Face Alignment across Large Pose via MT-CNN based 3D Shape Reconstruction

Gang Zhang[1,2], Hu Han[1], Shiguang Shan[1,2,3], Xingguang Song[4], Xilin Chen[1,2]
[1]*Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS),*
Institute of Computing Technology, CAS, Beijing 100190, China
[2]*University of Chinese Academy of Sciences, Beijing 100049, China*
[3]*CAS Center for Excellence in Brain Science and Intelligence Technology*
[4]*Huawei Technologies Co., Ltd*
gang.zhang@vipl.ict.ac.cn; {hanhu, sgshan, xlchen@ict}@ict.ac.cn, songxingguang@huawei.com

***Abstract*—Face alignment plays an important role for robust face recognition and analysis applications in the wild. While a number of face alignment methods are available, large-pose face alignment remains a very challenging problem due to the ambiguity of facial keypoints in 2D face images. Recent attempts to solve this problem via 3D model fitting show more robustness against large poses and 2D ambiguity, but their accuracy and speed are still limited. We propose a 3D reconstruction based method to quickly and accurately detect 2D facial landmarks and estimate their visibility. By designing a cascaded multi-task CNN model, we can efficiently reconstruct the 3D face shape, together with pose estimation as an auxiliary task. Finally, the landmarks on 3D shape are projected to the 2D face image to get the 2D landmarks and their visibility. Experimental results on the challenging 300W-LP, AFLW2000-3D, and AFLW databases show that the proposed approach can be comparable with the state-of-the-art methods and is able to run in real time (32ms per image) on 3.4 GHz CPU.**

***Keywords*-Landmark detection; 3D shape regression; multi-task CNN; large-pose face alignment;**

I. Introduction

Face alignment has attracted increasing attentions in recent years due to its widespread applications in face recognition [20], facial expression [3], and 3D face reconstruction [4]. While face alignment under frontal or near-frontal poses have been well addressed, face alignment under unconstrained scenarios remains a challenging problem due to variations of bad illumination, partial occlusion, and image blurring.

Early work on face alignment used an analysis-by-synthesis approach [28], [7]. For example, the Active Appearance Model (AAM) [7], detects facial landmarks by minimizing the difference between the synthesized face appearance and input face image. However, these methods are usually linear models, and thus are difficult to handle face alignment under complicated scenarios, such as pose, expression, and occlusion.

Regression-based methods are another widely used approach for face alignment, which use various facial appearance features, and learn a mapping from the facial appearance to the landmark positions [9], [30], [31]. The Cascaded Pose Regression (CPR) [9], which used pose-index features and cascaded random forest regressors to detect the keypoints of generic objects. However, CPR is not robust to occlusions and pose variations. The aforementioned methods used hand-craft features, which rely on expert knowledge, and often assume a linear model between the face appearance space and the landmark space. In order to model the possibly non-liner mapping between facial appearance and landmark positions. Deep convolutional neural networks (DCNNs) were then used to regress the coordinates of the facial landmarks from 2D face images [25], [31], [31]. For example, DCNN was proposed in [25] to directly regress the facial landmarks from 2D images, and achieved promising results on LFPW [2] dataset. However, even with CNN models, face alignment across large poses remains a challenging problem, because of the self-occlusions and landmark ambiguity in 2D face images.

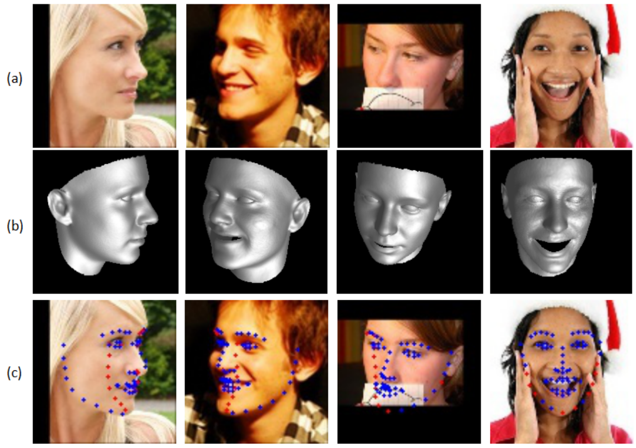

Figure 1. Facial landmarks detection from a single unconstrained face image. (a) the input face images with large variations of pose and expression, and partial occlusions; (b) the reconstructed 3D face shapes by the proposed multi-task cascaded CNN, which recovers the geometric details of the 2D face image, such as wrinkles and nasolabial folds; (c) the predicted facial landmarks on the 2D face images with red and blue points denoting the invisible and visible 2D landmarks respectively.

More recently, 3D based methods have been proposed in

H. Han is the corresponding author.

978-1-5386-2335-0/18/$31.00 ©2018 IEEE

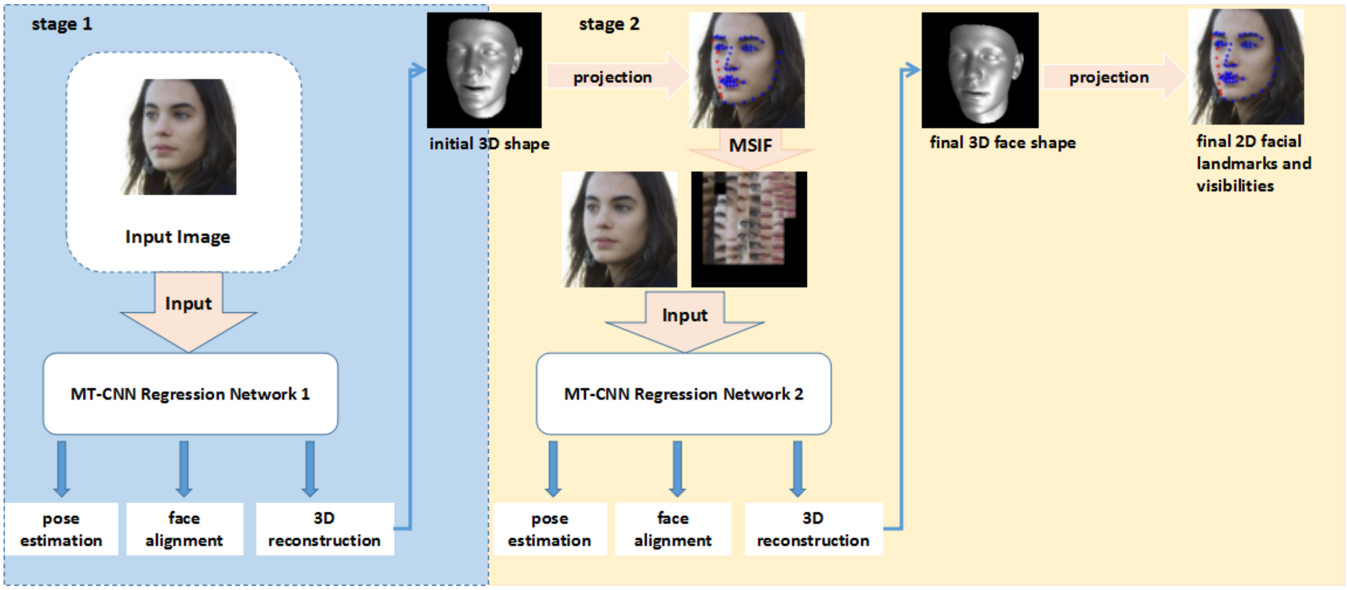

Figure 2. Overview of the proposed face alignment approach across large poses using CNN-based 3D shape reconstruction. In stage 1, multi-task CNN (MT-CNN) regression takes the whole face images as input and jointly estimate the 3D face shape, 2D landmarks and face poses. The initial 2D landmarks obtained in stage 1 are then used to extract Modified Shape Indexed Features (MSIFs) for MT-CNN regression of the 3D shape in stage 2.

order to cope with 2D keypoint ambiguity. For example, a cascaded CNN regressors was designed in [33], [15] to regress the PCA coefficients of the 3D morphable model, as well as the camera parameters. 2D landmarks were then obtained by projecting the 3D landmarks into face images. By incorporating 3D information, the 2D keypoint ambiguity and self-occlusions caused by 3D transformations can be inherently addressed. However, the current 3D based face alignment approaches are still facing several issues: (i) These 3DMM PCA coefficients cannot be treated equally, because each component of 3DMM have different influence on the resulted 3D face shape; (ii) The Projected Normalized Coordinate Code (PNCC) in [33] and Direct 3D projected feature (D3PF) in [15] are complicated. That is why they cannot run in real time.

In this paper, as shown in Figure 2, we propose a cascaded multi-task CNN (MT-CNN) to jointly regress the 3D face shape as well as the face poses. In each stage of our cascaded CNN, we first estimate the 3D keypoints, and then use a fully connected layer to predict the whole (dense) 3D face shape. Different from [33], [15], our method directly regresses the 3D vertex coordinates (x,y,z) instead of the PCA coefficients of 3DMM [4] and uses Modified Shape Indexed Features (MSIFs) as the input for second stage MT-CNN regression. Experiments on the 300W-LP, AFLW2000-3D and AFLW databases demonstrate the facial landmark detection accuracy and the computational efficiency (four times faster than [33]) of the proposed approach.

While the proposed method is also a cascaded CNN regressors, it differs from the existing 3D based face alignment methods [33], [15] in several important aspects: (i) The proposed method directly regresses the 3D vertex coordinates (x,y,z) instead of the PCA coefficients of 3DMM; (ii) The proposed method uses Modified Shape Indexed Features (MSIFs) as the features for the second-stage CNN regression, which is much faster than PNCC and D3PF; (iii) The proposed method jointly handles multiple tasks, which consists of shared feature learning and task-dependent feature optimization.

The rest of this paper is organized as follows: Section II briefly summarizes the related works on face alignment and multi-task learning. Section III details the proposed multi-task cascaded CNN for 3D face reconstruction and face alignment. Section IV presents the experimental results, and comparisons with the state-of-the-art methods on three different tasks. Section V makes a conclusion of this work.

II. Related Work

This work is related to face alignment and multi-task learning. In this section, we briefly review the most recent literatures on these two topics.

A. *Face Alignment*

Face alignment or face landmark detection aims to detect the facial keypoints like mouth corners, eye corners. Published methods for face alignment can be grouped into two main categories: synthesis-based method and regression-based method.

Synthesis-based face landmark detection methods, such as the Active Shape Model (ASM) [28] and the Active Appearance Model (AAM) [7], used Principal Component

Analysis (PCA) to represent the face appearance and minimize the difference between the synthesized face appearance and input face image. However, they are linear models, which can hardly handle non-linear scenarios, such as pose variations, expression variations and occlusion.

Regression-based face landmark detection methods, especially cascaded regression methods, directly regress the coordinates of the facial keypoints and are able to better cope with pose variations, expression variations and occlusion problems, compared with the synthesis-based methods. In [5], the researchers proposed the Robust Cascaded Pose Regression (RCPR) to explicitly handle occlusion scenarios by predicting the occlusion conditions of each keypoint. The Supervised Descent Method (SDM) [30], which is widely used in real-time face alignment, uses cascaded linear regressors to localize facial keypoints by using SIFT features around the facial landmarks detected by the previous stages. In order to improve the robustness to non-linear scenarios, such as pose variations, expression variations and occlusion, Zhang et al. [31] proposed a Coarse-to-Fine Auto-Encoder Networks (CFAN) to detect facial keypoints stage by stage and achieved promising results on LFPW. However, these 2D methods assume that all the landmarks are visible, which is not suitable for large pose scenarios. In [33], [15], 3D methods, which outperform the aforementioned 2D methods, especially under large pose scenarios, have been proposed. They both employed cascaded CNNs to regress the PCA coefficients of 3D face shape and the camera parameters. The main difference between [33], [15], is that they employed different shape indexed features. While PNCC features were used in [33], and PAWF and D3PF features were exploited in [15].

B. Multi-task Learning

Multi-task learning (MTL) allows feature sharing among individual tasks, and is able to learn more robust feature representations [18], [29], [12]. MTL optimizes the feature learning of individual tasks, and makes a balance between the computational cost and accuracy. Deep learning is well suited for MTL, and many researchers combine deep learning with MTL to address different joint optimization problems. For example, facial landmark detection, pose estimation, and facial attribute prediction were jointly handled in [32] via MTL. ROI classification and bounding box regression were jointly solved in Faster R-CNN [10]. Face identification and verification were jointly solved in [24]. Face alignment and 3D face reconstruction were jointly solved in [18], leading to promising performance in both tasks. Face alignment, pose estimation, and 3D face reconstruction are highly correlated tasks. In this paper, we use MTL to solve the problem, together with auxiliary problems, such as pose estimation.

III. PROPOSED METHOD

In this section, we provide the details of our face alignment approach via CNN-based 3D shape reconstruction.

A. 3D Shape Reconstruction via MT-CNN

Following the assumptions in many published methods [4], [19], [11], a 3D face with frontal pose and arbitrary expression can be represented as:

$$S = S_0 + \sum_{i=1}^{m} a_i U_i + \sum_{j=1}^{n} b_j V_j, \tag{1}$$

where S_0 is the average 3D face shape, U_i is the i_{th} principal shape component, and a_i is the i_{th} shape parameter; V_j is the j_{th} principal expression component, and b_j is the j_{th} expression parameter. S denotes a frontal face with expression.

Based on Eq. (1), a recovered 3D face shape with the same pose as an input 2D face image can be represented as:

$$\hat{S} = s * R * S + t, \tag{2}$$

where R is a 3D rotation matrix composed of rotations pitch, yaw, and roll; s is a scale factor; t is a 2D translation parameter.

$$\hat{S} = \begin{pmatrix} x_1 & x_2 & \dots & x_N \\ y_1 & y_2 & \dots & y_N \\ z_1 & z_2 & \dots & z_N \end{pmatrix} \tag{3}$$

The proposed approach directly predict the 3D shape (vertices) $\hat{S}$ given a single 2D face image I

$$\hat{S} = f(I), \tag{4}$$

where f can be seen as an inverse projecting from I to $\hat{S}$. While most of the published 3D reconstruction methods assume a very simple orthogonal or perspective projection from $\hat{S}$ to I, the proposed approach allows the projection to be a complicated process to replicate the projections in real scenarios..

The 3D face shape $\hat{S}$ can be of very high dimension. For example, for a 3D shape with 20,000 vertices, the target regression vector can be 60,000 dimensions. This may lead to difficulty in convergence during network training. Therefore, as shown in Figure 3, face alignment is also adopted as an auxiliary task to tackle the convergence problem.

Cascaded CNNs perform a coarse-to-fine regression, aiming to refine the landmark positions stage by stage. In each stage, it takes the feedback feature as input, such as pose-index feature in [9], PNCC in [33], PAWF and D3PF features in [15], and MSIFs in the proposed method, to regress the deviations between the ground-truth values and the values predicted by previous stages. In k_{th} stage, it can achieve better performance compared with the previous $(k-1)$ stages. Motivated by the success of coarse-to-fine regression methods [9], [31], [33], [15], we also employ a similar

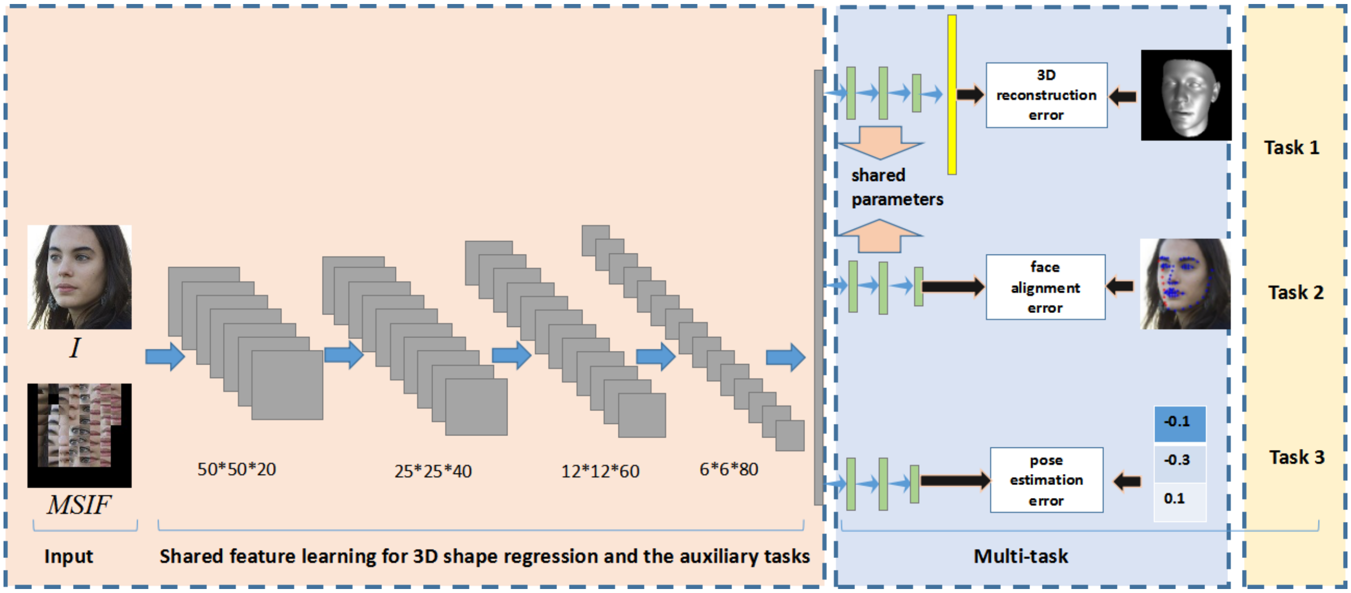

Figure 3. The architecture of the proposed multi-task CNN for directly regressing the 3D face shape vertices, as well as 2D face landmarks and pose estimation, consists of shared feature learning and task-dependent feature optimization. Such a architecture can be organized in a cascaded scheme in order to obtain more accurate 3D shape reconstruction results.

framework for 3D face shape regression. At the k_{th} stage, 3D face shape S^k can be computed by

$$[\hat{S}^k, o^k] = [\hat{S}^{k-1}, o^{k-1}] + [\Delta\hat{S^k}, \Delta o^k], \tag{5}$$

$$[\Delta\hat{S^k}, \Delta o^k] = Net^k(I, \Phi(I, \hat{S}^{k-1}_{keypoint2D})), \tag{6}$$

where Net^k is the our MT-CNN regression network at the k_{th} stage; I is the input 2D face image; $\Phi(I, \hat{S}^{k-1}_{keypoint})$ is the Modified Shape Indexed Features (MSIFs), which will be described in the subsection III-C; S^k represents the reconstructed 3D face shape at the k_{th} stage; o^k is the output of other auxiliary tasks at the k_{th} stage, such as face pose estimation.

In our experiments, we found that increasing the number stages of our cascaded MT-CNN improves the 3D face reconstruction accuracy slowly, but also lead to high computational cost. Therefore, we choose to use two stages in order to achieve a good balance between face alignment accuracy and the computational cost.

B. Face Alignment on 2D Images

Because face alignment is an auxiliary task in the proposed framework just to tackle the convergence problem, the resulted face landmarks are derived from the 3D face shape. Face landmarks for face alignment are a subset of the whole 3D shape vertices and thus can be obtained by projecting the 3D face vertices onto 2D image.

$$\hat{S}^k_{keypoint2D} = P\hat{S}^k_{keypoint3D}, \tag{7}$$

where $\hat{S}^k_{keypoint2D}$ is the 2D face keypoints at the k_{th} stage; $\hat{S}^k_{keypoint3D}$ is the 3D face keypoints at the k_{th} stage; P is the projection matrix.

$$P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \tag{8}$$

In order to compute the visibility of each 2D landmark, we calculate the angle between 3D surface normal of a landmark and the optical axis. Specifically, given the reconstructed 3D face shape $\hat{S}^k$, we compute the 3D surface normals of the i_{th} 3D landmarks $\vec{N_i}$ and then its visibility v_i as

$$v_i = \vec{N_c} \cdot \vec{N_i}, \tag{9}$$

where $\vec{N_c}$ represents camera optical axis direction, and usually is assumed to be

$$\vec{N_c} = \begin{pmatrix} 0 & 0 & 1 \end{pmatrix} \tag{10}$$

C. Implementation Details

MT-CNN Regression Network. In each stage of the cascaded network, we use a multi-task CNN (MT-CNN) architecture. Figure 3 shows the detailed architecture of MT-CNN. Generally speaking, the MT-CNN consists of a shared feature learning for all the tasks, and task-dependent feature optimization for each task. Shared feature learning uses four convolutional layers and three max-pooling layers, followed by several sub-networks (each with a few fully connected (FC) layers) for task-dependent feature optimization. We also use PReLU [13] and dropout [23] in FC layers to improve the model's generalization ability.

The objective function of the MT-CNN network can be formulated as

$$\arg\min_{W} \sum_{t=1}^{T} L_t, \tag{11}$$

where T is the total number of tasks that is going to be solved jointly; L_t is the loss function for t_{th} task. In particular, for tasks like 3D face shape reconstruction, face alignment and pose estimation, we use a regression loss (Euclidean)

$$L_t^R = |\hat{V} - V|_{L2}, \tag{12}$$

where $\hat{V}$ and V denote the predicted and the ground-truth values, respectively.

The final 3D face shape estimation result can be calculated as

$$\hat{S} = \hat{S}_0 + \sum_{i=1}^{2} \Delta\hat{S}^i, \tag{13}$$

where $\hat{S}_0$ is the initial 3D face shape, and in our experiments it is initialized with all zeros.

Modified Shape Indexed Features. To extract more relevant features to the 3D shape regression task, we propose a new feature named Modified Shape Indexed Features (M-SIFs). Specifically, we project the 3D keypoints $\hat{S}^{k-1}_{keypoint3D}$ obtained in the $(k-1)_{th}$ stage onto the 2D face image, and extract image patches centered at the 2D keypoints. These image patches are used as the input for the next stage. We also estimate the visibility of each 2D keypoint based on the surface normal of its 3D vertex. An image patch will be filled with zeros if the the 2D keypoint is invisible. The process for extracting MSIFs is shown in Figure 4.

Parameter Update Policy. Since the MT-CNN of all the stages have the same architecture, the following parameter update policy applies to every stage.

The loss function of MT-CNN can be formulated as

$$Loss = \sum_{i=1}^{T} L_i, \tag{14}$$

Let w_{shared} denote the parameters of the shared convolutional layers, w_i denote the parameters of the fully-connected layers only related to the i_{th} auxiliary task, and γ represent the learning rate, then we have

$$\omega^{n+1}_{shared} = \omega^{n}_{shared} - \gamma \frac{\partial Loss}{\partial \omega_{shared}}|_{\omega_{shared}=\omega^{n}_{shared}}, \tag{15}$$

$$\omega^{n+1}_{t} = \omega^{n}_{t} - \gamma \frac{\partial L_i}{\partial w_i}|_{w_i=w^{n}_{i}}, \tag{16}$$

We perform Adam [16] to optimize the network weights. We use an initial learning rate of 0.001, and a batch size of 400 in our experiments.

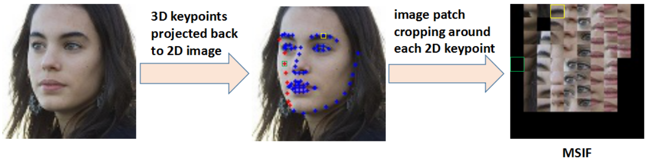
3D keypoints projected back to 2D image
image patch cropping around each 2D keypoint
MSIF

Figure 4. The predicted 3D keypoints on 3D face are first projected back to the 2D input image, and patches centered at each 2D keypoint are concatenated to form the Modified Shape Indexed Features (MSIFs). The MSIFs are used as input of the next stage.

Table I
COMPARISONS OF 2D FACE ALIGNMENT NME (IN %) BETWEEN THE PROPOSED APPROACH AND THE STATE-OF-THE-ART METHODS ON THE AFLW DATABASE.

Method	AFLW Database (21 pts)			
	[0,30]	[30,60]	[60,90]	Mean
RCPR [5]	5.43	6.58	11.53	7.85
ESR [6]	5.66	7.12	11.94	8.24
SDM [30]	4.75	5.55	9.34	6.55
3DDFA [33]	5.00	5.06	6.74	5.60
3DDFA+SDM [33]	4.75	4.83	6.38	5.32
HyperFace [22]	3.93	4.14	4.71	4.26
Ours	**3.90**	**4.10**	**4.70**	**4.24**

IV. EXPERIMENTAL RESULTS

In this section, we evaluate the proposed method using three different tasks, including 2D face alignment on the AFLW database, 3D face alignment on the AFLW2000-3D database, and head pose estimation on the AFLW database.

A. Databases, Baselines, and Metric

Databases. We provide evaluations on three widely used public domain databases: 300W-LP [33], AFLW2000-3D [33] and AFLW [17]. The 300W-LP database is an enlarged version of the 300W database, which contains 61,225 samples from four databases (1,786 from IBUG, 5,207 from

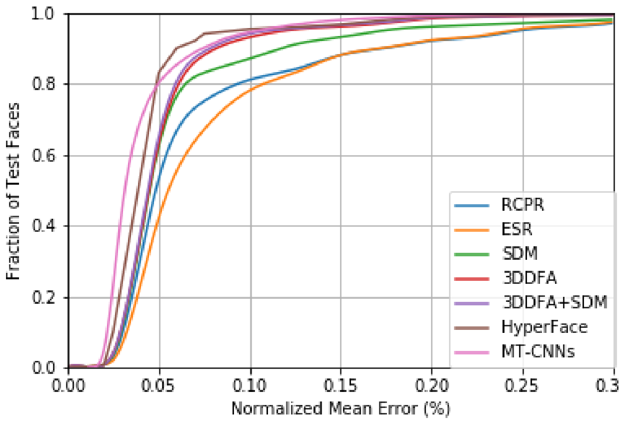

Figure 5. Comparisons of cumulative errors distribution (CED) curves on AFLW dataset. To balance the pose distribution, we plot the CED curves with a subset of 12,081 samples whose absolute yaw angles within $[0^0, 30^0]$,$[30^0, 60^0]$ and $[60^0, 90^0]$ are 1/3 each.

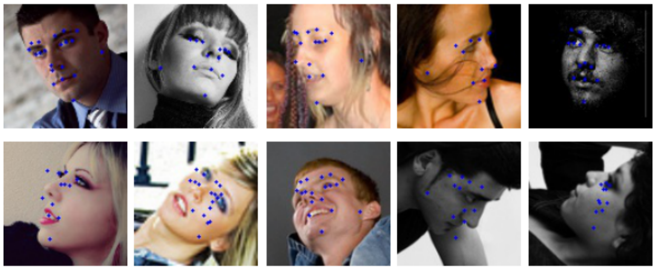

Figure 6. Examples of the 2D face alignment by the proposed approach on the AFLW database. The two rows show the promising cases with the detected visible facial landmarks on 2D images and some failure cases respectively.

AFW, 16,556 from LFPW and 37,676 from HELEN). The AFLW2000-3D database derived from the AFLW database [17], is constructed for 3D face alignment evaluation, which contains 2D face image and its corresponding 3D face shape. The AFLW database contains 21,080 in-the-wild faces with large-pose variations. And each face image is annotated with up to 21 visible keypoints. We follow the same evaluation protocol in [33] for 3D face alignment and 2D face alignment. For head pose estimation, we report our results on AFLW database and the comparisons with Approximate View Manifold (AVM) [26].

Metric. For measuring the face alignment accuracy, we employ Normalized Mean Error (NME) as follows

$$L_r(\hat{S}) = \frac{1}{N}\sum_{i=1}^{N}\frac{\sqrt{|\hat{s}_i - s_i|_{L2}}}{d}, \qquad (17)$$

where N is the number of the facial landmarks; d is the sqrt of the face image box area; $\hat{s}_i$ is the estimated i_{th} facial landmark coordinates, while s_i is the ground-truth i_{th} facial landmark coordinates.

For measuring face pose estimation accuracy, we employ Mean Angular Error (MAE) as follows

$$L_p = \frac{|\hat{p} - p|_{L1}}{3}, \qquad (18)$$

where $\hat{p}$ is the predicted face pose rotations, p is the ground-truth face pose rotations.

Baselines. We compare the proposed approach with a number of 2D face alignment and 3D face alignment methods including RCPR [5], ESR [6], SDM [30], 3DDFA [33] and HyperFace [22]. Since the proposed approach can jointly regress the face pose rotations, we also provide comparisons with the state-of-the-art face pose estimation methods AVM [26], Learning-manifold-in-the-wild [14], Feature-embedding [27], Patch-based [1] and Mixture of trees [21].

B. 2D Face Alignment

In this experiment, we use 300W-LP as the training set and the whole AFLW as the testing set. Each face image in AFLW database is annotated with up to 21 visible keypoints. And the 2D face alignment accuracy is reported using NME, which is the average of visible landmark error normalised by the bounding box size. Table I demonstrates the comparison results and Figure 5 shows the corresponding CED curves. The proposed method can be comparable with the state-of-the-art methods.

The reasons are two-folds: (i) Our method uses a cascased MT-CNNs architecture to directly regress the dense 3D face shape instead of 3DMM PCA coefficients like [33], which is able to recover more facial details, such as the wrinkles and nasolabial folds. These details improve the facial landmark detection accuracy; (ii) Face alignment, 3D face reconstruction and head pose estimation are highly correlated tasks. Therefore, jointly optimizing these tasks is helpful for learning more robust features compared with optimizing each single task separately. In Figure 6, the first row shows some successful cases to demonstrate that the proposed method can deal with large pose, expression variations and occlusions, while some failure cases are shown in the second row. These failure cases have very large in-plane or out-plane rotations. We find that there are a few face images with about 90^0 yaw or roll rotations. The proposed method will be more robust against such large pose variations if more training face images of the same conditions are available.

C. 3D Face Alignment

Since the proposed face alignment approach also reconstruct a dense 3D face shape, we can also use it for 3D face alignment. Following the commonly used protocols [33], we train our cascased MT-CNNs on the 300W-LP database, and use the AFLW2000-3D database for testing. AFLW2000-3D is divided into three parts according to the absolute yaw degrees: 1,306 samples in $[0^0, 30^0]$, 462 samples in $[30^0, 60^0]$, and 232 samples in $[60^0, 90^0]$. Table II

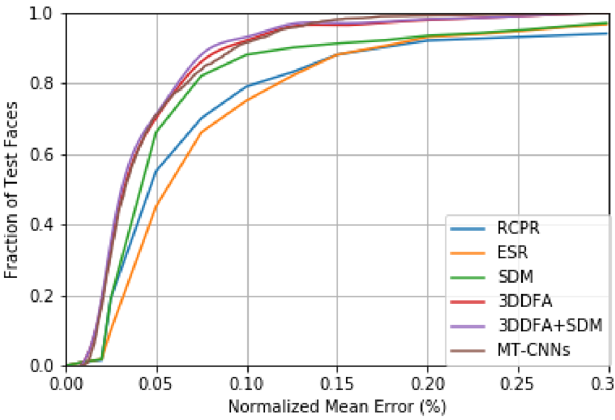

Figure 8. Comparisons of cumulative errors distribution (CED) curves on AFLW2000 dataset. To balance the pose distribution, we plot the CED curves with a subset of 696 samples whose absolute yaw angles within $[0^0, 30^0]$,$[30^0, 60^0]$ and $[60^0, 90^0]$ are 1/3 each.

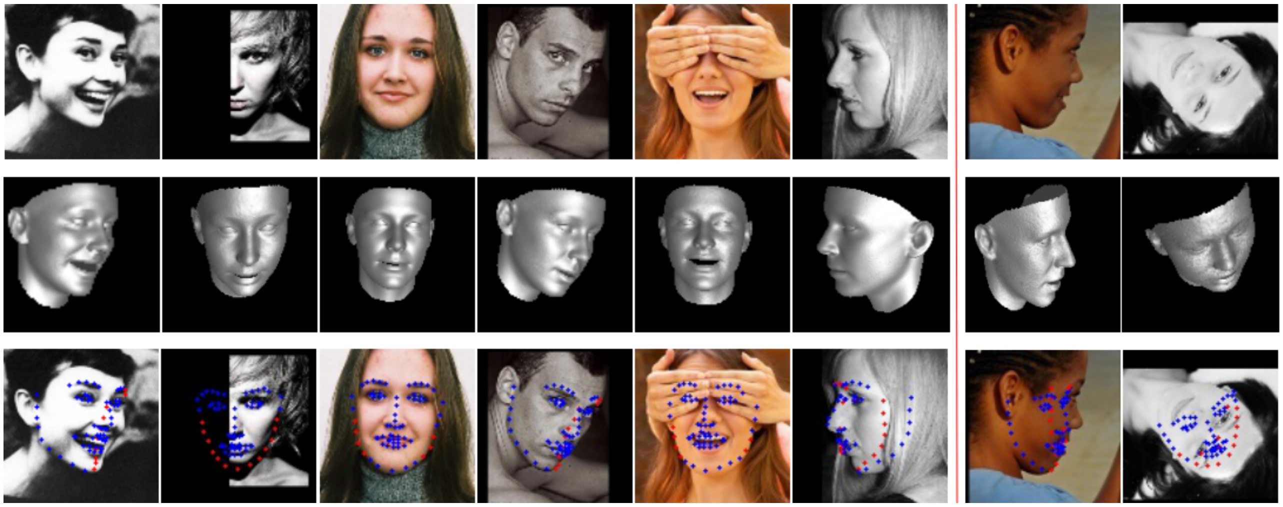

Figure 7. Examples of the face alignment by the proposed approach on the AFLW2000-3D database. The three rows show the input images, the reconstructed 3D face shape, and the landmark detection results on 2D images, respectively. The invisible landmarks on the 2D face image are shown in red. The successful cases (left side of the red line) demonstrate that our method shows promising robustness against large pose, expression and occlusion variations. Some failure cases (right side of the red line) indicate that the proposed method fails while the input 2D face image have a much larger in-plane rotation or out-plane rotation.

Table II
COMPARISONS OF 3D FACE ALIGNMENT NME (IN %) BETWEEN THE PROPOSED APPROACH AND THE STATE-OF-THE-ART METHODS ON THE AFLW2000-3D DATABASE.

Method	AFLW2000-3D Database (68 pts)			
	[0,30]	[30,60]	[60,90]	Mean
RCPR [5]	4.26	5.96	13.18	7.80
ESR [6]	4.60	6.70	12.67	7.99
SDM [30]	3.67	4.94	9.76	6.12
3DDFA [33]	3.78	4.54	7.93	5.42
3DDFA+SDM [33]	3.43	4.24	7.17	4.94
Ours	**3.24**	**5.85**	**9.24**	**6.11**

demonstrates the comparison results and Figure 8 shows the corresponding CED curves. Our method outperforms all the state-of-the-art methods [5], [6], [30], [33] when the absolute yaw rotations are smaller than 30^0. But when absolute yaw rotations is more than 30^0, 3DDFA performs better than our method. One possible reason is that while we only used a two-stage MT-CNN framework, 3DDFA employs a three-stage network and SDM for model learning, which is able to model the complicated non-linear relationship under large poses. In Figure 7, we present some face alignment results by the proposed approach under challenging conditions.

D. *Face Pose Estimation*

Head pose estimation is an auxiliary task of 2D face alignment, which aims to estimate the face poses in terms of $(pitch, yaw, roll)$. In this experiment, 300W-LP is used as the training set, and the evaluations and comparisons are preformed on the AFLW database. We use two evaluation metrics: one is Mean Angular Error (MAE), and the other one is accuracy (% of images with $\pm 15^0$ error). The results are reported in Table III. Our cascaded MT-CNNs notably outperforms other baseline methods [26], [14], [27], [1], [21]. In addition to multi-task learning, another possible reason why the proposed approach work better than the baselines is that our method works in a coarse-to-fine manner, and thus gradually obtains the final pose estimation with a high confidence.

E. *Running Time*

During testing, the proposed cascaded MT-CNNs takes 32ms per image on 3.4 GHz CPU, while 3DDFA [33] and HyperFace [22] take 75.72ms per image on a TITAN Black GPU and 100ms per image on a TITAN X GPU respectively.

V. CONCLUSION

Face alignment across large pose is a challenging task. We proposed an efficient approach for real-time 2D face alignment via a cascaded multi-task CNN (MT-CNN) based on 3D face reconstruction. Our MT-CNN contains a shared feature learning by all the tasks (3D reconstruction, pose

Table III
COMPARISONS OF HEAD POSE ESTIMATION MAE (IN DEGREES) AND ACCURACY BETWEEN THE PROPOSED APPROACH AND THE STATE-OF-THE-ART METHODS ON THE AFLW DATABASE.

Method	MAE	Accuracy (%)
Mixture of trees [21]	46.54	15.72
Patch-based [1]	38.39	23.87
Feature-embedding [27]	33.01	32.82
Learning-manifold-in-the-wild [14]	16.31	63.13
AVM [26]	17.48	58.15
Ours	**11.92**	**73.38**

estimation, and face alignment), and task-dependent feature optimizations w.r.t. individual tasks. Evaluations on the public domain AFLW2000-3D and AFLW database show that the proposed approach can not only run in real-time (32ms per image) on 3.4 GHz CPU, but also can be comparable with the state-of-the-art methods on 2D face alignment, 3D face alignment, and pose estimation tasks.

In our future work, we would like to provide additional annotations to some public domain database, so that they can be utilized for studying multi-task modeling for additional tasks, such as facial expression analysis. We would also like to generalize the proposed approach into the scenario of depth estimation from a 2D face image [8].

Acknowledgment

This research was supported in part by the National Key R&D Program of China (grant 2017YFB1002802), Natural Science Foundation of China (grants 61390511, 61732004, 61672496, 61650202 and 61702481), External Cooperation Program of Chinese Academy of Sciences (CAS) (grant GJHZ1843), Strategic Priority Research Program of CAS (grant XDB02070004), Beijing Science and Technology Program (Z151100000915072), and Youth Innovation Promotion Association CAS (2018135).

References

[1] J. Aghajanian and S. J. D. Prince. Face pose estimation in uncontrolled environments. In *Proc. BMVC*, pages 457–463, 2009.

[2] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(12):2930–2940, Dec.2013.

[3] V. Bettadapura. Face expression recognition and analysis: the state of the art. *arXiv preprint arXiv:1203.6722*, 2012.

[4] V. Blanz and T. Vetter. Face recognition based on fitting a 3D morphable model. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(9):1063–1074, Sep.2003.

[5] X. P. Burgos-Artizzu, P. Perona, and P. Dollár. Robust face landmark estimation under occlusion. In *Proc. ICCV*, pages 1513–1520, 2013.

[6] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. *International Journal of Computer Vision*, 107(2):177–190, Oct.2013.

[7] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(6):681–685, Jun.2001.

[8] J. Cui, H. Zhang, H. Han, S. Shan, and X. Chen. Improving 2D face recognition via discriminative face depth estimation. In *Proc. ICB*, pages 1–8, 2018.

[9] P. Dollár, P. Welinder, and P. Perona. Cascaded pose regression. In *Proc. IEEE CVPR*, pages 1078–1085, 2010.

[10] R. Girshick. Fast r-cnn. In *Proc. IEEE CVPR*, pages 1440–1448, 2015.

[11] H. Han and A. K. Jain. 3D face texture modeling from uncalibrated frontal and profile images. In *Proc. BTAS*, pages 223–230, 2012.

[12] H. Han, A. K. Jain, F. Wang, S. Shan, and X. Chen. Heterogeneous face attribute estimation: A deep multi-task learning approach. *IEEE Trans. Pattern Anal. Mach. Intell.*, PP(99):1–1, Aug.2017.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proc. IEEE ICCV*, pages 1026–1034, 2015.

[14] C. Hegde, A. C. Sankaranarayanan, and R. G. Baraniuk. Learning manifolds in the wild. *Journal of Machine Learning Research*, 1(2):4, Jul.2012.

[15] A. Jourabloo and X. Liu. Large-pose face alignment via CNN-based dense 3D model fitting. In *Proc. IEEE CVPR*, pages 4188–4196, 2016.

[16] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[17] M. Kstinger, P. Wohlhart, P. M. Roth, and H. Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *Proc. ICCV Workshops*, pages 2144–2151, 2011.

[18] F. Liu, D. Zeng, Q. Zhao, and X. Liu:. Joint face alignment and 3D face reconstruction. In *Proc. ECCV*, pages 545–560, 2016.

[19] K. Niinuma, H. Han, and A. K. Jain. Automatic multi-view face recognition via 3D model based pose regularization. In *Proc. IEEE BTAS*, pages 1–8, 2013.

[20] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *Proc. BMVC*, pages 41.1–41.12, 2015.

[21] D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Proc. IEEE CVPR*, pages 2879–2886, 2012.

[22] R. Ranjan, V. M. Patel, and R. Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, PP(99):1–1, Dec.2016.

[23] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, Jun.2014.

[24] Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. In *Proc. NIPS*, pages 1988–1996, 2014.

[25] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *Proc. IEEE CVPR*, pages 3476–3483, 2013.

[26] K. Sundararajan and D. L. Woodard. Head pose estimation in the wild using approximate view manifolds. In *Proc. IEEE CVPR Workshops*, pages 50–58, 2015.

[27] M. Torki and A. Elgammal. Regression from local features for viewpoint and pose estimation. In *Proc. IEEE ICCV*, pages 2603–2610, 2011.

[28] B. Van Ginneken, A. F. Frangi, J. J. Staal, B. M. ter Haar Romeny, and M. A. Viergever. Active shape model segmentation with optimal features. *IEEE Trans. Med. Imag.*, 21(8):924–933, Dec.2002.

[29] F. Wang, H. Han, S. Shan, and X. Chen. Deep multi-task learning for joint prediction of heterogeneous face attributes. In *Proc. IEEE FG*, pages 173–179, 2017.

[30] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *Proc. IEEE CVPR*, pages 532–539, 2013.

[31] J. Zhang, S. Shan, M. Kan, and X. Chen. Coarse-to-fine auto-encoder networks (CFAN) for real-time face alignment. In *Proc. ECCV*, pages 1–16, 2014.

[32] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In *Proc. ECCV*, pages 94–108, 2014.

[33] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3D solution. In *Proc. IEEE CVPR*, pages 146–155, 2016.