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

likeqiang2020@ia.ac.cn

State Key Laboratory for Management and Control of Complex Systems
Institute of Automation, Chinese Academy of Sciences

The 25th International Conference on Pattern Recognition
(ICPR2020)

Overview

➢ Introduction

- **Task:**

  Face Alignment and 3D Face Reconstruction

- **Problem：**

  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.

Overview

➢ Introduction

- **Challenge:**

  The problem is usually time consuming and difficult to learn parameters accurately.

- **Our Contribution:**

  (1) We propose a efficient network structure through Depthwise Separable Convolution and Muti-scale Representation and dual attention mechanisms together.

  (2) For training, two cost functions are used to constraint and optimize 3DMM parameters and 3D vertices. We finally provide a light-weighted framework.

  (3) 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.

Related Work

- A variety of solutions have been proposed to solve the problems

- Model-based
  - 3DDFA
  - DeFA
  - 2DASL
  - Deep3DFace

……

- Not Model-based
  - VRNet
  - PRNet

……

Proposed Method

➢ 3D morphable model

- Fitting a dense 3D morphable model(3DMM) instead of detecting landmarks, which describes the 3D face space with PCA

$$S = \overline{S} + A_s \alpha_s + A_{exp} \alpha_{exp}$$

$$\bar{S} \in \mathbb{R}^{3N}, A_s \in \mathbb{R}^{3N \times 40}, A_{exp} \in \mathbb{R}^{3N \times 10}, \alpha_s \in \mathbb{R}^{40}, \alpha_{exp} \in \mathbb{R}^{10}$$

- $\boldsymbol{S}$ can be projected onto the 2D image plane with the scale orthographic projection to generate a 2D face

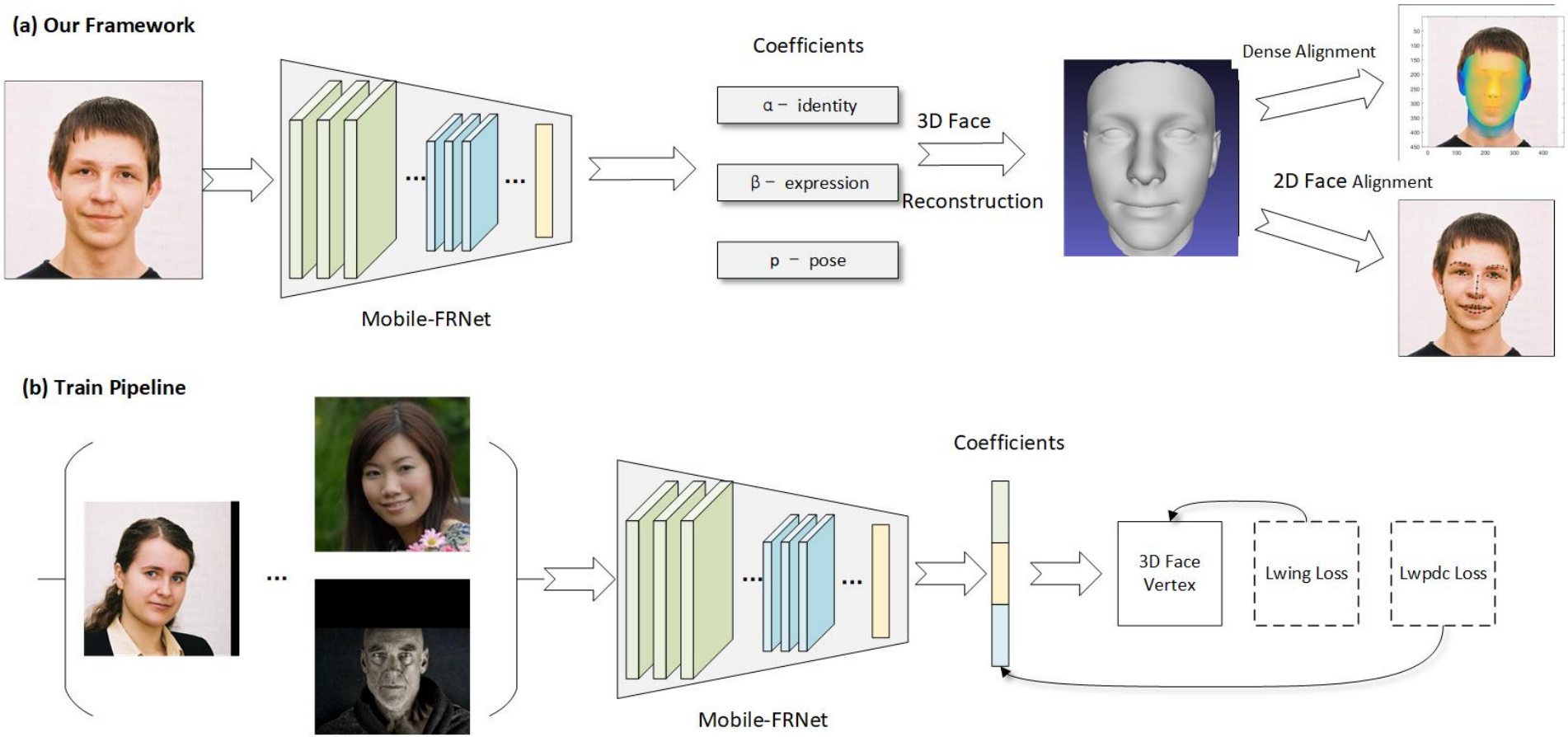
$$V = f * Pr * \Pi * S + t$$

- Putting them together, we have in total 62 parameters

$$p = \left[\alpha_s \quad \alpha_{exp} \quad f \quad t \quad \Pi \right]$$

Proposed Method

➢ Model Overview



Proposed Method

➢ Network Structure

- Based on Mobilenetv2 , we design a novel and efficient network structure named Mobile-FRNet.
- It applies depthwise separable convolution, muti-scale representation, channel attention, spatial attention mechanism.

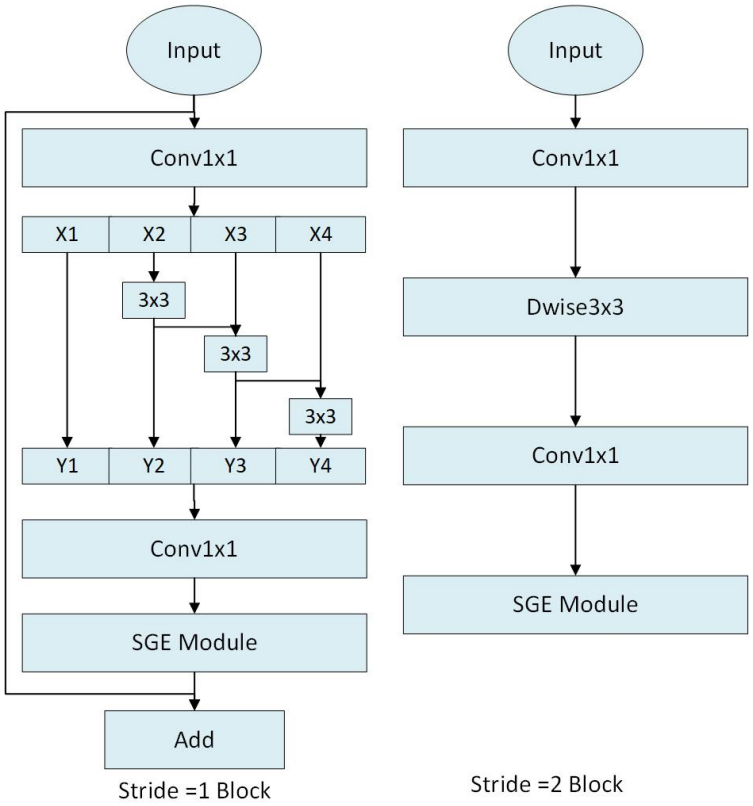
Operator	t	c	n	s
conv2d	-	32	1	2
Layer1	1	16	1	1
SE Module	-	-	-	-
Layer2	6	24	2	2
SE Module	-	-	-	-
Layer3	6	32	3	2
SE Module	-	-	-	-
Layer4	6	64	4	2
SE Module	-	-	-	-
Layer5	6	96	3	1
SE Module	-	-	-	-
Layer6	6	160	3	2
SE Module	-	-	-	-
Layer7	6	320	1	1
SE Module	-	-	-	-
conv2d1 × 1	-	1280	1	1
avgpool7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

- Each row describes a sequence consisting of one or more identical (stride) layers, repeated n times. All layers in the same sequence have the same number of output channels **c**. The first layer of each sequence has a stride **s**, all other layers use stride **1**. The expansion factor **t** is always applied to the input size

Proposed Method

➢ Network Structure

The convolution layers of a set of filters and SGE Module are called MobileBlock.



- channel dimension: SE module
- spatial dimension : SGE module
- MobileBlock are repeated n times for extracting deep features, an SE Module is added between each Layer
- multi-scale representation and SGE module are applied in MobileBlock.

Proposed Method

➢ Loss Function

Employ the Weighted Parameter Distance Cost (WPDC) to learn parameter

$$\mathcal{L}_{wpdc} = (p_g - p)^T W (p_g - p)$$

Make use of Wing Loss to constrain 3D face vertices as follows:

$$\mathcal{L}_{wing} = \begin{cases} \omega \ln(1 + |\Delta V(p)|/\epsilon) & \text{if} \quad |\Delta V(p)| < \omega \\ |\Delta V(p)| - C & \text{otherwise} \end{cases}$$

Thus the overall training loss

$$\mathcal{L} = \lambda_1 \mathcal{L}_{wpdc} + \lambda_2 \mathcal{L}_{wing}$$

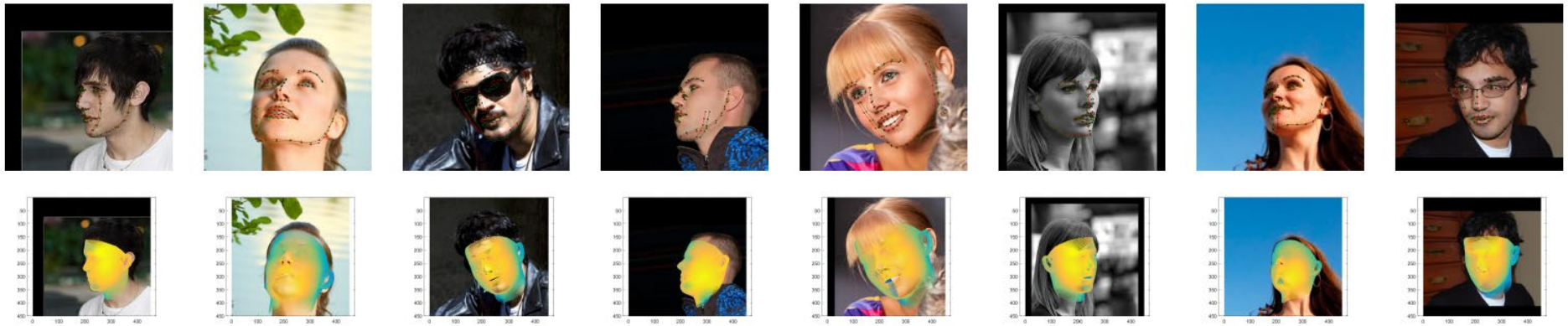
Experiment

- **Train data:**
  300W-LP
- **Test dataset :**
  (1) AFLW.
  (2) ALFW2000-3D.
- **The metric to evaluate the performance.**
  Normalized Mean Error (NME)

Experiment

➢ Face Alignment

Select some face images for qualitative testing in ALFW2000-3D dataset randomly.

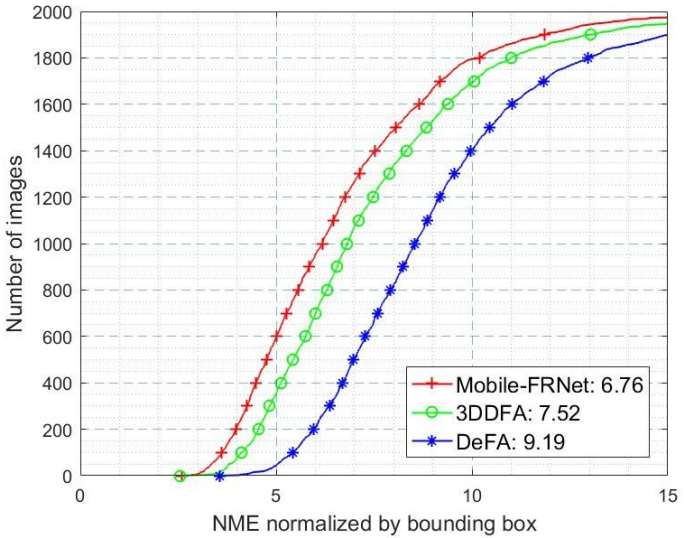


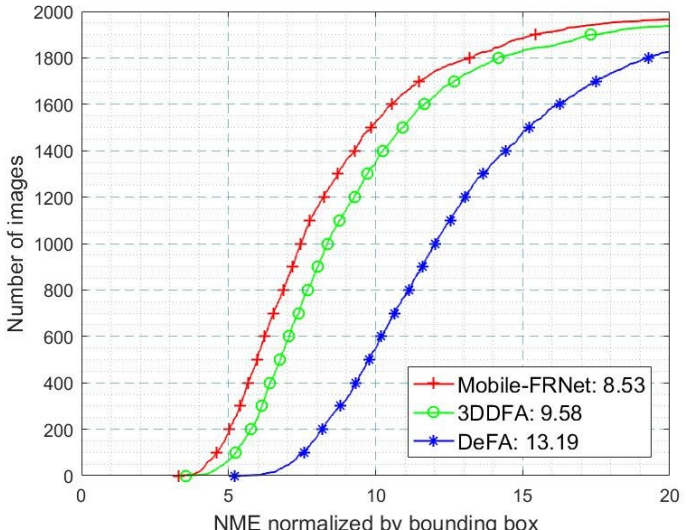
Experiment

Performance comparison on AFLW2000-3D(68 landmarks) and AFLW (21 landmarks).

	AFLW DataSet (21 pts)					AFLW2000-3D Dataset (68 pts)				
Method	$[0^o - 30^o]$	$[30^o - 60^o]$	$[60^o - 90^o]$	Mean	Std	$[0^o - 30^o]$	$[30^o - 60^o]$	$[60^o - 90^o]$	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**

Evaluation is performed on all points with both the 2D (left) and 3D (right) coordinates.

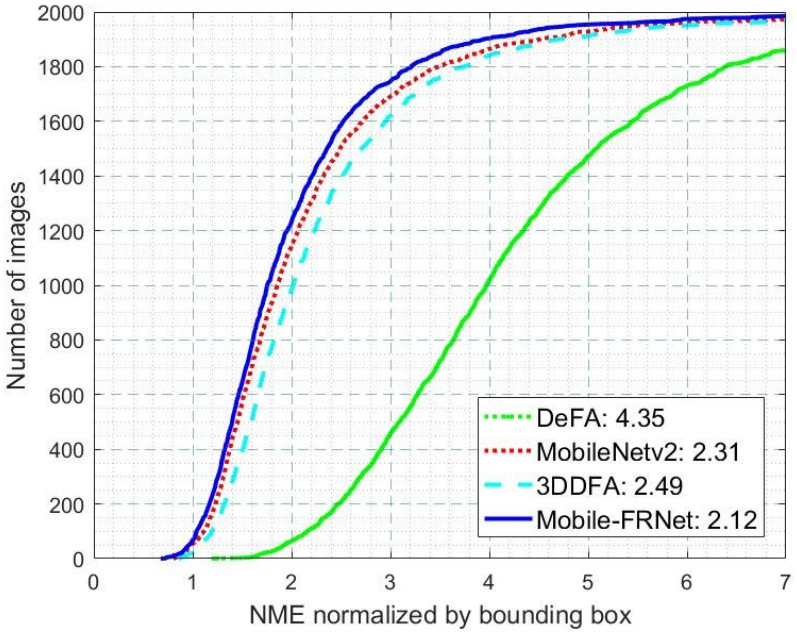




Experiment

➢ 3D Face Reconstruction

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



Experiment

➢ Comparisons of Different Networks Structures

The experimental network structures include ResNeXt50,MobileNetV2, DenseNet121 and our proposed Mobile-FRNet.

			AFLW DataSet(21 pts)					AFLW2000-3D DataSet(68 pts)				
Net	Params(M)	GFLOPs	$[0^o - 30^o]$	$[30^o - 60^o]$	$[60^o - 90^o]$	Mean	Std	$[0^o - 30^o]$	$[30^o - 60^o]$	$[60^o - 90^o]$	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**

Conclusion

- We propose a method Mobile-FRNet which simultaneously completes 3D face reconstruction and provides dense alignment results from an input 2D face image.
- Quantitative and qualitative results show that our method is robust to large poses and occlusions. Experiments on two challenging face datasets illustrate the effectiveness of Mobile-FRNet on both 3D face reconstruction and face alignment by comparing with other methods.
- Our method also makes a good compromise between accuracy and efficiency.

Reference

[1] 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, pp. 146–155, 2016.

[2] 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), pp. 504–513, IEEE, 2019.

[3] Y. Deng, J. Yang, S. Xu, D. Chen, Y. Jia, and X. Tong, “Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set,” in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 285–295, IEEE, 2019.

[4] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou, “Joint 3d face reconstruction and dense alignment with position map regression network,” in Proceedings of the European Conference on Computer Vision (ECCV), pp. 534–551, 2018.

[5] Y. Liu, A. Jourabloo, W. Ren, and X. Liu, “Dense face alignment,” in Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 1619–1628, 2017.

[6] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141, 2018.

[7] X. Li, X. Hu, and J. Yang, “Spatial group-wise enhance: Enhancing semantic feature learning in convolutional networks,” arXiv preprint arXiv:1905.09646, 2019.

Thanks

Acknowledgment

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).