3D Face Shape Regression From 2D Videos with Multi-reconstruction and Mesh Retrieval

Xiaohu Shao[1,2] Jiangjing Lyu[3] Junliang Xing[4] Lijun Zhang[1]
Xiaobo Li[3] Xiangdong Zhou[1] Yu Shi[1]
[1]Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences [3]Alibaba Group
[4]Institute of Automation, Chinese Academy of Sciences
{shaoxiaohu, zhanglijun, zhouxiangdong, shiyu}@cigit.ac.cn
{jiangjing.ljj, xiaobo.lixb}@alibaba-inc.com, jlxing@nlpr.ia.ac.cn

Abstract

This paper introduces our submission to the 2^{nd} 3DFAW Challenge. To get a high-accuracy 3D dense face shape based on 2D videos or multiple images, a framework which consists of multi-reconstruction branches and a mesh retrieval module, is proposed to effectively utilize the information of all frames and the results predicted by all branches. The recent state-of-the-art methods based on single-view and multi-view are introduced to form an ensemble of independent regression networks. The candidate 3D shape of each branch is synthesized by weighted linear combination of the results on all frames to boost the depth estimation and invisible regions reconstruction. Finally, the best fitting mesh is retrieved according to the distance between the synthesized texture and the ground truth texture. Experiment results show that our approach obtains competitive results near the accuracy of "pseudo" ground truths, and achieves superior performance over most of submissions by other teams in the testing phases.

1. Introduction

Face alignment is critical to face analysis applications, such as face identification [2, 29], face tracking [25], and face synthesis [8, 6, 13]. Compared with 2D face alignment methods [24, 18, 14, 23], 3D face alignment is more robust to variation of occlusions and out-of-plane rotations, and has stronger representational power for describing face shapes [20, 27, 5, 19]. 3D face alignment and reconstruction have made rapid advances in recent years, especially after the utilization of deep convolution neural networks (CNN) for solving the problem [5, 1, 27, 7, 21]. Previous approaches can be divided into two categories according to input data type, the methods based on single-view and the methods based on multi-view. Due to space constrains, we mainly focus on the recently proposed methods related with our work from the above two categories, then discuss their advantages and drawbacks, respectively.

3DDFA [28] stacks a 2D image and projected normalized coordinate codes (PNCC) as the input of a cascaded CNN network to regress the 3D Morphable Model (3DMM) [3] parameters iteratively. Besides of 3DMM parameters, landmark heatmaps are also used as the representation for regressing 3D face shapes [4, 5]. [4] builds a two-stage convolutional part heatmap regression for 3D face alignment, and ranks first in the 1^{st} 3DFAW Challenge [10]. PRNet [7] introduces a light-weighted encoder-decoder network to solve the problems of face alignment and 3D face reconstruction together from a single 2D facial image. Tu *et al.* [19] take the sparse 2D facial landmarks as additional information to to substantially improve 3D face model from an single image. These methods using CNN have shown their remarkable progress of obtaining 3D facial shapes compared with traditional regressors. However, recovering 3D facial parameters from a single view always suffers from the lack of reliable depth information, and it is not robust enough to handle difficult scenarios caused by extreme poses, facial expressions and complex lighting conditions.

The approach of [6] deploys a displaced dynamic expression regression to reconstruct a performance-driven face shape based a single video without user-specific calibration. Jeni *et al.* [9] utilize a fast cascade regression framework, in which the facial landmarks on all frames are consistent across all poses, to enable real-time person-independent 3D registration from a 2D video. [16] proposes a method that selects the most plausible reconstructions according to their visual qualities, operates on different region of the face separately, and merges them into a single 3D face. MVF-Net [22] learns features from three-view face images by a shared

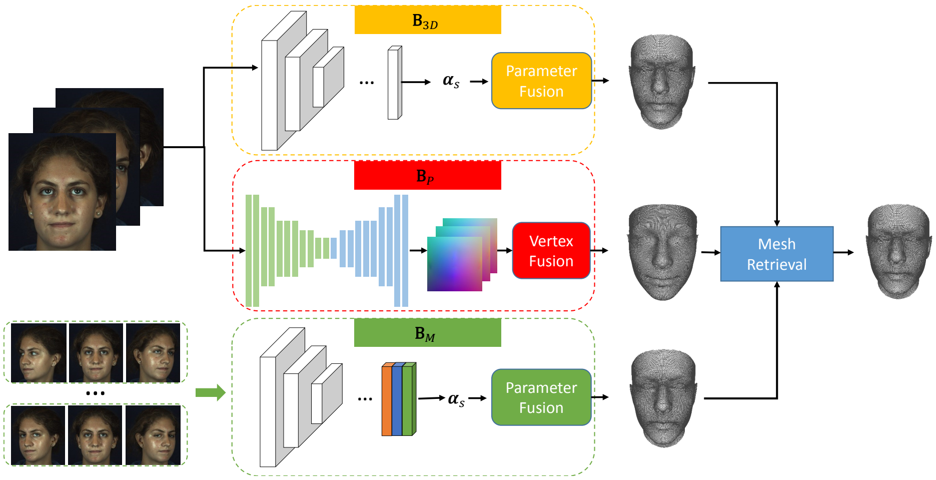

Figure 1. The overview of our proposed framework.

weight CNN, and then concatenates the features by minimizing the photometric reprojection error between different views to regress the 3DMM parameters for the face. The work [26] proposes a deep neural network that predicts the intermediate representation of the deep appearance model (DAM) from a single image and its self-supervised domain adaptation method, thus enabling facial reconstruction from a monocular video. The methods based on multi-view have the advantage of utilizing multi-view geometric constraints and thus are more robust to the wild face images compared with the methods based on single-view. However, only a single type of representation is used by these methods, *e.g.*, 3DMM parameters, vertex positions, or DAMs, it is not sufficient for describing the 3D face shapes comprehensively in complex environments.

In this paper, we propose a novel framework which consists of multi-reconstruction branches and a mesh retrieval module (see Figure 1), to overcome the representing limitation of previous single-reconstruction methods and enhance the accuracy of 3D shape reconstruction from 2D videos. The recent state-of-the-art 3D face networks which have shown their advantages based on a single image or a small amount of images, are introduced to form the ensemble of independent regression branches. In each branch, the meshes of all frames are reconstructed individually, then the candidate 3D shape is synthesized by weighted linear combination. This step helps reduce the errors of depth estimation and invisible regions reconstruction under a single camera view. At last, the best fitting mesh is retrieved according to the distance between the synthesized texture and the ground truth texture of the real 2D face image. To evaluate our algorithm, we implement different settings of our method to make a complete analysis. Extensive experiments show that our approach is able to obtain competitive results near the best theoretical predictions, which are reconstructed by the external 3DMM model and the ground truth 3D sparse landmarks. In the testing phases of the 2^{nd} 3DFAW challenge [11], our method also achieves superior performance over most of submissions by other participating teams.

To summarize, in this work we make the following main contributions: 1) A novel framework consisting of multi-reconstruction branches and a mesh retrieval module is presented to handle 3D face reconstruction from 2D videos. 2) We conduct comprehensive experiments on 3DFAW-Video dataset to evaluate our method, and our approach performs very well on solving the 3D face alignment and reconstruction problem from 2D videos.

2. Our Method

In this section, we first review the goal of 3DMM Reconstruction from 2D images and videos. Then we introduce the detail of the multi-reconstruction branches and the combination formulations of multiple frames for each branch. At last, we discuss how to retrieve the best fitting mesh according to the outputs of different branches.

2.1. 3D Model

The geometry of a 3D face is denoted as a shape vector $\mathbf{S} \in \Re^{3\times n}$ with total n vertices. A common assumption is that a new shape of 3D face can be modeled by 3DMM [3] with a linear combination of the average shape and the principal components:

$$\mathbf{S} = \bar{\mathbf{S}} + \boldsymbol{\alpha}_{id}\mathbf{A}_{id} + \boldsymbol{\alpha}_{exp}\mathbf{A}_{exp}, \tag{1}$$

where $\bar{\mathbf{S}}$ is the average shape, $\boldsymbol{\alpha}_{id}$ and $\boldsymbol{\alpha}_{exp}$ are the coefficients of the identity and expression eigenvectors, respectively, $\mathbf{A}$ is the principal component.

Given the 3DMM model parameter $\zeta = \{\bar{\mathbf{S}}, \mathbf{A}_{id}, \mathbf{A}_{exp}\}$, the sparse 2D shape $\mathbf{s} \in \Re^{2\times l}$ with l landmarks on the static image I, the goal of 3D reconstruction is to estimate the intrinsic camera parameters $\mathbf{A} \in \Re^{3\times 3}$, rotation matrix

$\mathbf{R} \in \Re^{3\times3}$, translation vector $\mathbf{t} \in \Re^{3\times1}$, and the face parameter $\mathbf{\Phi} = \{\boldsymbol{\alpha}_{id}, \boldsymbol{\alpha}_{exp}\}$. To find $\mathbf{\Phi}$ and the projection parameter $\mathbf{\Omega} = \{\mathbf{A}, \mathbf{R}, \mathbf{t}\}$ that best fits the 3D face model to the 2D landmarks, the following nonlinear least squares optimization problem can be solved by:

$$\begin{aligned} \left[\hat{\mathbf{\Omega}}, \hat{\mathbf{\Phi}}\right] &= \min_{\mathbf{\Omega},\mathbf{\Phi}} \left\|\mathbf{f}(\mathbf{\Omega}, \mathbf{\Phi}) - \mathbf{s})\right\|^2, \\ \mathbf{f} &= \mathbf{f}_1 \circ \mathbf{f}_2, \\ \mathbf{f}_1(\mathbf{\Omega}, \mathbf{\Phi}) &= [\mathbf{A}(\mathbf{RS} + \mathbf{T})]_{\text{sparse}}, \\ \mathbf{f}_2(\mathbf{S}) &= \begin{bmatrix} \mathbf{S}_1^\top \oslash \mathbf{S}_3^\top \\ \mathbf{S}_2^\top \oslash \mathbf{S}_3^\top \end{bmatrix}, \end{aligned} \tag{2}$$

where $\mathbf{T} = [\mathbf{t}, \mathbf{t}, ...] \in \Re^{3\times n}$ consists of n copies of $\mathbf{t}$, the subscript sparse means that only the sparse 3D vertices corresponding with $\mathbf{s}$ are selected, $\mathbf{f}$ projects the 3D vertices into the coordinate space of I, $\oslash$ denotes element-wise division, $\mathbf{S}_i$ is the i^{th} row vector.

The 3DFAW challenge only uses the face mesh reconstructed from a 2D video $\boldsymbol{\nu} = \{I_1, I_2, ..., I_K\}$ with K frames while excluding each projection parameter $\mathbf{\Omega}_k$ on each frame for submission, so the solution of face reconstruction based on multiple frames can be formulated as:

$$\hat{\mathbf{\Phi}} = \min_{\mathbf{\Phi}} \sum_{k=1}^{K} \left\|\mathbf{f}(\mathbf{\Omega}_k, \mathbf{\Phi}) - \mathbf{s}_k)\right\|^2. \tag{3}$$

2.2. Multi-Reconstruction Branches

Figure 1 shows the flowchart of three reconstruction branches, $\mathbf{B}_{3D}$, $\mathbf{B}_P$, $\mathbf{B}_M$, they are built on the recent successful 3D face alignment networks, 3DDFA [28], PRNet [7] and MVF-Net [22] respectively. Given the video $\boldsymbol{\nu}$, the multi-reconstruction branches regress face meshes independently. They have different structures with different representations of 3D face shapes. $\mathbf{B}_{3D}$ deploys an unified CNN structure across the cascade to regress the 3DMM parameter $\mathbf{\Theta}_{3D}^k = \left[\mathbf{\Omega}_{3D}^k, \mathbf{\Phi}_{3D}^k\right]$ from the input of each single frame I_k. $\mathbf{B}_P$ uses an encoder-decoder structure to generate a UV position map from I_k, and the map can be converted directly to the 3D face shape $\mathbf{S}_P^k$. $\mathbf{B}_M$ takes a triplet T consisting of a front, left, and right view frame as its input. For the current frame I_k, two frames with the other views are randomly selected from $\boldsymbol{\nu}$ to construct the triplet T_k. Three features are extracted from a shared weight CNN separately, and they are concatenated together to regress the 3D shape $\mathbf{\Theta}_M^k = \left[\mathbf{\Omega}_M^k, \mathbf{\Phi}_M^k\right]$.

2.3. Fusion of Multi-frame Parameters

Three sets of 3D face shapes, $\Gamma_{3D} = \{\mathbf{\Theta}_{3D}^1, ..., \mathbf{\Theta}_{3D}^K\}$, $\Gamma_P = \{\mathbf{S}_P^1, ..., \mathbf{S}_P^K\}$, $\Gamma_M = \{\mathbf{\Theta}_M^1, ..., \mathbf{\Theta}_M^K\}$ are predicted by the the multi-reconstruction branches separately. For the shape set of each branch, all the elements are weighted linear combined together to form an optimal overall solution to improve the quality of face reconstruction.

For the sets Γ_{3D} and Γ_M, only identity parameters are utilized to fuse the optimal face shape for the input video $\boldsymbol{\nu}$:

$$\mathbf{S}_{3D} = \bar{\mathbf{S}} + \frac{\sum_{k=1}^{K} \omega_{3D,id}^k \boldsymbol{\alpha}_{3D,id}^k}{\sum_{k=1}^{K} \omega_{3D,id}^k} \mathbf{A}_{3D,id}, \tag{4}$$

$$\mathbf{S}_{M} = \bar{\mathbf{S}} + \frac{\sum_{k=1}^{K} \omega_{M,id}^k \boldsymbol{\alpha}_{M,id}^k}{\sum_{k=1}^{K} \omega_{M,id}^k} \mathbf{A}_{M,id}, \tag{5}$$

where ω_{id} represents the weight value for identity coefficients. For Γ_{3D}, the weight value is inversely proportional to the absolute value of the facial yaw angle, because the reconstruction shape becomes less accurate on a single frame as the facial pose increases. For Γ_M based on multi-view frames, their weight values are all set to 1.

For the set Γ_P, all the shapes are firstly rotated to the frontal view and aligned to the reference shape, and then fused to one shape with the weights:

$$\mathbf{S}_P = \frac{\sum_{k=1}^{K} \omega_P^k \mathbf{S}_P^k}{\sum_{k=1}^{K} \omega_P^k}, \tag{6}$$

where the value of ω_P^k is also inversely proportional to the absolute value of facial yaw angle on I_k.

2.4. Mesh Retrieval

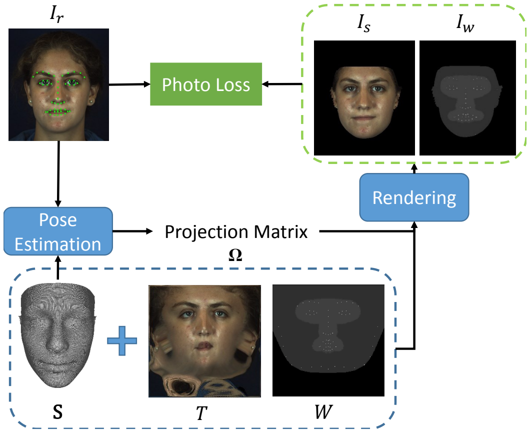

Figure 2. The diagram of the mesh retrieval module.

After the fused meshes $\mathbf{S}_{3D}$, $\mathbf{S}_P$, and $\mathbf{S}_M$ are obtained by the multi-reconstruction branches, inspired by [22], we derive a weighted photo distance in the mesh retrieval module to select the best fitting face shape as the final result. As shown in Figure 2, according to the annotated 51 2D facial landmarks $\mathbf{s}$, the projection matrix $\mathbf{\Omega}_c$ is estimated between the candidate shape $\mathbf{S}_c \in \{\mathbf{S}_{3D}, \mathbf{S}_P, \mathbf{S}_M\}$ and the real 2D frontal frame I_r. For each branch, the synthesized image I_s is reconstructed by $\mathbf{S}_c$, and its corresponding UV texture map T_c which is the weighted mean of the UV texture maps of all frames, and the weighted mask W. The pixel values of W follow the setting of [7]. We assume that the optical

3D mesh has the best photometric consistency between I_r and I_s. The weighted photo distance is defined as:

$$D = \sum_{u=0}^{W-1} \sum_{v=0}^{H-1} \left\| I_r\left(u, v\right) - I_s\left(u, v\right) \right\| I_w\left(u, v\right), \tag{7}$$

where I_r, I_s and I_w are the input frontal face, the synthesized face and facial mask image respectively, they have the same width W and height H. The mesh with the minimum distance is selected as the final reconstruction result of $\boldsymbol{\nu}$.

3. Experiments

3.1. Datasets

Training Datasets. The 300W-LP dataset [28], which contains over 60,000 images across different identities, poses and expressions with fitted 3DMM parameters, is used for training the models based on the proposed framework.

Evaluation Datasets. Our method is evaluated on the 3DFAW-Video dataset, which consists of a large corpora of profile-to-profile face videos annotated with corresponding high-resolution 3D ground truth meshes. It is divided into the training, validation and testing sets. The training set contains 26 subjects with neutral expression and available ground-truth 3D meshes. Each subject contains two videos, a high-definition in-the-lab video, and an unconstrained video captured from an iPhone device. The validation and testing sets contain 14 and 26 subjects respectively. Contrast to the training set, only one type video, the in-the-lab or the unconstrained video is provided for each subject. For each video, a frame with frontal face is annotated with 51 facial landmarks, which are extracted from the standard 68 dlib [12] landmarks. The qualities of reconstructed meshes in this challenge are evaluated by using the metric of Average Root Mean Square Error (ARMSE).

3.2. Implementation Details

For the branch training of $\mathbf{B}_{3D}$, we select MobileNetV2 [17] as the backbone network. The weight of the loss for regressing identity parameters is set to 3, while the weight of the loss for learning expression parameters is set to 1, it makes the model training pay more attentions on shape parameters regression. For optimization, we use SGD optimizer with a learning rate begins at 0.001 and decays half after each 10 epoches. The batch size is set to 512 and the total epoch is set to 50. As for the branches $\mathbf{B}_P$ and $\mathbf{B}_M$, their initial public released models are used directly for the subsequent parameter fusion and mesh retrieval.

3.3. Experimental Results

We first conduct a series of experiments on the training set to demonstrate the effectiveness of the multi-reconstruction branches and the mesh retrieval module. Table 1 lists the comparison of our models with different settings on the training set. It is noted that the fusion method based on multi-view performs much better than the method based on single-frame, and the mesh retrieval module significantly improves the reconstruction accuracy compared with the single-branch methods.

Method	w/o Fusion		w Fusion	
	HiRes	iPhone	HiRes	iPhone
$\mathbf{B}_{3D}$	2.24	2.28	2.22	2.23
$\mathbf{B}_P$	2.25	2.47	2.04	2.18
$\mathbf{B}_M$	2.32	2.32	2.24	2.26
Our Method	-	-	**1.89**	**2.02**

Table 1. The ARMSE comparison of different settings based our method on the training dataset of 3DFAW-Video.

Method	$\mathbf{B}_{3D}$	$\mathbf{B}_P$	$\mathbf{B}_M$	Our Method
ARMSE	2.28	2.02	2.33	**1.86**

Table 2. The ARMSE comparison on the testing dataset of 3DFAW-Video.

In the training set, 51 sparse 3D landmarks of each subject can be extracted from the ground truth mesh. So the maximum reconstruction accuracy of the methods based on external 3D models can be approximated by the ”pseudo” ground truth meshes which are reconstructed based on the sparse landmarks and the BFM model [15]. These meshes are evaluated and get the ARMSE of 1.76. It shows that our method on the in-the-lab videos is able to obtain a competitive performance (ARMSE = 1.89) near the best theoretical results.

At last, we submit the results predicted by our method on the testing dataset. Table 2 shows the performance achieved by our methods with different settings. The ARMSE of 1.86 outperforms most of the results submitted by other teams.

4. Conclusion

This paper introduces the submission to the 2^{nd} 3DFAW Challenge. The multi-reconstruction branches and the mesh retrieval module are presented to integrate the advantages of some recent 3D alignment methods with different feature representations and enhance the robustness of 3D face reconstruction from 2D videos. The experiments on the dataset of 3DFAW-Video show the significant performance of our method.

Acknowledgement

This work is partially supported by the Project from National Key Research and Development Program of China (Grant No. 2018YFC0808300), National Natural Science Foundation of China (Grant No. 6180021609, 6180070559, 61672519, 61602433), and the CAS “Light of West China” Program.

References

[1] Riza Alp Guler, George Trigeorgis, Epameinondas Antonakos, Patrick Snape, Stefanos Zafeiriou, and Iasonas Kokkinos. Densereg: Fully convolutional dense shape regression in-the-wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6799–6808, 2017. 1

[2] Thomas Berg and Peter N Belhumeur. Tom-vs-pete classifiers and identity-preserving alignment for face verification. In *Proceedings of British Machine Vision Conference*, 2012. 1

[3] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194. ACM Press/Addison-Wesley Publishing Co., 1999. 1, 2

[4] Adrian Bulat and Georgios Tzimiropoulos. Two-stage convolutional part heatmap regression for the 1st 3d face alignment in the wild (3dfaw) challenge. In *European Conference on Computer Vision*, pages 616–624. Springer, 2016. 1

[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 *Proceedings of the IEEE International Conference on Computer Vision*, pages 1021–1030, 2017. 1

[6] Chen Cao, Qiming Hou, and Kun Zhou. Displaced dynamic expression regression for real-time facial tracking and animation. *ACM Transactions on Graphics (TOG)*, 33(4):43, 2014. 1

[7] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *European Conference on Computer Vision*, pages 534–551, 2018. 1, 3

[8] Yuxiao Hu, Dalong Jiang, Shuicheng Yan, Lei Zhang, and Hongjiang Zhang. Automatic 3d reconstruction for face recognition. In *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on*, pages 843–848. IEEE, 2004. 1

[9] László A Jeni, Jeffrey F Cohn, and Takeo Kanade. Dense 3d face alignment from 2d videos in real-time. In *2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, volume 1, pages 1–8. IEEE, 2015. 1

[10] László A Jeni, Sergey Tulyakov, Lijun Yin, Nicu Sebe, and Jeffrey F Cohn. The first 3d face alignment in the wild (3dfaw) challenge. In *European Conference on Computer Vision*, pages 511–520. Springer, 2016. 1

[11] Laszlo A. Jeni, Huiyuan Yang, Rohith K. Pillai, Zheng Zhang, Jeffrey Cohn, and Lijun Yin. 3d dense face reconstruction from video (3dfaw-video) challenge. In *2nd Workshop and Challenge on 3D Face Alignment in the Wild Dense Reconstruction from Video (3DFAW-Video) 2019, in conjunction with IEEE International Conference on Computer Vision (ICCV), 2019*. 2

[12] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1867–1874, 2014. 4

[13] Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis. Fast face-swap using convolutional neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3677–3685, 2017. 1

[14] Jiangjing Lv, Xiaohu Shao, Junliang Xing, Cheng Cheng, and Xi Zhou. A deep regression architecture with two-stage re-initialization for high performance facial landmark detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1

[15] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 296–301. Ieee, 2009. 4

[16] Marcel Piotraschke and Volker Blanz. Automated 3d face reconstruction from multiple images using quality measures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3418–3427, 2016. 1

[17] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018. 4

[18] Baoguang Shi, Xiang Bai, Wenyu Liu, and Jingdong Wang. Deep regression for face alignment. *arXiv preprint arXiv:1409.5230*, 2014. 1

[19] Xiaoguang Tu, Jian Zhao, Zihang Jiang, Yao Luo, Mei Xie, Yang Zhao, Linxiao He, Zheng Ma, and Jiashi Feng. Joint 3d face reconstruction and dense face alignment from a single image with 2d-assisted self-supervised learning. *arXiv preprint arXiv:1903.09359*, 2019. 1

[20] Sergey Tulyakov and Nicu Sebe. Regressing a 3d face shape from a single image. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3748–3755, 2015. 1

[21] Huawei Wei, Shuang Liang, and Yichen Wei. 3d dense face alignment via graph convolution networks. *arXiv preprint arXiv:1904.05562*, 2019. 1

[22] Fanzi Wu, Linchao Bao, Yajing Chen, Yonggen Ling, Yibing Song, Songnan Li, King Ngi Ngan, and Wei Liu. Mvf-net: Multi-view 3d face morphable model regression. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 959–968, 2019. 1, 3

[23] Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. Look at boundary: A boundary-aware face alignment algorithm. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2129–2138, 2018. 1

[24] Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2013. 1

[25] Xuehan Xiong and Fernando De la Torre. Global supervised descent method. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 1

[26] Jae Shin Yoon, Takaaki Shiratori, Shoou-I Yu, and Hyun Soo Park. Self-supervised adaptation of high-fidelity face models for monocular performance tracking. *arXiv preprint arXiv:1907.10815*, 2019. 2

[27] Xiangyu Zhu, Zhen Lei, Stan Z Li, et al. Face alignment in full pose range: A 3d total solution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 1

[28] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In *IEEE conference on computer vision and pattern recognition*, pages 146–155, 2016. 1, 3, 4

[29] Xiangyu Zhu, Zhen Lei, Junjie Yan, Dong Yi, and Stan Z Li. High-fidelity pose and expression normalization for face recognition in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 787–796, 2015. 1