Robust 3D Face Alignment with Efficient Fully Convolutional Neural Networks

Lei Jiang[1,3], Xiao-Jun Wu[1,3](✉), and Josef Kittler[2]

[1] School of IoT Engineering, Jiangnan University, Wuxi 214122, China
ljiang_jnu@outlook.com, xiaojun_wu_jnu@163.com

[2] Center for Vision, Speech and Signal Processing (CVSSP), University of Surry,
Guildford GU2 7XH, UK
j.kittler@surrey.ac.uk

[3] Jiangsu Provincial Engineering, Laboratory of Pattern Recognition
and Computational Intelligence, Jiangnan University, Wuxi 214122, China

Abstract. 3D face alignment from monocular images is a crucial process in computer vision with applications to face recognition, animation and other areas. However, most algorithms are designed for faces in small to medium poses (below 45°), lacking the ability to align faces in large poses up to 90°. At the same time, many methods are not efficient. The main challenge is that it is time consuming to determine the parameters accurately. In order to address this issue, this paper proposes a novel and efficient end-to-end 3D face alignment framework. We build an efficient and stable network model through Depthwise Separable Convolution and Densely Connected Convolutional, named MobDenseNet. Simultaneously, different loss functions are used to constrain 3D parameters based on 3D Morphable Model (3DMM) and 3D vertices. Experiments on the challenging AFLW, AFLW2000-3D databases show that our algorithm significantly improves the accuracy of 3D face alignment. Model parameters and complexity of the proposed method are also reduced significantly.

Keywords: 3D face alignment · 3D Morphable Model · Computer vision

1 Introduction

Face alignment, which fits a face model to an image and extracts the semantic meanings of facial pixels. Traditional face alignment is to locate the feature points of human face. Such as corners of the eyes, corners of the mouth, tip of the nose, etc. This is a fundamental processing process for many computer vision tasks, e.g., face recognition [3], facial expression analysis [2], facial animation [6,7] and

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-34110-7_23) contains supplementary material, which is available to authorized users.

Y. Zhao et al. (Eds.): ICIG 2019, LNCS 11902, pp. 266–277, 2019.
https://doi.org/10.1007/978-3-030-34110-7_23

so on. In view of the importance of this problem, face alignment has been widely studied since the Active Shape Model (ASM) of Cootes in the early 1990s [10].

Despite the continuous improvement on the alignment accuracy, face alignment is still a very challenging problem. Traditional 2D face alignment can achieve satisfactory accuracy in small to medium poses, but this does not meet the changing conditions in real-world applications, non-frontal images, low image resolution, variable illumination and occlusion, etc. 3D face alignment aims to reconstruct 3D face structure through 2D image and estimated the position of 3D and 2D face feature points after 3D face alignment to 2D image.

Motivated by the needs to address the efficient model, pose variation, and the lack of prior work in handling poses, the paper proposes a novel and efficient network structure, and uses different loss functions to optimize the 3D parameters and 3D vertices. The purpose is to calculate the positions of 2D and 3D facial feature points under arbitrary postures. The reason for the efficiency of MobileNet [18] is that the Depthwise Separable Convolution is used in the network structure. Because of the Densely Connected between convolutional layers, DenseNet [19] strengthened the transmission of feature, made more effective use of feature and reduced the number of parameters to a certain extent. Inspired by the above two network structures, our network structure has high efficiency of both Depthwise Separable Convolution and feature reuse of Densely Connected. To achieve a balance between high efficiency and high precision. Finally, extensive experiments are conducted on a large subset of AFLW dataset [23] with a wide range of poses, and the AFLW2000-3D dataset [35] with the comparison with a number of methods. An overview of our method is shown in Fig. 1.

In summary, our contributions are summarized as follows:

(1). *We proposes a novel and efficient network structure (MobDenseNet). To the best of our knowledge, this is the first that Depthwise Separable Convolution and Densely Connected are combined in a network leading to a new structure of DNN.*

(2). *Different loss functions are used to optimize the parameters of 3D Morphable Model and 3D vertices. Meanwhile, face alignment that can estimate 2D/3D landmarks with an arbitrary pose.*

(3). *We experimentally verified that our algorithm has significantly improved performance of 3D face alignment compared to the previous algorithms, The proposed face alignment method can deal with arbitrary pose and it is more efficient.*

2 Related Work

In this section, we will review the prior work in generic face alignment and 3D face alignment.

2.1 Generic Face Alignment

Face alignment has achieved many achievements, including the classic AAM [9,26] and ASM [8] models. This method considers face alignment as an optimization problem to find the best shape and appearance parameters, which make

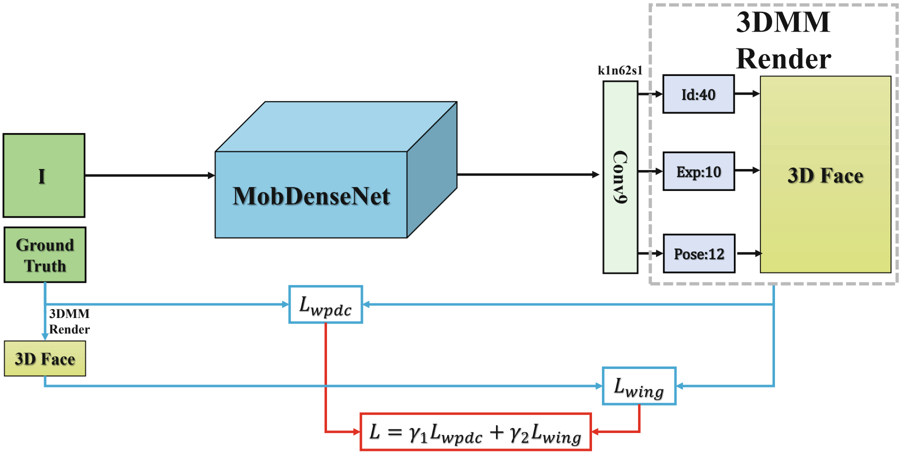

Fig. 1. Overview of the ours method. Efficient full convolutional neural networks (MobDenseNet). Figure 2 describes the details of MobDenseNet. The 3D parameters and 3D vertices are constrained using different loss functions.

the appearance model best fit the input face. The basic idea of Constrained Local Model (CLM) [1,11,27] in Discriminative methods is to learn a set of local appearance models, one for each landmark, and the decision from the local models are combined with a global shape model. Cascaded regression gradually refines a specified initial prediction value through a series of regressions. Each regression unit relies on the output of the previous regression unit to perform simple image operations, and the entire system can automatically learn from the training samples [12]. The ESR [7] (Explicit Shape Regression) proposed by Sun et al. includes three methods, namely two-level boosted regression, shape-indexed features and correlation-based feature selection method.

Besides traditional models, deep convolutional neural networks have recently been used for feature point localization of faces. Sun et al. [28] firstly use CNN to regress landmark locations with the raw face image, accurately positioning of 5 key points of faces from coarse to fine. The work of [16] using the human body pose estimation, the boundary information is introduced into the key point regression. In recent years, most of the landmark detections of faces have been studied on "coarse to fine", while Feng et al. [14] have taken a different approach, using the idea of cascaded convolutional neural networks. And [14] compared the commonly used loss functions in face landmark detection, and based on this, the concept of wing loss is proposed.

2.2 3D Face Alignment

Although the traditional method has achieved many achievements in face alignment, it will be affected by non-frontal face, illumination and occlusion in real-life applications. The most common method is the multi-view framework [29], which

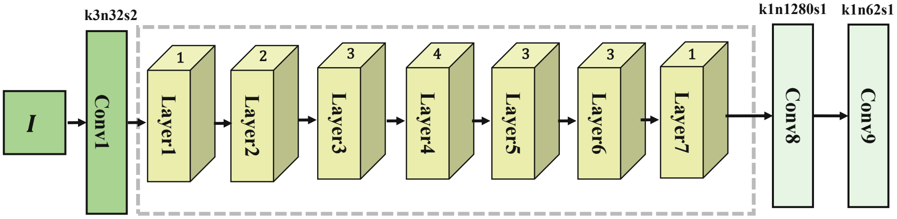

Fig. 2. Details of MobDenseNet. k3n64s1 corresponds to the kernel size(k), number of feature maps(n) and stride(s) of conv1.

uses different landmark configurations for different views. For example, TSPM [34] and CDM [33] use the DPM-like [15] method to align faces of different shape models, and finally select the most probable model as the final result. However, since each view requires testing, the computational cost of the multiview approach is always high.

In addition to multi-view solutions, 3D face alignment is a more common approach. 3D face alignment [16,20], which aims to fit a 3D morphable model (3DMM) [3] from a 2D image. The 3D Morphable Model is a typical statistical 3D face model. It has a clear understanding of the prior knowledge of 3D faces through statistical analysis. Zhu et al. [35] proposed a localization method based on 3D face shape, which solved the problem that some feature points were invisible under extreme postures (such as side faces). Liu et al. [21] used the cascade of 6 convolutional neural networks to solve the problem of locating facial feature points in a large pose by using 3D face modeling. This paper [13] designed a UV position map to achieve 3D shape features of a complete human face in a 2D UV space.

Our approach is also based on convolutional neural networks, but we have redesigned the network structure to make it efficient and robust. At the same time, we use different loss functions for 3D parameters and 3D vertices to constrain the semantic information of 3D parameters and 3D vertices respectively.

3 Proposed Method

In this section we introduce the proposed robust 3D face alignment (R3FA) which fits 3D morphable model with efficient fully convolutional neural networks.

3.1 3D Morphable Model

The 3D Morphable model is one of the most successful methods for describing 3D face space. Blanz et al. [3] proposed a 3D morphable model (3DMM) of 3D face space with PCA. It is expressed as follows:

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

where S is a specific 3D face, $\overline{S}$ is the mean face, A_{id} is the principle axes trained on the 3D face scans with neutral expression and α_{id} is the shape parameter, A_{exp} is the principle axes trained on the offsets between expression scans and neutral scans and α_{exp} is the expression parameter. So the coefficient $\{\alpha_{id}, \alpha_{exp}\}$ defines a unique 3D face . In this work A_{id} comes from the BFM [24] model and A_{exp} comes from the FaceWarehouse model [5].

In the process of 3DMM fitting, we use the Weak Perspective Projection to project 3DMM onto the 2D face plane. This process can be expressed as follows:

$$S_{2d} = f * Pr * R * \{S + t_{3d}\} \tag{2}$$

where S_{2d} is the 2D coordinate matrix of the 3D face after Weak Perspective Projection, rotation and translation. f is the scaling factor. Pr is a perspective projection matrix $\begin{pmatrix} 1\,0\,0 \\ 0\,1\,0 \end{pmatrix}$. R is a rotation matrix constructed according to three rotation angles of pitch, yaw and roll respectively. t_{3d} is the translation transformation matrix of 3D points. Therefore, for the modeling of a specific face, we only need to solve the 3D parameter $P = [f, pitch, yaw, roll, t_{3d}, \alpha_{id}, \alpha_{exp}]$.

3.2 MobDenseNet Structure

The reason MobileNet [18] is effective is the use of Depthwise Separable Convolution technology in the network structure. Based on MobileNetV1 [18], the design of MobileNetV2 [25] combines with the recent popular residual ideas. But the idea of residuals is achieved by the direct addition of elements. [19] a phenomenon that many layers of the ResNet [17] network, the first performer of residual thinking, contribute less and can be randomly discarded during training. This shows that residual ideas are prone to redundant information. In order to solve this problem, DenseNet [19] proposes any layer of the network, the feature map of all the layers in front of the layer is the input of this layer. The feature map of the layer is the input of all the layers behind. However, DenseNet has many parameters and the network structure is not efficient. So combining with MobileNet's efficiency and DenseNet's feature enhancement, we build a new network structure MobDenseNet by combining DenseNet's dense connections on the overall framework of MobileNet. Our network structure includes both MobileNet's high efficiency and enhance feature representation.

The architecture of MobDenseNet is illustrated in Fig. 2. MobDenseNet is a fully convolutional neural network without full connection layer. Conv1 is a convolution layer with kernel size(k) of 3, stride(s) of 2 and number of feature maps(n) of 32 to extract rough features. $Layer1$ to $Layer7$ are 7 dense blocks for extracting depth features. Figure 3 shows the details of one of the DenseBlock, $Layer3$. The convolution layer of a set of $1 \times 1, 3 \times 3, 1 \times 1$ filters in MobDenseNet as a basic unit called MobileBlock. As shown in Fig. 3, this set of basic units is consistent with MobileNetV2. DenseLayer3 contains three sets of Mobile Blocks (each MobileBlock output is cascaded as the input of the next MobileBlock). As such, MobDenseNet retains the simplicity and efficiency of MobileNet.

As shown in Fig. 3, Layer3 contains three sets of MobileBlocks. In order to match the number of channels connected to the Dense connection, we added a transition layer after each MobileBlock (the convolution layer filter is 1 × 1), the purpose is adjust the number of channels in the preview MobileBlock output feature map. We use both real face images and generated face images to train our MobDenseNet (details can be found in the suppl. material).

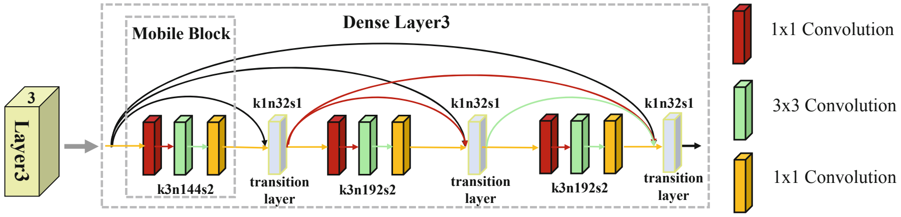

Fig. 3. The details of one of the DenseBlock, *Layer*3. The convolution layer of a set of $1 \times 1, 3 \times 3, 1 \times 1$ filters in MobDenseNet as a basic unit called MobileBlock. The transition layer is the number of channels to match the input and output feature maps.

3.3 Loss Function

We chose two different Loss Functions to jointly train MobDenseNet. For 3D parameters and 3D vertices we use different loss functions for training. We follow the Weighted Parameter Distance Cost (WPDC) of Zhu et al. [35] to calculate the difference between the ground truth of 3D parameters and the predicted 3D parameters. The basic idea is explicitly modeling the importance of each parameter:

$$L_{wpdc} = (P_{gt} - \overline{P})^T W (P_{gt} - \overline{P}) \tag{3}$$

where $\overline{P}$ is the estimation and P_{gt} is the ground truth. The diagonal matrix W contains the weights. For each element of the shape parameter p, its weight is the inverse of the standard deviation that was obtained from the data used in 3DMM training. Because our ultimate goal is to accurately obtain 68 landmarks of human faces. So for 3D face vertices reconstructed with 3D parameters, we use Wing Loss [14] which is defined as:

$$L_{wing}(\varDelta V(P)) = \begin{cases} \omega \ln(1 + |\varDelta V(P)|/ \in) & if\ |\varDelta V(P)| < \omega \\ |\varDelta V(P)| - C & otherwise \end{cases} \tag{4}$$

where $\varDelta V(P) = V(P_{gt}) - V(\overline{P})$,$V(P_{gt})$ and $V(\overline{P})$ are the ground truth of the 3D facial vertices and the 3D facial vertices reconstructed using the 3D parameters predicted by the network, respectively. ω and $\in$ are parameters. $C = \omega - \omega \ln(1 + \omega/ \in)$ is a constant that smoothly links the piecewise-defined linear and nonlinear parts.

Overall, the framework is optimized by the following loss function:

$$L_{loss} = \lambda_1 L_{wpdc} + \lambda_2 L_{wing} \tag{5}$$

where λ_1 and λ_2 are parameters, which balance the contribution of L_{wpdc} and L_{wing}. The selection of those parameters will be discussed in the next section.

4 Experiments

In this section, we evaluate the performance of R3FA on three common face alignment tasks, face alignment in small and medium poses, face alignment in large poses, and face reconstruction in extreme poses (±90° yaw angles), respectively.

4.1 Implementation Details

We use the Pytorch deep learning framework to train the MobDenseNet models. The loss weights of R3FA are empirically set to $\lambda_1 = 0.5$ and $\lambda_2 = 1$. In our experiments, we set the parameters of the Wing loss as $\omega = 10$ and $\in = 2$. The Adam solver [22] is employed with the mini-batch size and the initial learning rate set to 128 and 0.01, respectively. There are 680,000 face images in our training data set, including 430,000 real face images and 250,000 synthetic face images. Real face images come from 300W-LP [35] datasets, and various data enhancement algorithms are adopted to expand the datasets. We run the training for a total of 40 epochs. After 15, 25 and 30 epochs, we reduced the learning rate to 0.002, 0.0004 and 0.00008 respectively.

4.2 Evaluation Databases

We evaluate the performance of R3FA on two publicly available face data sets AFLW [23] and AFLW2000-3D [35]. These two data sets contain small and medium poses, large poses and extreme poses (±90° yaw angles). We divide the dataset AFLW and AFLW2000-3D into three intervals of $[0^\circ, 30^\circ]$, $[30^\circ, 60^\circ]$, and $[60^\circ, 90^\circ]$ according to the face absolute yaw angle, and each interval is about 1/3 of the total.

AFLW. AFLW face database is a large-scale face database including multi-pose and multi-view, and each face is marked with 21 feature points. This database has a very large amount of information, including pictures of various poses, expressions, lighting, and ethnicity. The AFLW face database consists of approximately 250 million hand-labeled face images, of which 59% are women and 41% are men. Most of the images are color, images only a few are gray images. We only use part of the extreme pose face images of the AFLW database for qualitative analysis.

AFLW2000-3D. AFLW2000-3D is constructed by [35] to evaluate 3D face alignment on challenging unconstrained images. This database contains the first 2000 images from AFLW and expands its annotations with fitted 3DMM parameters and 68 3D landmarks. We use this database to evaluate the performance of our method on face alignment tasks.

4.3 Evaluation Metric

Given the ground truth 2D landmarks U_i, their visibility v_i, and estimated landmarks $\hat{U_i}$ of N_t testing images. Normalized Mean Error (NME), which is the average of the normalized estimation error of visible landmarks, i.e.,

$$NME = \frac{1}{N_t}\sum_{i}^{N_t}(\frac{1}{d_i|v_i|_1}\sum_{j}^{N} v_i(j)||\hat{U_i}(:,j) - U_i(:,j)||) \tag{6}$$

where d_i is the square root of the face bounding box size, as used by [37]. Note that normally d_i is the distance of two centers of eyes in most prior face alignment work dealing with near-frontal face images.

4.4 Comparison Experiments

Comparison on AFLW. In the AFLW dataset, 21,080 images were selected as test samples, with 21 landmarks in each sample. During testing, we divide the testing set into 3 subsets according to their absolute yaw angles: $[0°, 30°], [30°, 60°]$ and $[60°, 90°]$ with 11,596, 5,457 and 4,027 samples respectively. Since few experiment has been conducted on AFLW, we choose some baseline methods with released codes, including CDM [33], RCPR [4], ESR [7], SDM [32], 3DDFA [35] and nonlinear 3DMM [30]. Table 1 demonstrates the comparison results. The NME(%) of face alignment results on AFLW with the first and the second best results highlighted. The results of provided alignment models are marked with their references. Figure 4 shows the corresponding CED curves. Our CED curve is only compared to the best method in Table 1. The results show that our R3FA algorithm significantly improves the face alignment accuracy in full pose. The minimum standard deviation of R3FA also proves its robustness to posture changes.

Table 1. The NME(%) of face alignment results on AFLW and AFLW2000-3D.

	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
CDM	8.150	13.020	16.170	12.440	4.040	-	-	-	-	-
RCPR	5.430	6.580	11.530	7.850	3.240	4.260	5.960	13.180	7.800	4.740
ESR	5.660	7.120	11.940	8.240	3.290	4.600	6.700	12.670	7.990	4.190
SDM	4.750	5.550	9.340	6.550	2.450	3.670	4.940	9.760	6.120	3.210
3DDFA(CVPR16)	5.000	**5.060**	6.740	5.600	0.990	3.780	4.540	7.930	5.420	2.210
Nonlinear 3DMM(CVPR18)	-	-	-	-	-	-	-	-	4.700	-
Ours-R3FA	**4.549**	5.427	**6.204**	**5.393**	**0.676**	**3.149**	**4.010**	**5.270**	**4.143**	**0.871**

Table 2. The NME(%) of face alignment results on AFLW and AFLW2000-3D with the different network structures.

	Extracting Params Time(ms/pic)		Params	AFLW DataSet(21 pts)					AFLW2000-3D DataSet(68 pts)				
Method	AFLW(21 pts)	AFLW2000-3D(68 pts)		$[0^o, 30^o]$	$[30^o, 60^o]$	$[60^o, 90^o]$	Mean	Std	$[0^o, 30^o]$	$[30^o, 60^o]$	$[60^o, 90^o]$	Mean	Std
RestNeXt50	0.799ms	2.012ms	90.585M	4.599	5.516	6.297	5.471	0.694	3.122	4.065	5.351	4.179	0.913
MobileNetV2	**0.316ms**	**0.956ms**	**9.487M**	4.643	5.581	6.397	5.540	0.716	3.236	4.080	**5.181**	4.165	0.796
DenseNet121	0.684ms	2.221ms	27.9M	**4.442**	**5.249**	**6.168**	**5.286**	0.705	**3.051**	**3.912**	5.297	**4.087**	0.925
MobDenseNet	0.395ms	1.024ms	10.900M	4.549	5.427	6.204	5.393	0.676	3.149	4.010	5.27	4.143	0.871

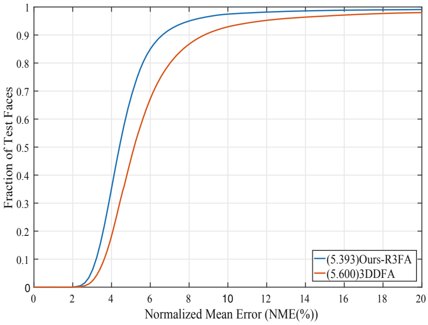

Fig. 4. Comparisons of cumulative errors distribution (CED) curves on AFLW.

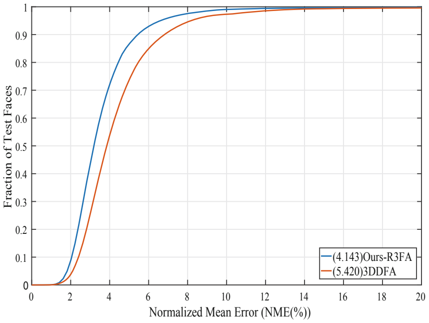

Fig. 5. Comparisons of cumulative errors distribution (CED) curves on AFLW2000-3D.

Comparison on AFLW2000-3D. In the AFLW2000-3D dataset, 2000 images were selected as test samples, with 68 landmarks in each sample. Considering the visible and invisible evaluation, 3D face alignment evaluation can be downgraded to a full landmark evaluation. we divide the testing set into 3 subsets according to their absolute yaw angles: $[0°, 30°]$, $[30°, 60°]$, $[60°, 90°]$ with 1,312, 383 and 305 samples respectively. Table 1 demonstrates the comparison results. The NME(%) of face alignment results AFLW2000-3D with the first and the second best results highlighted. The results of provided alignment models are marked with their references. Figure 5 shows the corresponding CED curves. Our CED curve is only compared to the best method in Table 1. Table1 and Fig. 5 demonstrate that our algorithm also has a significant improvement in the prediction of invisible regions, showing good robustness for face alignment in arbitrary poses.

Comparison on Different Network Structures. We selected a variety of different network structures for comparison during the experiment. The experimental network structure includes ResNeXt [31], MobileNetV2 [25], DenseNet121 [19], and our proposed MobDenseNet. To the best of our knowledge, these three popular and efficient network structures are the first to be used in the field of 3D face alignment. Table 2 demonstrates the comparison results. The NME(%) of face alignment results on AFLW and AFLW2000-3D with the different network structures. The table shows the time when each sample extracts parameters through the network model and the parameter size of the network model. Extracting params time (ms/pic) is calculated on GTX 1080Ti and 64 GB RAM. These three network structures can be divided into two categories, one is the efficient network structure represented by MobileNetV2, and the other is the high-precision network structure of ResNeXt50 and DenseNet121. In order to balance efficiency and high precision, we have designed MobDenseNet independently. The experimental results demonstrate the motivation and expected results of our original design. Our network structure achieves a balance between high efficiency and high precision. Comparison and analysis with MobileNetV2 and DesenNet can be found in suppl. material. The 2D/3D alignment results of our method are shown in Fig. 6.

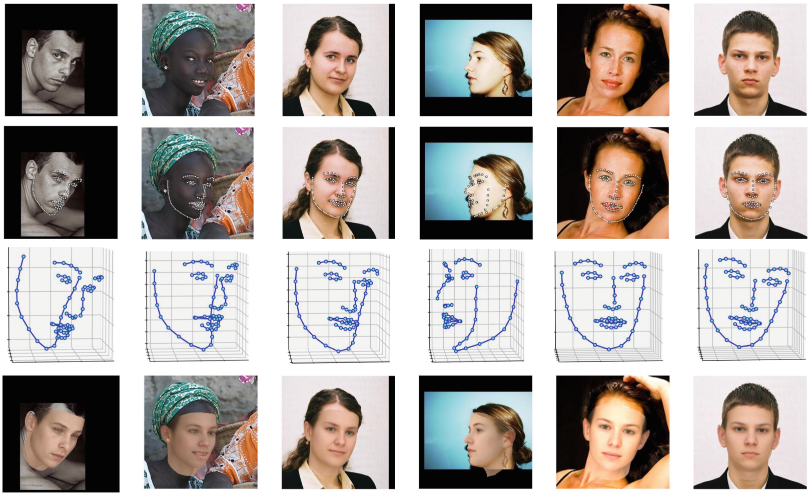

Fig. 6. The results of 2D/3D face alignment of our method. Result of 2D face alignment (second rows), 3D face alignment (third rows), Align 3D face mesh to 2D image (fourth rows).

5 Conclusions

In this paper, we propose a novel and efficient framework (R3FA), which solves the problem of 2D/3D face alignment with full pose. In order to balance the computational efficiency and alignment accuracy of the model, we propose a new deep network MobDenseNet. We innovatively use two loss functions to jointly optimize 3D reconstruction parameters and 3D vertices. At the same time, we use real and synthetic images to train our network together. We have achieved the best accuracy on both AFLW and AFLW2000-3D datasets compared to existing algorithms. Comparing experiments with several popular networks, our algorithm can achieve a good balance between accuracy and efficiency. In the future, we will further improve the accuracy of 2D/3D face alignment, and at the same time the algorithm will have higher efficiency.

Acknowledgments. The paper is supported by the National Natural Science Foundation of China (Grant No. 61672265,U1836218), the 111 Project of Ministry of Education of China (Grant No. B12018), and UK EPSRC Grant EP/N007743/1, Muri/EPSRC/ Dstl Grant EP/R018456/1,

References

1. Asthana, A., Zafeiriou, S., Cheng, S., Pantic, M.: Robust discriminative response map fitting with constrained local models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3444–3451 (2013)
2. Bettadapura, V.: Face expression recognition and analysis: the state of the art. arXiv preprint arXiv:1203.6722 (2012)
3. Blanz, V., Vetter, T.: Face recognition based on fitting a 3D morphable model. IEEE Trans. Pattern Anal. Mach. Intell. **25**(9), 1063–1074 (2003)
4. Burgos-Artizzu, X.P., Perona, P., Dollár, P.: Robust face landmark estimation under occlusion. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1513–1520 (2013)
5. Cao, C., Weng, Y., Zhou, S., Tong, Y., Zhou, K.: Facewarehouse: a 3D facial expression database for visual computing. IEEE Trans. Vis. Comput. Graph. **20**(3), 413–425 (2014)
6. Cao, C., Wu, H., Weng, Y., Shao, T., Zhou, K.: Real-time facial animation with image-based dynamic avatars. ACM Trans. Graph. **35**(4) (2016)
7. Cao, X., Wei, Y., Wen, F., Sun, J.: Face alignment by explicit shape regression. Int. J. Comput. Vis. **107**(2), 177–190 (2014)
8. Cootes, T., Baldock, E.R., Graham, J.: An introduction to active shape models. Image Process. Anal. 223–248 (2000)
9. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. IEEE Trans. Pattern Anal. Mach. Intell. **6**, 681–685 (2001)
10. Cootes, T.F., Taylor, C.J., Lanitis, A.: Active shape models: evaluation of a multiresolution method for improving image search. In: BMVC, vol. 1, pp. 327–336 (1994)
11. Cristinacce, D., Cootes, T.F.: Feature detection and tracking with constrained local models. In: BMVC, vol. 1, p. 3 (2006)
12. Dollár, P., Welinder, P., Perona, P.: Cascaded pose regression. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1078–1085. IEEE (2010)
13. Feng, Y., Wu, F., Shao, X., Wang, Y., Zhou, X.: Joint 3D face reconstruction and dense alignment with position map regression network. arXiv preprint arXiv:1803.07835 (2018)
14. Feng, Z.-H., Kittler, J., Awais, M., Huber, P., Wu, X.-J.: Wing loss for robust facial landmark localisation with convolutional neural networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2235–2245. IEEE (2018)
15. Forsyth, D.: Object detection with discriminatively trained part-based models. Computer **2**, 6–7 (2014)
16. Gu, L., Kanade, T.: 3D alignment of face in a single image. In: Null, pp. 1305–1312. IEEE (2006)
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)

18. Howard, A.G., et al.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
19. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: CVPR, vol. 1, p. 3 (2017)
20. Jourabloo, A., Liu, X.: Pose-invariant 3D face alignment. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3694–3702 (2015)
21. Jourabloo, A., Liu, X.: Large-pose face alignment via CNN-based dense 3D model fitting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4188–4196 (2016)
22. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
23. Koestinger, M., Wohlhart, P., Roth, P.M., Bischof, H.: Annotated facial landmarks in the wild: a large-scale, real-world database for facial landmark localization. In: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pp. 2144–2151. IEEE (2011)
24. Paysan, P., Knothe, R., Amberg, B., Romdhani, S., Vetter, T.: A 3D face model for pose and illumination invariant face recognition. In: Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance. AVSS 2009, pp. 296–301. IEEE (2009)
25. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation. arXiv preprint arXiv:1801.04381 (2018)
26. Saragih, J., Goecke, R.: A nonlinear discriminative approach to AAM fitting. In: IEEE 11th International Conference on Computer Vision. ICCV 2007, pp. 1–8. IEEE (2007)
27. Saragih, J.M., Lucey, S., Cohn, J.F.: Deformable model fitting by regularized landmark mean-shift. Int. J. Comput. Vis. **91**(2), 200–215 (2011)
28. Sun, Y., Wang, X., Tang, X.: Deep convolutional network cascade for facial point detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3476–3483 (2013)
29. Tran, A.T., Hassner, T., Masi, I., Medioni, G.: Regressing robust and discriminative 3D morphable models with a very deep neural network. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1493–1502. IEEE (2017)
30. Tran, L., Liu, X.: Nonlinear 3D face morphable model. arXiv preprint arXiv:1804.03786 (2018)
31. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5987–5995. IEEE (2017)
32. Yan, J., Lei, Z., Yi, D., Li, S.: Learn to combine multiple hypotheses for accurate face alignment. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 392–396 (2013)
33. Yu, X., Huang, J., Zhang, S., Metaxas, D.N.: Face landmark fitting via optimized part mixtures and cascaded deformable model. IEEE Trans. Pattern Anal. Mach. Intell. **11**, 2212–2226 (2016)
34. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2879–2886. IEEE (2012)
35. Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S.Z.: Face alignment across large poses: a 3D solution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 146–155 (2016)