.

Zollhöfer, M.; Thies, J.; Garrido, P.; Bradley, D.; Beeler, T.; Pérez, P.; Stamminger, M.; Nießner, M.; and Theobalt, C. 2018. State of the art on monocular 3D face reconstruction, tracking, and applications. In *Computer Graphics Forum*, volume 37, 523–550. Wiley Online Library.