MMFace: A Multi-Metric Regression Network for Unconstrained Face Reconstruction

Hongwei Yi[1]*, Chen Li[2]†, Qiong Cao[2], Xiaoyong Shen[2], Sheng Li[3]†, Guoping Wang[3], Yu-Wing Tai[2]
[1]Shenzhen Graduate School, Peking Univ. [2]Tencent
[3]School of Electronics Engineering and Computer Science, Peking Univ.
{hongweiyi, lisheng, wgp}@pku.edu.cn {chaselli, freyaqcao, dylanshen, yuwingtai}@tencent.com

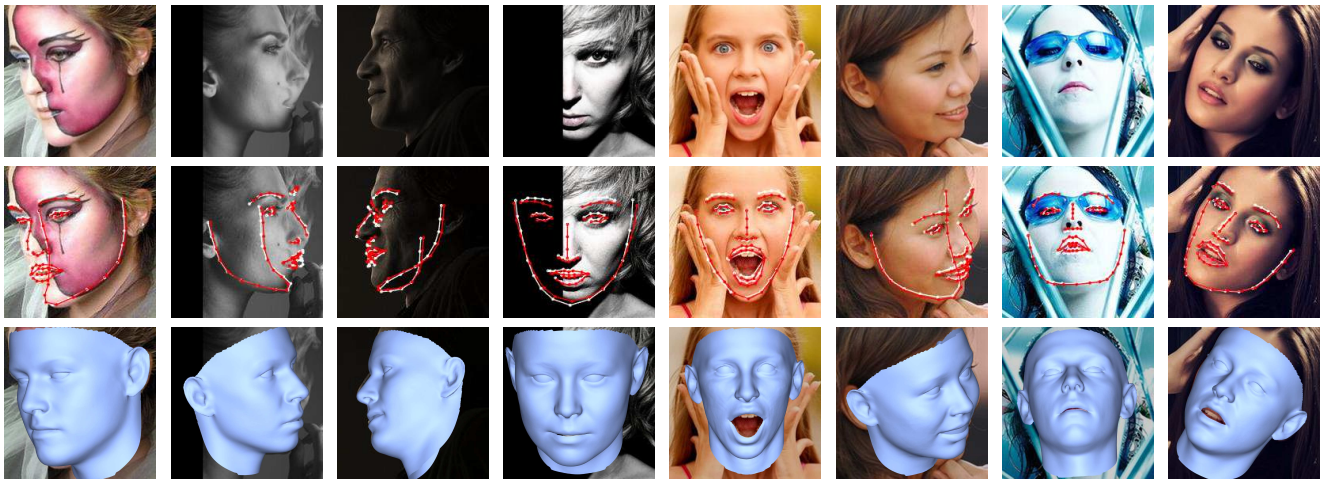

Figure 1. Face reconstruction and alignment in the wild. Our method estimates very accurate face geometries as well as 3D facial landmarks under large variations of facial expressions, poses, illumination conditions, and partial occlusions. The ground truth landmarks are denoted as red and our estimated landmarks are denoted as white.

Abstract

We propose to address the face reconstruction in the wild by using a multi-metric regression network, ***MMFace****, to align a 3D face morphable model (3DMM) to an input image. The key idea is to utilize a volumetric sub-network to estimate an intermediate geometry representation, and a parametric sub-network to regress the 3DMM parameters. Our parametric sub-network consists of identity loss, expression loss, and pose loss which greatly improves the aligned geometry details by incorporating high level loss functions directly defined in the 3DMM parametric spaces. Our high-quality reconstruction is robust under large variations of expressions, poses, illumination conditions, and even with large partial occlusions. We evaluate our method by comparing the performance with state-of-the-art approaches on latest 3D face dataset* ***LS3D-W*** *and* ***Florence****. We achieve significant improvements both quantitatively and qualitatively. Due to our high-quality reconstruction, our method can be easily extended to generate high-quality geometry sequences for video inputs.*

1. Introduction

Avatar digitization aims to produce a virtual avatar by reconstructing one's facial geometry and appearance from individual or multiple images. In the real world scenarios, face reconstruction from single unconstrained image is very challenging, especially under large head pose orientations, extreme facial expressions, severe partial occlusions, and complex illumination conditions.

To address the aforementioned limitations, we propose to train a novel neural network to align a 3D face morphable model (3DMM) [2] with an input image by regressing the corresponding model parameters. The 3DMM represents a face geometry in a multiple PCA-based linear space counting for identity and expression variations. The PCA-based linear face model has laid the foundations for the modern

[1]This work was done while Hongwei Yi was an intern at Tencent.
[2]Chen Li and Sheng Li are the joint corresponding authors.

image-based 3D face modeling [16] and it has three major advantages over other representations. First, the biometric constraints of human face structure are already embedded in the model. Second, one can easily manipulate the face identity parameter and/or face expression parameter for creating animations or achieving other dramatic VR/AR effects. Last but not least, the facial topology is preserved during the reconstruction and there are explicit point-to-point correspondences in the model space. The semantic facial landmarks can be associated with the corresponding points in the reconstructed geometry and thus the 3D facial landmarks can be easily obtained as a by-product of our method.

The parameters of 3DMM can be iteratively optimized by using optimization-based methods [27, 7, 16, 28, 29, 8, 17, 31]. These methods all rely on 2D face landmark detectors which limit their utilization since 2D landmarks detection cannot be very accurate for non-frontal faces, and in the presence of partial occlusions. Very recently, significant improvement has been achieved in 2D/3D face alignment [5, 36], which utilizes convolutional neural network (CNN) to solve the face alignment and the 3DMM parameter regression problems in one unified framework. Under the CNN framework, the parameter can be directly [12, 33, 32] (end-to-end) or iteratively [38, 19] estimated. However, due to the lacking of 3D information, these methods still cannot handle large poses and extreme expression adequately.

In contrast to the previous methods, we design a novel face reconstruction network, **MMFace**, as a multi-metric regression network which consists of two metrical regressing sub-networks, namely a volumetric network and a parametric network. The volumetric network aims to estimate a volumetric representation of the 3D facial geometry from an input 2D image and the parametric network aims to predict the corresponding 3DMM parameters upon the former volumetric representation. These two sub-networks are cascaded and supervised by two metric losses: a volumetric loss and a parametric loss. The volumetric loss restricts the volumetric network from a geometry attention perspective by representing the ground truth facial geometry in a volume space rather than its original 3D coordinate space. In this way, the head pose orientation and coarse geometry can be robustly estimated. In the second stage, the parametric loss numerically restricts the parametric network in the 3DMM attention perspective where the ground truth of its 3DMM parameter is provided. Thus, parameters that correspond to face identity and expression can be accurately estimated. By jointly restricting the entire framework using two metric losses from different attention perspectives, the two sub-networks mutually benefit each other.

We perform comprehensive experiments using the latest 3D face dataset **LS3D-W** [5] and **Florence** [1] to quantitatively evaluate the performance of our method. Our method achieves significant improvement over the state-of-the-art methods [31, 5, 36, 18, 13, 32] in both 3D face reconstruction and 3D facial landmark detection. Some of our results are shown in Fig. 1. Since our method is very robust and accurate, our method can be easily extended to video inputs by processing each frame individually. We further employ our parametric reconstruction results on 3D Animoji blendshapes to demonstrate how our method benefits other VR/AR applications.

2. Related Work

A comprehensive survey for face modeling is out of our scope in this paper. In this section, we limit the scope to the face reconstruction from single image and review the most representative recent work.

3D Face Reconstruction from single image. The PCA-based 3D morphable model of face geometry was first proposed in [2], but it was not widely used because of its high complexity. Other approaches use shading cues [22, 30, 23, 9], internet image collections [21], and/or data-driven methods [15, 14] for face reconstruction. Recently, the 3DMM has shown very impressive results by aligning the 3DMM with known 2D facial landmarks [27, 7, 16, 28, 29, 8, 17, 31]. Thanks to the recent deep learning based approaches [37, 11, 35], 2D landmarks can be robustly estimated which makes a great advancement in the face reconstruction [3, 10]. However, the results of these landmark-based alignment approaches are limited by the accuracy of the estimated 2D landmarks which are inaccurate under large head pose orientations, extreme facial expression, and/or partial occlusions.

The 3DMM parameter can also be regressed from the input image using CNN directly [12, 33] or iteratively [5, 19]. In [12], an end-to-end approach is proposed to directly estimate the facial identity and expression from 2D images based on the **VGG-Face** features [25]. Besides the single image, multiple images for the same person are used to restrict the estimated face identity during the training stage[32]. Compared with these methods, our method differs in two aspects. First, we additionally estimate the pose parameter besides the identity and expression. Second, inspired by [18], our face parameters are learned from an intermediate 3D volumetric representation to achieve a more accurate regression rather than directly regress the parameters from the input 2D images. Besides regressing 3DMM parameters, the dense point cloud of a face geometry can be reconstructed by predicting 3D position maps in the face texture coordinate space [13]. However, such geometry representation in the texture space lacks the capability for further geometry manipulation.

3D Face Alignment. State-of-the-art 3D face alignment [5, 6, 4] applies a two-step strategy: it first estimates the 2D landmarks, and then predicts the corresponding depth value

of the estimated 2D landmarks. A 2D-to-3D network takes an RGB image and 2D landmarks as inputs and outputs the corresponding 3D landmarks in [5]. However, directly extending the 2D heat maps to 3D heat maps is memory- and computation-demanding, especially when the number of landmarks increases.

JVCR[36], on the other hand, proposes to jointly regress a 3D volumetric representation and the 3D facial landmark coordinates. The network is divided into two cascaded sub-networks: a compact voxel regressor and a coordinate regressor. The compact voxel regression sub-network regresses a compact volumetric representation from coarse to fine via multiple stacked "hourglass" networks, then the coordinate regression sub-network adopts a 3D convolution to further estimate the 3D landmark coordinates from the volumetric representation. Different from our approach, **JVCR**[36] is not based on the 3DMM model, and its volumetric representation is relatively sparse compared with ours.

These methods mentioned above have the same limitation: they can only handle limited quantity of landmarks, e.g. 68 landmarks, which are too few in many 3D face applications. Although our method has a similar framework with **JVCR**[36], we propose to regress the 3DMM parameter from a complete 3D volumetric representation instead of only 68 facial landmarks. The volume of **JVCR**[36] encodes only the 3D Gaussian distribution of the specific 68 3D landmarks while our volume encodes the entire face geometry. This is significantly more challenging, and the 3DMM representation has far more applications than the 3D landmark representation as demonstrated in the experiment of our paper.

3. Multi-Metric Regression Network

The framework of our method is shown in Fig. 2. After we obtain the estimation of 3DMM parameters, the face pose are further refined using ICP (iterative closest point) as post-processing. The detailed specification of our entire network structure is described in the supplementary material.

3.1. 3D morphable model

We use the 3D morphable PCA model proposed by [2] to represent the facial geometry $\mathbf{S}$ with $n = 53215$ vertices as:

$$\mathbf{S}(\alpha_{\mathbf{id}}, \alpha_{\mathbf{exp}}) = \bar{\mathbf{S}} + \mathbf{U}_{\mathbf{id}}\alpha_{\mathbf{id}} + \mathbf{U}_{\mathbf{exp}}\alpha_{\mathbf{exp}}, \quad (1)$$

where $\bar{\mathbf{S}} \in \mathcal{R}^{3n}$ is the mean geometry, $\mathbf{U}_{\mathbf{id}} \in \mathcal{R}^{\mathbf{3n\times 199}}$ and $\mathbf{U}_{\mathbf{exp}} \in \mathcal{R}^{\mathbf{3n\times 29}}$ are the basis of face identity and expression, $\alpha_{id} \in \mathcal{R}^{199}$ and $\alpha_{exp} \in \mathcal{R}^{29}$ are the corresponding face identity parameter and face expression parameter. $\bar{\mathbf{S}}$ and $\mathbf{U}_{\mathbf{id}}$ are learned from *Basel Face Model* [26] and $\mathbf{U}_{\mathbf{exp}}$ is obtained from *FaceWarehouse* [7].

In order to further project the face geometry $\mathbf{S}$ onto the image coordinate, we utilize the weak perspective projection to simplify the projection model as:

$$\mathbf{V}(\mathbf{p}) = \mathbf{f} * \mathbf{P} * \mathbf{R} * \mathbf{S} + \mathbf{t}, \quad (2)$$

where $\mathbf{V}$ is the projected geometry in image coordinate, f is a scale factor, $\mathbf{P} = \begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\end{bmatrix}$ is the orthographic projection, $\mathbf{R}$ is the rotation matrix constructed from rotation parameter $\mathbf{r} = \{\phi, \gamma, \theta\}$, and $\mathbf{t}$ is a 2D translation. Thus, the 3D face reconstruction problem is transformed into a facial parameter regression problem where $\mathbf{p} = [\mathbf{f}, \mathbf{r}, \mathbf{t}, \alpha_{\mathbf{id}}, \alpha_{\mathbf{exp}}]^{\mathbf{T}}$ are the parameters we want to estimate.

3.2. Proposed framework

We propose to restrict 3DMM parameters from two different attention perspectives, namely a geometry perspective by using the volumetric sub-network **VMN** and a 3DMM perspective by using the parametric sub-network **PMN**.

3.2.1 Volumetric network

We stack two "hourglass network" [24] with identical structure together to encode the input to a feature space and then decode this feature representation to the volumetric domain $VMN : I \rightarrow \mathbb{V}$ [18].

Using this geometry representation, the **VMN** can be intermediately supervised by the **Volumetric loss** $E_{\mathbb{V}}$ as:

$$E_{\mathbb{V}} = \sum_{w=1}^{W}\sum_{h=1}^{H}\sum_{d=1}^{D}[\hat{\mathbb{V}}_{whd}\log(\mathbb{V}_{whd}) + (1-\hat{\mathbb{V}}_{whd})\log(1-\mathbb{V}_{whd})], \quad (3)$$

where $\{w, h, d\}$ denotes a voxel position in the volumetric space, $\mathbb{V}$ and $\hat{\mathbb{V}}$ are the corresponding sigmoid output and the ground truth of volumetric representation, respectively.

3.2.2 Parametric network

The parametric sub-network takes the output feature map of volumetric sub-network as input, and predicts the 3DMM parameter $\mathbf{p}$. We employ five 3D convolution layers to extract the 3D geometry information from the intermediate volumetric representation. After extracting the 3D features, we incorporate three independent branches with 2D fully-connected layers to regress the face identity parameter α_{id}, face expression parameter α_{exp}, and the pose parameter $\{f, \mathbf{r}, \mathbf{t}\}^{\mathbf{T}}$, respectively. Because the dimension of shape parameter α_{id} is much larger than that of expression

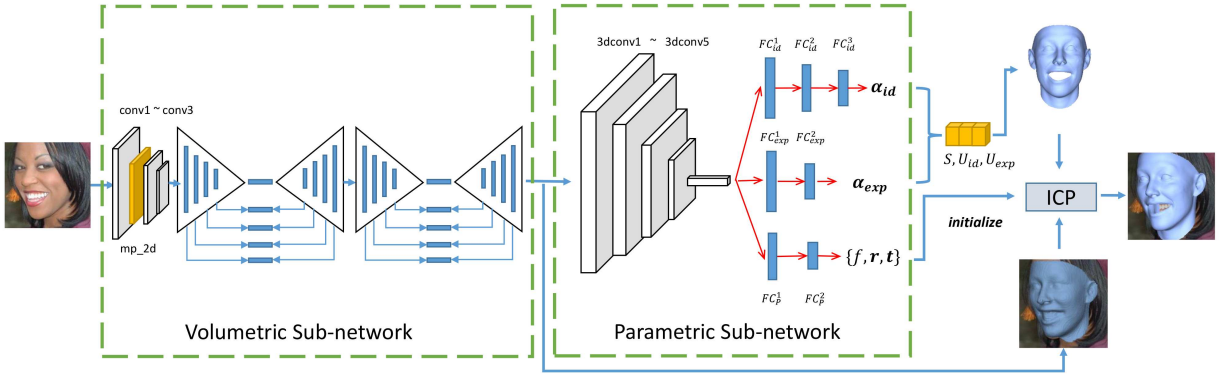

Figure 2. The framework of the proposed single image face reconstruction network **MMFace**. The 3DMM parameters are constrained from a geometry perspective by using the volumetric sub-network and a 3DMM perspective by using the parametric sub-network. The face pose is refined using ICP as post-processing.

and face pose, we use one additional fully-connected layer in its branch to better regress its value.

Our parametric sub-network can be restricted from the 3DMM attention perspective by directly incorporating the ground truth 3DMM parameters. The parametric loss consists of three loss functions, namely **Identity loss** E_{id}, **Expression loss** E_{exp} and **Pose loss** E_P.

Identity loss E_{id}**.** A trivial identity loss can be defined by using the Euclidean distance between our estimation and ground truth as::

$$E'_{id} = \|(\alpha_{id} - \hat{\alpha}_{id})\|_2^2, \tag{4}$$

where α_{id} denotes the predicted face identity parameter and $\hat{\alpha}_{id}$ is the ground truth parameter. Because the 3DMM is a PCA-based model, different dimensions in α_{id} with different singular value influence the face geometry differently. However, a uniform restriction in Eq. (4) may decrease the influence from important dimensions, especially those with large singular values.

To solve this problem, we propose to use the mean square error (MSE) between predict face geometry and ground truth directly as the constraints [12]:

$$\begin{aligned} E_{id} &= \|\mathbf{S}(\alpha_{\mathbf{id}}, \hat{\alpha}_{\mathbf{exp}}) - \mathbf{S}(\hat{\alpha}_{\mathbf{id}}, \hat{\alpha}_{\mathbf{exp}})\|_{\mathbf{2}}^{\mathbf{2}}, \\ &= \|\mathbf{U}_{\mathbf{id}}(\alpha_{\mathbf{id}} - \hat{\alpha}_{\mathbf{id}})\|_{\mathbf{2}}^{\mathbf{2}}, \end{aligned} \tag{5}$$

where $\hat{\alpha}_{exp}$ is the ground truth of face expression parameter.

Expression loss E_{exp}**.** Similar with the identity loss E_{id}, the expression loss E_{exp} can also be represented using the MSE as:

$$E_{exp} = \|\mathbf{U}_{\mathbf{exp}}(\alpha_{\mathbf{exp}} - \hat{\alpha}_{\mathbf{exp}})\|_{\mathbf{2}}^{\mathbf{2}}. \tag{6}$$

Pose loss E_P**.** The MSE loss used in identity loss E_{id} and expression loss E_{exp} can also be used to restrict the pose estimation. However, since pose has limited number of freedom, we only consider the MSE loss on its 68 facial landmarks $f_i \in \mathcal{F}$ to simplify the computation as:

$$E_P = \frac{1}{|\mathcal{F}|} \sum_{f_i \in \mathcal{F}} \|f * P * R * S_{f_i}(\hat{\alpha}_{id}, \hat{\alpha}_{exp}) + t - \hat{f} * P * \hat{R} * S_{f_i}(\hat{\alpha}_{id}, \hat{\alpha}_{exp}) - \hat{t}\|_2^2, \tag{7}$$

where $\hat{f}$, $\hat{R}$ and $\hat{t}$ are the ground truth of pose parameter.

One alternative way to restrict the parameter subnetwork could be using a single MSE loss with entire facial parameter $\mathbf{p}$ instead of three independent losses as:

$$E'_{\mathbf{p}} = \|f * P * R * S(\alpha_{id}, \alpha_{exp}) + t - \hat{f} * P * \hat{R} * S(\hat{\alpha}_{id}, \hat{\alpha}_{exp}) - \hat{t}\|_2^2. \tag{8}$$

However, in practice, the facial identity parameter α_{id}, expression parameter α_{exp}, and pose parameter $\{f, \mathbf{r}, \mathbf{t}\}^{\mathbf{T}}$ within the same loss term will affect each other and make it very difficult to converge to a good solution.

3.2.3 Final objective function

The final loss function E for training our face reconstruction network **MMFace** is define as:

$$E = \lambda_{\mathbb{V}} E_{\mathbb{V}} + E_{id} + \lambda_{exp} E_{exp} + \lambda_P E_P, \tag{9}$$

where $\lambda_{\mathbb{V}}$, λ_{exp} and λ_P balance the weights for these constraints.

Because the input of our **MMFace** network is a cropped facial region and scaled to an uniform 256×256 resolution, we can assume the scale and translation are roughly accurate with scale factor $f \simeq 0.001$ (aligning the size of mean

shape $\bar{\mathbf{S}}$ with pixel unit) and $t \simeq \{128, 128\}$. Although some methods eliminated the estimation of face pose from their frameworks [12, 18, 13], we found an accurate pose estimation is still very important. Thus, we further incorporate an ICP post-processing to align the predicted facial geometry $\mathbf{S}(\alpha_{\mathbf{id}}, \alpha_{\mathbf{exp}})$ with the predicted volume representation $\mathbb{V}$ for refining the pose estimation. Considering ICP is known to be susceptible to local minima and its performance critically relies on the quality of the initialization [34], the pose parameter predicated by our parametric sub-network is shown to be a very reasonable initialization to ensure the ICP stage converges effectively and stably.

Compared with other relevant approaches [38, 18, 12, 33, 13, 32], our network architecture benefits from the supervised training in two ways. First, both the intermediate volumetric representation and the final 3DMM parameter prediction are directly restricted by the corresponding metric losses from different attention perspectives. Second, the constraints for predicted parameters further affect the two "hourglass network" in the volumetric regression network via the backward propagation. Through such cascaded network, our **MMFace** not only predicts more accurate 3DMM parameters, but also gets a better inferred volume result.

4. Implementation

In this section, we describe the datasets and training scheme. More details are included in the supplementary material.

The 3D face dataset **LS3D-W** [5] consists of three sub-datasets: namely **300W-LP-3D** [38], **AFLW2000-3D** [38], and **300-VW-3D** [5]. **300W-LP-3D** [38] contains $122,450$ face images with 68 synthesized 3D facial landmark annotations. **AFLW2000-3D** [38] is a dataset for 3D face alignment in the wild and the images show large variations in pose, expression, illumination and occlusion etc. **300-VW-3D** [5] dataset contains 114 facial videos and $218,595$ frames in total. **Florence** [1] consists of 53 subjects and the ground truth of their 3D face geometry is scanned by a structured-light system.

Following the common protocol [38, 5, 36, 13], we use the entire **300W-LP-3D** [38] as the training set and directly evaluate our model on **AFLW2000-3D** [38] and **Florence** [1]. Because the image condition in **300-VW-3D** [5] is different from **300W-LP-3D** [38], we fine-tune our network on the training set of **300-VW-3D** [5] to handle such variations of data distribution.

To train our framework, the face regions are cropped according to the ground truth 3D facial landmarks and then scaled to 256×256 as the input to our network. The ground truth of pose parameter are modified accordingly. We first train the two sub-networks, **VMN** and **PMN**, independently. The input to **PMN** is the ground truth volume $\hat{\mathbb{V}}$ and the weights are set as $\{\lambda_{exp}, \lambda_P\} = \{5.0, 5.0\}$. Both **VMN** and **PMN** are trained using the Adam solver with initial learning rate $1.0e^{-4}$ and the learning rate is reduced to $1.0e^{-5}$ after 40 epochs for another 20 epochs. After training each sub-network, we concatenate them and fine-tune the entire network for 10 more epochs with $\{\lambda_{\mathbb{V}}, \lambda_{exp}, \lambda_P\} = \{1.0e^6, 5.0, 5.0\}$.

Table 1. The quantitative evaluation on **AFLW2000-3D** [38]. Different face orientation along Y-axis are averaged separately for $NME_{2D}^{68}\%$.

	NME_{2D}^{68}					NME_{3D}
	0-30	30-60	60-90	Mean	Std.	Mean
3DFAN [5]	2.77	**3.48**	4.61	3.62	0.86	-
JVCR [36]	2.94	3.46	4.53	3.64	**0.65**	-
Face2Face [31]	3.22	8.79	19.7	10.5	8.4	9.95
DisFace [32]	4.90	12.16	45.0	20.7	9.82	11.6
PRN [13]	2.75	3.51	4.61	3.63	0.87	4.40
3DMMITW [3]	3.09	9.21	17.20	9.83	7.07	8.33
ExpNet [10]	4.01	5.46	6.23	5.23	1.13	7.39
MMFace-PMN	5.05	6.23	7.05	6.11	1.00	8.29
MMFace-ICP-64	2.98	3.83	4.89	3.90	0.91	4.23
MMFace-ICP-128	2.61	3.65	4.43	3.56	0.83	3.78
MMFace-ICP-192	**2.50**	3.63	**4.25**	**3.46**	0.78	**3.66**
E2FAR-GT [12]	2.65	2.79	2.83	2.76	**0.1**	3.12
MMFace-GT	**1.33**	**1.64**	**1.83**	**1.61**	0.25	**2.01**

5. Experiments and Results

We compare the performance of our proposed method **MMFace** with the state-of-the-art approaches, including three face reconstruction approaches **Face2Face** [31], **DisFace** [32],**PRN** [13], as well as two 3D face alignment approaches **3DFAN** [5] and **JVCR** [36].

5.1. Evaluation

We use two measurements, NME_{2D}^{68} and NME_{3D} to quantitatively evaluate the performance from both alignment and reconstruction perspectives:

$$NME_{2D}^{68} = \frac{1}{|\mathcal{F}|\mathbb{F}} \sum_{f_i \in \mathcal{F}} \| f * P * R * \mathbf{S}_{\mathbf{f_i}}(\alpha_{\mathbf{id}}, \alpha_{\mathbf{exp}}) + \mathbf{t} - \hat{f} * P * \hat{R} * \mathbf{S}_{\mathbf{f_i}}(\hat{\alpha}_{\mathbf{id}}, \hat{\alpha}_{\mathbf{exp}}) - \hat{\mathbf{t}}) \|_{\mathbf{2}}^{\mathbf{2}}, \quad (10)$$

$$NME_{3D} = \frac{1}{|\mathbf{S}|\mathbb{S}} \| f * R * \mathbf{S}(\alpha_{\mathbf{id}}, \alpha_{\mathbf{exp}}) + \mathbf{t} - \hat{f} * \hat{R} * \mathbf{S}(\hat{\alpha}_{\mathbf{id}}, \hat{\alpha}_{\mathbf{exp}}) - \hat{\mathbf{t}}) \|_{\mathbf{2}}^{\mathbf{2}} \quad (11)$$

where $\mathbb{F}$ and $\mathbb{S}$ are the the diagonal size of the face region in image space and 3D coordinate space, respectively. NME_{2D}^{68} evaluates the normalized 2D facial landmarks prediction error and the NME_{3D} evaluates the normalized 3D face geometry estimation accuracy. Due to the ambiguity caused by the weak perspective projection model, the reconstruction results of different methods have ambiguities along Z-axis. We employ a rigid translation along Z-axis to align each result within the ground truth range.

Table 1 lists the quantitative evaluation on **AFLW2000-3D** [38]. Images with different face orientations are ana-

lyzed separately for the metric NME_{2D}^{68}. The optimization-based approach, **Face2Face** [31], shows unsatisfactory results in both reconstruction and alignment measurements. This is due to the fact that the 2D landmark detectors, *e.g.* [20], which **Face2Face** [31] heavily relies on, usually estimate inappropriate results or even fail to localize a face region under large face orientations. **DisFace** [32] aims to reconstruct a more distinguishable face identity parameter but optimizes the face pose and expression by fitting the pre-detected 2D facial landmarks. So it suffers from the same drawbacks with optimization-based approaches, such as **Face2Face** [31], that it cannot handle large pose orientations and partial occlusions. Although **3DFAN** [5], **JVCR** [36], and **PRN** [13] have different architectures, because they are all restricted only from a 3D perspective, they perform similarly in face alignment. **3DMMITW** [3] and **ExpNet** [10] aim more than reconstructing the face geometry, so less attention is paid on handling extreme face poses and lead to unsatisfied errors once the face orientation is larger than 30 degrees.

Ablation study and computational cost. In order to better analyze the performance of our method, we incorporate several variants of our method with different estimations of face pose, namely **MMFace**-PMN, **MMFace**-ICP64/128/192, and **MMFace**-GT. Specifically, **MMFace**-PMN directly uses the estimated pose from the parametric sub-network; **MMFace**-ICP-64/128/192 refine the pose estimation by doing ICP with the volume geometry from our volumetric network in different resolutions; **MMFace**-GT replaces the estimated face pose by the corresponding ground truth. Not surprisingly, **MMFace**-GT gives the best results. The mean error of NME_{2D}^{68} and NME_{3D} of **MMFace**-GT demonstrates our parametric sub-network achieves an accurate estimation of face identity parameter and expression parameter. As we have discussed, the face pose is difficult to learn and **MMFace**-PMN performs worse than some other methods. In practice, the face pose can be refined via ICP with a volumetric geometry representation in a higher resolution and the results of **MMFace**-ICP-128/192 significantly improve over **MMFace**-PMN. We additionally compare with **E2FAR**-GT [12] which only estimates face identity and expression. It performs consistently in varying face poses and achieves the lowest standard deviation. Although our results with known ground truth pose, denoted as **MMFace**-GT, the overall metrics, NME_{2D}^{68} and NME_{3D}, are improved about 30%, thanks to more accurate estimation of face identity and expression by our multi-metric regression network.

Comparing with other methods, **MMFace** achieves significant improvements in both 3D face alignment and reconstruction. Even for **MMFace**-ICP-64, it generates better reconstruction results than **PRN** [13]. Through guiding the proposed multi-metric regression network by a volumetric loss and a parametric loss from different attention perspectives, we obtain high quality results robust to large variation of facial expressions, poses, illumination conditions, and even large partial occlusions.

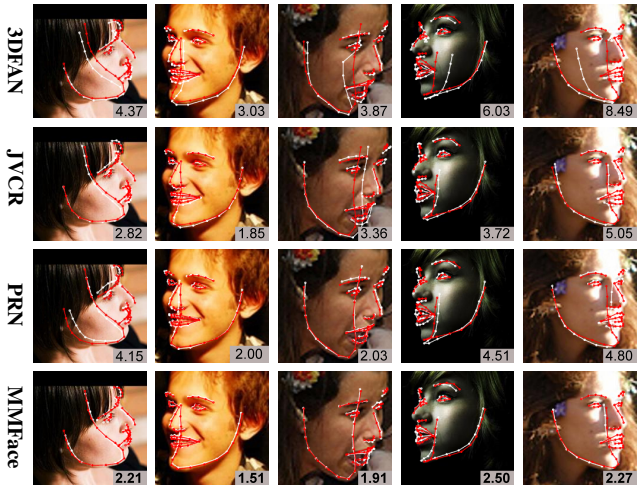

Figure 3. Comparison on 3D facial landmark detection with **3DFAN** [5], **JVCR** [36] and **PRN** [13] on **AFLW2000-3D** [38]. The NME_{2D}^{68} is listed in the lower right corner. The ground truth landmarks are denoted as red and the estimated landmarks are denoted as white.

Our forward network takes 20ms on NVIDIA TitanV GPU and ICP128 takes 15ms on Intel i7-6700 with 3.40GHz. We include additional analysis on 3D face alignment and face reconstruction as follows. The subsequent results presented in our paper are generated using **MMFace**-ICP-128 and we denote it as **MMFace** for simplification.

Comparison on 3D face alignment. One benefit of using the 3DMM rather than other geometry representations, e.g. normal map, is that the semantic facial landmarks can be associated with the corresponding points in reconstructed geometry. Thus, besides the reconstructed face geometry, the 3D facial alignment result is a by-product of our method. As a supplementary to the quantitative evaluation in Table. 1, we demonstrate some results from **AFLW2000-3D** [38] in Fig. 3. Since the images from **AFLW2000-3D** [38] are captured in the wild and show large variations in pose and appearance, it is the most challenging 3D face alignment dataset so far. Because our multi-metric regression network is restricted by both the volumetric loss and parametric loss, it outperforms the state-of-the-art 3D face alignment approaches **JVCR** [36], **3DFAN** [5], and **PRN** [13] which only consider training loss from one perspective, especially for images whose face orientation along Y-axis ranges from 60° to 90°.

Comparison on 3D face reconstruction. In addition to the quantitative evaluation in Table. 1, we further evaluate our face reconstruction results on **Florence** [1] as shown in Fig. 4. Because **Florence** [1] only provides the ground truth

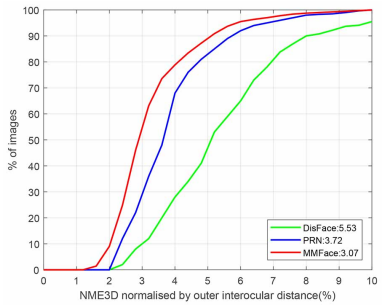

Figure 4. Evaluation for 3D reconstruction accuracy on **Florence** [1]. The mean NME% of each method is listed in the legend.

of face geometry but no camera geometry calibrations, we employ the ICP to align our results to the ground truth as done in [13]. Although our improvement in the estimation of face pose is eliminated due to this alignment, our method still outperforms others and achieves 17.5% improvement over **PRN** [13] in the NME_{3D} measurement.

Figure 6 presents a detailed comparison with **Face2Face** [31], **DisFace** [32], and **PRN** [13]. Our results are generally better in the estimation of face pose and expression. **Face2Face** [31] generally works well for frontal faces which are stable to a 2D landmark detector. Even though, it may estimate incorrect face geometry, for example the first case in Fig. 6, especially incorrect head orientations when the face is partially occluded. **DisFace** [32] can only estimate discriminative face identity parameters, but does not perform well on the estimation of face pose and expressions. So its result for the first and second cases are obviously not good. Because **PRN** [13] predicts smooth position maps rather than 3DMM parameters, it can estimate accurate face orientation but infers blurred position maps which lack facial structures in the reconstruction results. Our method, shown in the column (e) in Figure 6, provides a good approximation to the ground truth geometry. In the fourth and fifth row, we show the results on **Florence** [1]. Although our results lack of high-frequency details due to our low-dimensional 3DMM model, it still approximates better face shape and expression than **DisFace** [32] and **PRN** [13].

5.2. Video results

Besides handling the static unconstrained images across extreme facial expressions, large head poses, partial occlusions, and complex illumination, our method can be easily extended to handle video inputs.

We evaluate the results using face video dataset **300-VW-3D** [5] and show some results from one testing video in Fig. 5. Although the results of **Face2Face** [31] and **DisFace** [32] look generally ok, they generate unacceptable facial poses or expressions under large face orientations or severe occlusions due to the poor 2D facial landmarks. In frame 2217/2393, the face orientation is wrongly estimated

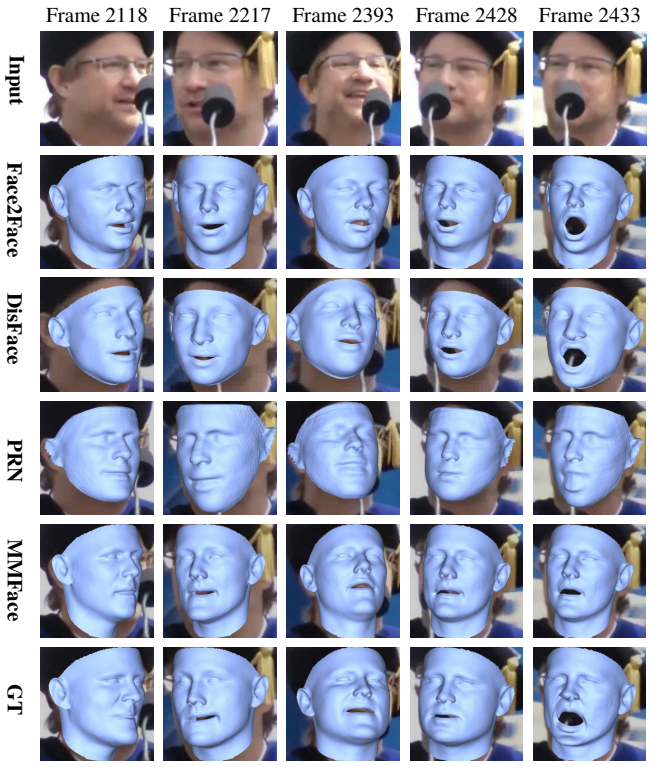

Figure 5. Comparison on video input. From the first row to the fifth row are namely, the input frames, the results of **Face2Face** [31], **DisFace** [32], **PRN** [13], our **MMFace** and the corresponding ground truth. Please refer to our supplementary material for the results on the entire video sequence.

and the expressions in frame 2118/2428/2433 are incorrect. Because **PRN** [13] directly learns to regress point clouds rather than 3DMM parameters, the internal biometric geometry constraints are not well preserved. Consequently, it cannot correctly handle the occlusions and generates unstable estimations for occluded regions caused by the voice tube.

In contrast to pervious methods, our method generates very similar results when comparing with the ground truth. In frame 2433, the voice tube completely occludes the mouth and nose region, and the ground truth synthesized by **3DFAN** [5] does not handle such occlusion correctly. In contrast, our method has not been affected by partial occlusions and produces a good approximation of the expression at occluded area. We refer readers to our supplementary material for more results.

5.3. Application

As shown in Sec. 5.1 and Sec. 5.2, our method generates accurate and robust estimation of face pose, identity and expression parameters for both static images and video streams. Such parametric reconstruction results benefit many downstream VR/AR applications, for example, the 3D Animoji. We directly employ the estimated face

(a) Input image (b) **Face2Face** (c) **DisFace** (d) **PRN** (e) **MMFace** (f) Ground truth

Figure 6. Comparison on 3D face reconstruction. (a) The input image. (b-f) The result and close-up views of **Face2Face** [31], **DisFace** [32], **PRN** [13], our **MMFace** and the ground truth. Close-up views for better visualization are aligned right to their corresponding results. The first three images are from **AFLW2000-3D** [38] and the rest images are from **Florence** [1].

Figure 7. 3D Animoji application driven by our method. The first row shows input images; the second row shows reconstructed face geometries; the rest show deformed Animoji models using our estimated face poses and expressions.

poses and expressions on Animoji blendshapes created by 3D artists, and the deformed Animoji models shown in Fig. 7 have highly consistent expressions and orientations with the input images even under large variation of face poses. Please refer to our supplementary martial for more interesting Animoji animations driven by our method.

6. Conclusion

In this paper, we present a multi-metric regression network, **MMFace**, for unconstrained 3D face reconstruction. The challenges arise from various facial expressions, head pose orientations, illumination conditions, and partial occlusions are addressed by aligning a 3D face morphable model to the input image through the proposed multi-metric regression network. This network consists of two sub-networks: a volumetric sub-network to estimate an intermediate face geometry representation in 3D volume space and a parametric sub-network to infer the corresponding 3DMM parameters. By further incorporating the volumetric loss and parametric loss, the entire framework is jointly restricted from different attention perspectives and achieves an accurate parametric 3D face reconstruction results. Our method performs significantly better than other state-of-the-art methods both quantitatively and qualitatively. Because of the robustness of our method, we also demonstrate stable and accurate results for video inputs.

Acknowledgements This project was supported by the National Key R&D Program of China (No.2017YFB1002700, No.2017YFB0203000) and NSFC of China (No.61632003, No.61661146002, No.61631001).

References

[1] Andrew D. Bagdanov, Alberto Del Bimbo, and Iacopo Masi. The florence 2d/3d hybrid face dataset. In *Proc. Joint ACM Workshop on Human Gesture and Behavior Understanding ACM Multimedia Workshop*, 2011. 2, 5, 6, 7, 8

[2] V. Blanz and T. Vetter. Face recognition based on fitting a 3d morphable model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1063–1074, 2003. 1, 2, 3

[3] James Booth, Epameinondas Antonakos, Stylianos Ploumpis, George Trigeorgis, Yannis Panagakis, Stefanos Zafeiriou, et al. 3d face morphable models "in-the-wild". In *Proc. of Computer Vision and Pattern Recognition*, 2017. 2, 5, 6

[4] Adrian Bulat and Georgios Tzimiropoulos. Two-stage convolutional part heatmap regression for the 1st 3d face alignment in the wild (3dfaw) challenge. In *Proc. of European Conference on Computer Vision*, 2016. 2

[5] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *Proc. of Internatoinal Conference on Computer Vision*, 2017. 2, 3, 5, 6, 7

[6] Adrian Bulat and Yorgos Tzimiropoulos. Convolutional aggregation of local evidence for large pose face alignment. In *Proc. of British Machine Vision Conference*, 2016. 2

[7] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2014. 2, 3

[8] Chen Cao, Hongzhi Wu, Yanlin Weng, Tianjia Shao, and Kun Zhou. Real-time facial animation with image-based dynamic avatars. *ACM Transactions on Graphics*, 35(4):126:1–126:12, 2016. 2

[9] Menglei Chai, Linjie Luo, Kalyan Sunkavalli, Nathan Carr, Sunil Hadap, and Kun Zhou. High-quality hair modeling from a single portrait photo. *ACM Transactions on Graphics*, 34(6):204:1–204:10, 2015. 2

[10] Feng-Ju Chang, Anh Tuan Tran, Tal Hassner, Iacopo Masi, Ram Nevatia, and Gerard Medioni. Expnet: Landmark-free, deep, 3d facial expressions. In *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on*, 2018. 2, 5, 6

[11] Xuanyi Dong, Shoou-I Yu, Xinshuo Weng, Shih-En Wei, Yi Yang, and Yaser Sheikh. Supervision-by-Registration: An unsupervised approach to improve the precision of facial landmark detectors. In *Proc. of Computer Vision and Pattern Recognition*, 2018. 2

[12] P. Dou, S. K. Shah, and I. A. Kakadiaris. End-to-end 3d face reconstruction with deep neural networks. In *Proc. of Computer Vision and Pattern Recognition*, 2017. 2, 4, 5, 6

[13] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *Proc. of European Conference on Computer Vision*, 2018. 2, 5, 6, 7, 8

[14] T. Hassner. Viewing real-world faces in 3d. In *Proc. of Internatoinal Conference on Computer Vision*, 2013. 2

[15] T. Hassner and R. Basri. Example based 3d reconstruction from single 2d images. In *Proc. 2006 Conference on Computer Vision and Pattern Recognition Workshop*, 2006. 2

[16] Liwen Hu, Shunsuke Saito, Lingyu Wei, Koki Nagano, Jaewoo Seo, Jens Fursund, Iman Sadeghi, Carrie Sun, Yen-Chun Chen, and Hao Li. Avatar digitization from a single image for real-time rendering. *ACM Transactions on Graphics*, 36(6):195:1–195:14, 2017. 2

[17] Patrik Huber, Guosheng Hu, Rafael Tena, Pouria Mortazavian, P Koppen, William J Christmas, Matthias Ratsch, and Josef Kittler. A multiresolution 3d morphable face model and fitting framework. In *Proc. of the International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2016. 2

[18] Aaron S. Jackson, Adrian Bulat, Vasileios Argyriou, and Georgios Tzimiropoulos. Large pose 3d face reconstruction from a single image via direct volumetric cnn regression. In *Proc. of Internatoinal Conference on Computer Vision*, 2017. 2, 3, 5

[19] Amin Jourabloo and Xiaoming Liu. Large-pose face alignment via cnn-based dense 3d model fitting. In *Proc. of Computer Vision and Pattern Recognition*, 2016. 2

[20] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proc. of Computer Vision and Pattern Recognition*, 2014. 6

[21] Ira Kemelmacher-Shlizerman. Internet-based morphable model. In *Proc. of Internatoinal Conference on Computer Vision*, 2013. 2

[22] Ira Kemelmacher-Shlizerman and Ronen Basri. 3d face reconstruction from a single image using a single reference face shape. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(2):394–405, 2011. 2

[23] Chen Li, Kun Zhou, and Stephen Lin. Intrinsic face image decomposition with human face priors. In *Proc. of European Conference on Computer Vision*, 2014. 2

[24] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *Proc. of European Conference on Computer Vision*, 2016. 3

[25] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *BMVC*, 2015. 2

[26] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *Proc. of the IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2009. 3

[27] Sami Romdhani and Thomas Vetter. Estimating 3d shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In *Proc. of Computer Vision and Pattern Recognition*, 2005. 2

[28] Joseph Roth, Yiying Tong, and Xiaoming Liu. Unconstrained 3d face reconstruction. In *Proc. of Computer Vision and Pattern Recognition*, 2015. 2

[29] J. Roth, Y. Tong, and X. Liu. Adaptive 3d face reconstruction from unconstrained photo collections. 2016. 2

[30] Supasorn Suwajanakorn, Ira Kemelmacher-Shlizerman, and Steven M. Seitz. Total moving face reconstruction. In *Proc. of European Conference on Computer Vision*, 2014. 2

[31] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proc. of Computer Vision and Pattern Recognition*, 2016. 2, 5, 6, 7, 8

[32] Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gerard Medioni. Regressing robust and discriminative 3d morphable models with a very deep neural network. In *Proc. of Computer Vision and Pattern Recognition*, 2017. 2, 5, 6, 7, 8

[33] Luan Tran and Xiaoming Liu. Nonlinear 3d face morphable model. In *Proc. of Computer Vision and Pattern Recognition*, 2018. 2, 5

[34] J. Yang, H. Li, D. Campbell, and Y. Jia. Go-icp: A globally optimal solution to 3d icp point-set registration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(11):2241–2254, 2016. 5

[35] J. Yang, Q. Liu, and K. Zhang. Stacked hourglass network for robust facial landmark localisation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017. 2

[36] Hongwen Zhang, Qi Li, and Zhenan Sun. Joint voxel and coordinate regression for accurate 3d facial landmark localization. *CoRR*, abs/1801.09242, 2018. 2, 3, 5, 6

[37] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Proc. of European Conference on Computer Vision*, 2014. 2

[38] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z. Li. Face alignment across large poses: A 3d solution. In *Proc. of Computer Vision and Pattern Recognition*, 2016. 2, 5, 6, 8