# Joint Face Alignment and 3D Face Reconstruction with Efficient Convolution Neural Networks

Keqiang Li, Huaiyu Wu, Xiuqin Shang, Zhen Shen,
Gang Xiong, Xisong Dong, Bin Hu and Fei-Yue Wang

Institute of Automation, Chinese Academy of Sciences, Beijing, China.

Corresponding author: Huaiyu Wu. E-mail: huaiyu.wu@ia.ac.cn

## 1 Abstract

***Goal:*** Face Alignment and 3D Face Reconstruction

***Challenge:*** Recent methods based on CNN typically aim to learn parameters of 3D Morphable Model (3DMM) from 2D images to render face alignment and 3D face reconstruction. Most algorithms are designed for faces with small, medium yaw angles, which is extremely challenging to align faces in large poses. At the same time, they are not efficient usually.The main challenge is that it takes time to determine the parameters accurately.

***Contribution:*** This paper proposes a efficient network structure through Depthwise Separable Convolution and Muti-scale Representation and dual attention mechanisms together.

For training, two loss functions are used to constraint and optimize 3DMM parameters and 3D vertices. We finally provide a light-weighted framework which can produce dense face alignment and 3D reconstruction results with strong robustness to large poses, illuminations and occlusions.

Comparison on the challenging AFLW2000-3D and AFLW datasets shows that our method achieves significant performance on both tasks of 3D face reconstruction and face alignment.

## 2 Peoposed Method

In [1], the 3D Morphable Model (3DMM) which describes the 3D face space with PCA is proposed. It is expressed as follows:

$$S = \overline{S} + A_{id}\alpha_{id} + A_{exp}\alpha_{exp} \tag{1}$$

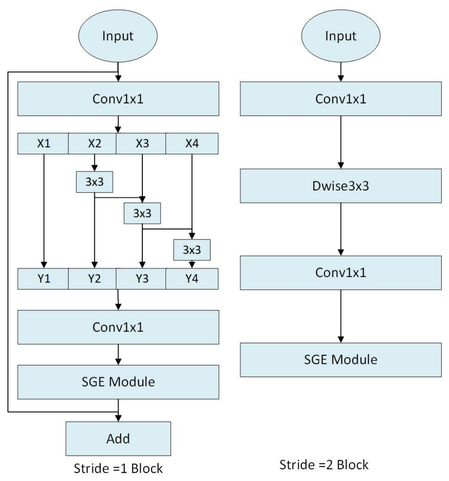
the face $S$ can be projected onto the 2D image plane with the scale orthographic projection. This process can be expressed as follows:

$$V = f * Pr * R * \left(\overline{S} + A_{id}\alpha_{id} + A_{exp}\alpha_{exp}\right) + t \tag{2}$$

Putting them together, we have in total 62 parameters $P = [f, pitch, yaw, roll, t, \alpha_{id}, \alpha_{exp}]$ for 3D face reconstruction. Following [2], we employ the Weighted Parameter Distance Cost (WPDC) to supervise the model training where The basic idea is explicitly modeling the importance of each parameter. We make use of Wing Loss to constrain 3D face vertices.

## 3 Network Structure

Based on Mobilenetv2 [3], we design a novel and efficient network structure named Mobile-FRNet which transfers the input RGB image into parameters. It applies depthwise separable convolution, muti-scale representation, we introduce a lightweight attention mechanism for space and channel dimensions respectively as in [4].The Mobile-FRNet architecture is illustrated in **Figure 1**.




| Operator | t | c | n | s |
|---|---|---|---|---|
| conv2d | - | 32 | 1 | 2 |
| Layer1 | 1 | 16 | 1 | 1 |
| SE Module | - | - | - | - |
| Layer2 | 6 | 24 | 2 | 2 |
| SE Module | - | - | - | - |
| Layer3 | 6 | 32 | 3 | 2 |
| SE Module | - | - | - | - |
| Layer4 | 6 | 64 | 4 | 2 |
| SE Module | - | - | - | - |
| Layer5 | 6 | 96 | 3 | 1 |
| SE Module | - | - | - | - |
| Layer6 | 6 | 160 | 3 | 2 |
| SE Module | - | - | - | - |
| Layer7 | 6 | 320 | 1 | 1 |
| SE Module | - | - | - | - |
| conv2d1 × 1 | - | 1280 | 1 | 1 |
| avgpool7 × 7 | - | - | 1 | - |
| conv2d1 × 1 | - | k | - | - |

**Figure 1:** The details of MobileBlock and Network Architecture

The convolution layers of a set of filters and SGE Module are called MobileBlock in **Figure 1**. We introduce multi-scale representation [5], these smaller filter groups are connected similar to residuals.

## 4 Experiments

Follow the work [2], we choose baseline methods including 3DDFA, 3DSTN, DeFA, Nonlinear 3DMM, DAMDNet to evaluate the face alignment performance.

**Table 1:** Performance comparison on AFLW2000-3D(68 landmarks) and AFLW (21 landmarks).

| | AFLW DataSet (21 pts) | | | | | AFLW2000-3D Dataset (68 pts) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Method | $[0° - 30°]$ | $[30° - 60°]$ | $[60° - 90°]$ | Mean | Std | $[0° - 30°]$ | $[30° - 60°]$ | $[60° - 90°]$ | Mean | Std |
| 3DDFA [6] | 5.000 | 5.060 | 6.740 | 5.600 | 0.990 | 3.780 | 4.540 | 7.930 | 5.420 | 2.210 |
| 3DDFA+SDM [6] | 4.750 | 4.830 | 6.380 | 5.320 | 0.920 | 3.430 | 4.240 | 7.170 | 4.940 | 1.970 |
| 3DSTN [11] | - | - | - | - | - | 3.150 | 4.330 | 5.980 | 4.490 | - |
| DeFA [3] | - | - | - | - | - | - | - | - | 4.500 | - |
| Nonlinear 3DMM [42] | - | - | - | - | - | - | - | - | 4.700 | - |
| DAMDNet [15] | 4.359 | 5.209 | 6.028 | 5.199 | 0.682 | **2.907** | 3.830 | 4.953 | 3.897 | 0.837 |
| **Mobile-FRNet** | **4.199** | **4.862** | **5.668** | **4.910** | **0.601** | 2.930 | **3.799** | **4.768** | **3.832** | **0.751** |

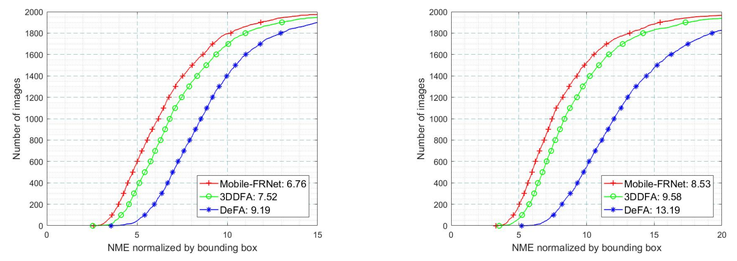



**Figure 2:** Performance on all points with both the 2D (left) and 3D (right) coordinates.

We also compare our dense alignment result with other baseline methods on the task of dense alignment in **Figure 2**.

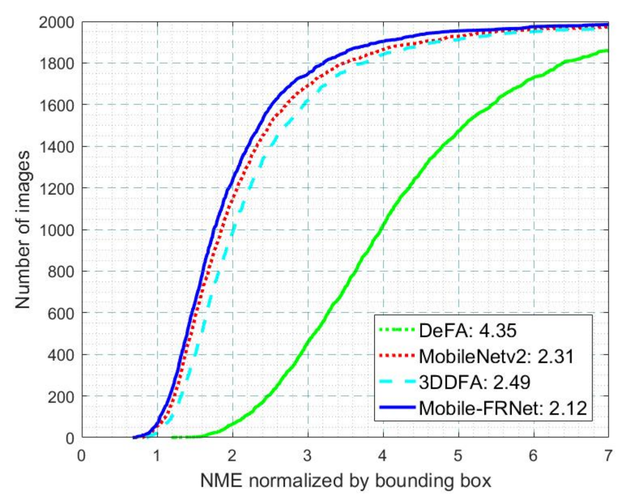



**Figure 3:** 3D reconstruction performance on AFLW2000-3D dataset.

We employ NME to evaluate our method on the task of 3D face reconstruction. We choose baseline methods including 3DDFA, DeFA, MobileNet_v2 in **Figure 3**.

The experimental network structures include ResNeXt50, MobileNetV2, DenseNet121 and our proposed Mobile-FRNet. Params and GFLOPs with the different network structures in **Table 2**.

**Table 2:** different network structures.

| | | | AFLW DataSet(21 pts) | | | | | AFLW2000-3D DataSet(68 pts) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Net | Params(M) | GFLOPs | $[0° - 30°]$ | $[30° - 60°]$ | $[60° - 90°]$ | Mean | Std | $[0° - 30°]$ | $[30° - 60°]$ | $[60° - 90°]$ | Mean | Std |
| ResNeXt50 [43] | 23.11 | 1.319 | 4.599 | 5.516 | 6.297 | 5.471 | 0.694 | 3.122 | 4.065 | 5.351 | 4.179 | 0.913 |
| Mobilenet_v2 [37] | **2.38** | **0.109** | 4.643 | 5.581 | 6.397 | 5.540 | 0.716 | 3.236 | 4.080 | 5.181 | 4.165 | 0.796 |
| DenseNet121 [44] | 7.02 | 0.800 | 4.442 | 5.249 | 6.168 | 5.286 | 0.705 | 3.051 | 3.912 | 5.297 | 4.087 | 0.925 |
| Mobile-FRNet(no attention) | 2.40 | 0.110 | **4.371** | **5.199** | **6.031** | **5.201** | **0.678** | **2.962** | **3.856** | **4.991** | **3.936** | **0.830** |
| Mobile-FRNet | 2.60 | 0.120 | **4.199** | **4.862** | **5.668** | **4.910** | **0.601** | **2.930** | **3.799** | **4.768** | **3.832** | **0.751** |

## References

[1] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in *Proceedings of the 26$^{th}$ Annual Conference on Computer Graphics and Interactive Techniques*, 1999, pp. 187–194.

[2] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3d solution," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2016, pp. 146–155.

[3] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.

[4] L. Jiang, X.-J. Wu, and J. Kittler, "Dual attention mobdensenet (damdnet) for robust 3d face alignment," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. IEEE, 2019, pp. 504–513.

[5] S. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. H. Torr, "Res2net: A new multi-scale backbone architecture," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

## Acknowledgements

National Natural Science Foundation of China (under Grants No. 61872365, U1909204, 61773381, 61773382, U1909218); Zhong-Shan Talent Plan, Guangdong; Chinese Guangdong's S&T project (2019B1515120030).