Face Alignment Across Large Poses: A 3D Solution

Xiangyu Zhu[1] Zhen Lei[1] Xiaoming Liu[2] Hailin Shi[1] Stan Z. Li[1]
[1]Institute of Automation, Chinese Academy of Sciences
[2]Department of Computer Science and Engineering, Michigan State University
{xiangyu.zhu,zlei,hailin.shi,szli}@nlpr.ia.ac.cn liuxm@msu.edu

Abstract

Face alignment, which fits a face model to an image and extracts the semantic meanings of facial pixels, has been an important topic in CV community. 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°. The challenges are three-fold: Firstly, the commonly used landmark-based face model assumes that all the landmarks are visible and is therefore not suitable for profile views. Secondly, the face appearance varies more dramatically across large poses, ranging from frontal view to profile view. Thirdly, labelling landmarks in large poses is extremely challenging since the invisible landmarks have to be guessed. In this paper, we propose a solution to the three problems in an new alignment framework, called 3D Dense Face Alignment (3DDFA), in which a dense 3D face model is fitted to the image via convolutional neutral network (CNN). We also propose a method to synthesize large-scale training samples in profile views to solve the third problem of data labelling. Experiments on the challenging AFLW database show that our approach achieves significant improvements over state-of-the-art methods.

1. Introduction

Traditional face alignment aims to locate face fiducial points like "eye corner", "nose tip" and "chin center", based on which the face image can be normalized. It is an essential preprocessing step for many face analysis tasks, e.g., face recognition [41], expression recognition [5] and inverse rendering [1]. The researches in face alignment can be divided into two categories: the analysis-by-synthesis based [12, 42, 15] and regression based [11, 17, 27, 45]. The former simulates the process of image generation and achieves alignment by minimizing the difference between model appearance and input image. The latter extracts features around key points and regresses it to the ground truth landmarks. With the development in the last decade, face alignment across medium poses, where the yaw angle is less than 45° and all the landmarks are visible, has been well addressed [45, 51, 54]. However, face alignment across large poses (±90°) is still a challenging problem without much attention and achievements. There are three main challenges:

Figure 1. Fitting results of 3DDFA. For each pair of the four results, on the left is the rendering of the fitted 3D shape with the mean texture, which is made transparent to demonstrate the fitting accuracy. On the right is the landmarks overlayed on the 3D face model, in which the blue/red ones indicate visible/invisible landmarks. The visibility is directly computed from the fitted dense model by [21]. More results are demonstrated in supplemental material.

Modelling: Landmark shape model [13] implicitly assumes that each landmark can be robustly detected based on its distinctive visual patterns. However, when faces deviate from the frontal view, some landmarks become invisible due to self-occlusion [53]. In medium poses, this problem can be addressed by changing the semantic positions of face contour landmarks to the silhouette, which is termed landmark marching [55]. However, in large poses where half of face is occluded, some landmarks are inevitably invisible and have no image data. As a result, the landmark shape model no longer works well.

Fitting: Face alignment across large poses is more challenging than medium poses due to the dramatic appearance variations when close to the profile views. The cascaded linear regression [45] or traditional nonlinear models [27, 50, 10] are not sophisticated enough to cover such complicated patterns in a unified way. The view-based

framework, which adopts different landmark configurations and fitting models for each view category [53, 49, 56, 38], may significantly increase computation cost since every view has to be tested.

Data Labelling: The most serious problem comes from the data. Manual labelling landmarks on large-pose faces is a very tedious task. Firstly, no algorithm can provide a good initialization to reduce the workload. Secondly, the occluded landmarks have to be "guessed" which is impossible for most of people. As a result, almost all public face alignment databases such as AFW [56], LFPW [22], HELEN [26] and IBUG [35] are collected in medium poses. Existing large-pose databases such as AFLW [25] only contains visible landmarks, which could be ambiguous in invisible landmarks and hard to train a unified face alignment model.

In this paper, we address all the three challenges with the goal of improving the face alignment performance across large poses.

1. To address the problem of invisible landmarks in large poses, we propose to fit the 3D dense face model rather than the sparse landmark shape model to the image. By incorporating 3D information, the appearance variations and self-occlusion caused by 3D transformations can be inherently addressed. We call this method 3D Dense Face Alignment (3DDFA). Some results are shown in Fig. 1.

2. To resolve the fitting process in 3DDFA, we propose a cascaded convolutional neutral network (CNN) based regression method. CNN has been proved of excellent capability to extract useful information from images with large variations in object detection [48] and image classification [40]. In this work, we adopt CNN to fit the 3D face model with a specifically designed feature, namely Projected Normalized Coordinate Code (PNCC). Besides, Weighted Parameter Distance Cost (WPDC) is proposed as the cost function. To the best of our knowledge, this is the first attempt to solve the 3D face alignment with CNN.

3. To enable the training of the 3DDFA, we construct a face database containing pairs of 2D face images and 3D face models. We further propose a face profiling algorithm to synthesize $60k+$ training samples across large poses. The synthesized samples well simulate the face appearances in large poses and boost the performance of both prior and our proposed face alignment algorithms.

The database, face profiling code and 3DDFA code are released at http://www.cbsr.ia.ac.cn/users/xiangyuzhu/.

2. Related Works

Generic Face Alignment: Face alignment in 2D aims at locating a sparse set of fiducial facial landmarks. A number of achievements have been made including the classic Active Appearance Model (AAM) [12, 36, 42] and Constrained Local Model (CLM) [16, 37, 2]. Recently, the regression based method, which maps the discriminative features around landmarks to the desired landmark positions [43, 45, 46, 10, 50, 27], has been proposed. By utilizing the feedback characteristic that the the output (landmark positions) of the regression has an influence on the input (features at landmarks), the cascaded regression [17] cascades a list of weak regressors to reduce the alignment error progressively and reaches the state of the art [46, 54].

Besides traditional models, convolutional neutral network (CNN) has also been employed in face alignment recently. Sun et al. [39] firstly use CNN to regress landmark locations with the raw face image. Liang et al. [28] improve the flexibility by estimating the landmark response map. Zhang et al. [51] further combine face alignment with attribute analysis through multi-task CNN to boost the performance of both tasks. Although with considerable achievements, most CNN methods only detect a sparse set of landmarks (5 points in [39, 51, 28]) with limited descriptive power of face shape.

Large Pose Face Alignment: Despite the great attentions on face alignment, literature on large-pose scenario is rather limited. The most common method is the multi-view framework [14], which uses different landmark configurations for different views. For example, TSPM [56] and CDM [49] employ DPM-like [18] method to align faces with different shape models, among which the highest possibility is chosen as the final result. However, since every view has to be tested, the computation cost of multi-view method is always high.

Besides 2D methods, 3D face alignment [19], which aims to fit a 3D morphable model (3DMM) [6] from a 2D image, also has the potential to deal with large poses. It models the 3D face shape with a linear subspace (PCA [6] or Tensor [8]) and achieves fitting by minimizing the difference between image and model appearance. 3DMM can cover arbitrary poses [6, 33] but suffers from the one-minute-per-image computation cost. Recently, regression based 3DMM fitting, which estimates the model parameters by regressing the features at landmark positions [49, 24, 8, 23], has been proposed to improve the efficiency. However, since the features at landmarks may be self-occluded as in 2D methods, the fitting algorithm is no longer pose-invariant and suffers from the three problems in Section 1. A relevant but different problem is the 3D face reconstruction [1, 44, 55, 20], which recovers a 3D face from given 2D landmarks. Interestingly, based on that 2D/3D face alignment results can be mutually transformed, where 3D

to 2D is made by selecting x, y coordinates of landmark vertexes and 2D to 3D is made by 3D face reconstruction.

3. 3D Dense Face Alignment (3DDFA)

In this section we introduce the 3D Dense Face Alignment (3DDFA) which fits 3D morphable model with cascaded CNN.

3.1. 3D Morphable Model

Blanz et al. [6] propose the 3D morphable model (3DMM) which describes the 3D face space with PCA:

$$\mathbf{S} = \overline{\mathbf{S}} + \mathbf{A}_{id}\boldsymbol{\alpha}_{id} + \mathbf{A}_{exp}\boldsymbol{\alpha}_{exp}, \quad (1)$$

where $\mathbf{S}$ is a 3D face, $\overline{\mathbf{S}}$ is the mean shape, $\mathbf{A}_{id}$ is the principle axes trained on the 3D face scans with neutral expression and $\boldsymbol{\alpha}_{id}$ is the shape parameter, $\mathbf{A}_{exp}$ is the principle axes trained on the offsets between expression scans and neutral scans and $\boldsymbol{\alpha}_{exp}$ is the expression parameter. In this work, the $\mathbf{A}_{id}$ and $\mathbf{A}_{exp}$ come from BFM [31] and FaceWarehouse [9] respectively. The 3D face is then projected onto the image plane with Weak Perspective Projection:

$$V(\mathbf{p}) = f * \mathbf{Pr} * \mathbf{R} * (\overline{\mathbf{S}} + \mathbf{A}_{id}\boldsymbol{\alpha}_{id} + \mathbf{A}_{exp}\boldsymbol{\alpha}_{exp}) + \mathbf{t}_{2d}, \quad (2)$$

where $V(\mathbf{p})$ is the model construction and projection function, leading to the 2D positions of model vertexes, f is the scale factor, $\mathbf{Pr}$ is the orthographic projection matrix $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$, $\mathbf{R}$ is the rotation matrix constructed from rotation angles $pitch$, yaw, $roll$ and $\mathbf{t}_{2d}$ is the translation vector. The collection of all the model parameters is $\mathbf{p} = [f, pitch, yaw, roll, \mathbf{t}_{2d}, \boldsymbol{\alpha}_{id}, \boldsymbol{\alpha}_{exp}]^T$.

3.2. Network Structure

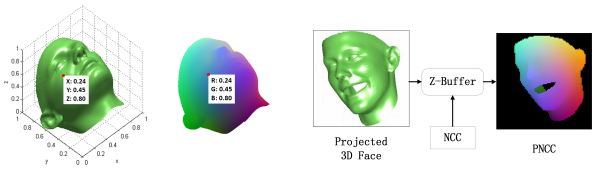
The purpose of 3D face alignment is estimating $\mathbf{p}$ from a single face image $\mathbf{I}$. Unlike existing CNN methods [39, 28] which apply different networks for different fitting stages, 3DDFA employ a unified network structure across the cascade. In general, at iteration k ($k = 0, 1, ..., K$), given an initial parameter $\mathbf{p}^k$, we construct a specially designed feature PNCC with $\mathbf{p}^k$ and train a convolutional neutral network Net^k to predict the parameter update $\Delta\mathbf{p}^k$:

$$\Delta\mathbf{p}^k = Net^k(\mathbf{I}, \text{PNCC}(\mathbf{p}^k)). \quad (3)$$

Afterwards, a better medium parameter $\mathbf{p}^{k+1} = \mathbf{p}^k + \Delta\mathbf{p}^k$ becomes the input of the next network Net^{k+1} which has the same structure as Net^k. Fig. 2 shows the network structure. The input is the $100 \times 100 \times 3$ color image stacked by PNCC. The network contains four convolution layers, three pooling layers and two fully connected layers. The first two convolution layers share weights to extract low-level features. The last two convolution layers do not share weights to extract location sensitive features, which is further regressed to a 256-dimensional feature vector. The output is a 234-dimensional parameter update including 6-dimensional pose parameters $[f, pitch, yaw, roll, t_{2dx}, t_{2dy}]$, 199-dimensional shape parameters $\boldsymbol{\alpha}_{id}$ and 29-dimensional expression parameters $\boldsymbol{\alpha}_{exp}$.

3.3. Projected Normalized Coordinate Code

The special structure of the cascaded CNN has three requirements of its input feature: Firstly, the **feedback property** requires that the input feature should depend on the CNN output to enable the cascade manner. Secondly, the **convergence property** requires that the input feature should reflect the fitting accuracy to make the cascade converge after some iterations [57]. Finally, the **convolvable property** requires that the convolution on the input feature should make sense. Based on the three properties, we



(a) NCC (b) PNCC

Figure 3. The Normalized Coordinate Code (NCC) and the Projected Normalized Coordinate Code (PNCC). (a) The normalized mean face, which is also demonstrated with NCC as its texture (NCC_x = R, NCC_y = G, NCC_z = B). (b) The generation of PNCC: The projected 3D face is rendered by Z-Buffer with NCC as its colormap.

design our features as follows: Firstly, the 3D mean face is normalized to $0-1$ in x, y, z axis as Equ. 4. The unique 3D coordinate of each vertex is called its Normalized Coordinate Code (NCC), see Fig. 3(a).

$$\text{NCC}_d = \frac{\overline{\mathbf{S}}_d - \min(\overline{\mathbf{S}}_d)}{\max(\overline{\mathbf{S}}_d) - \min(\overline{\mathbf{S}}_d)} \quad (d = x, y, z), \quad (4)$$

where the $\overline{\mathbf{S}}$ is the mean shape of 3DMM in Equ. 1. Since NCC has three channels as RGB, we also show the mean face with NCC as its texture. Secondly, with a model parameter $\mathbf{p}$, we adopt the Z-Buffer to render the projected 3D face colored by NCC as in Equ. 5, which is called the Projected Normalized Coordinate Code (PNCC), see Fig. 3(b):

$$\begin{aligned} \text{PNCC} &= \textit{Z-Buffer}(V_{3d}(\mathbf{p}), \text{NCC}) \\ V_{3d}(\mathbf{p}) &= f * \mathbf{R} * \mathbf{S} + [\mathbf{t}_{2d}, 0]^T \\ \mathbf{S} &= \overline{\mathbf{S}} + \mathbf{A}_{id}\boldsymbol{\alpha}_{id} + \mathbf{A}_{exp}\boldsymbol{\alpha}_{exp}, \end{aligned} \quad (5)$$

where $\textit{Z-Buffer}(\boldsymbol{\nu}, \boldsymbol{\tau})$ renders an image from the 3D mesh $\boldsymbol{\nu}$ colored by $\boldsymbol{\tau}$ and $V_{3d}(\mathbf{p})$ is the current 3D face. Afterwards,

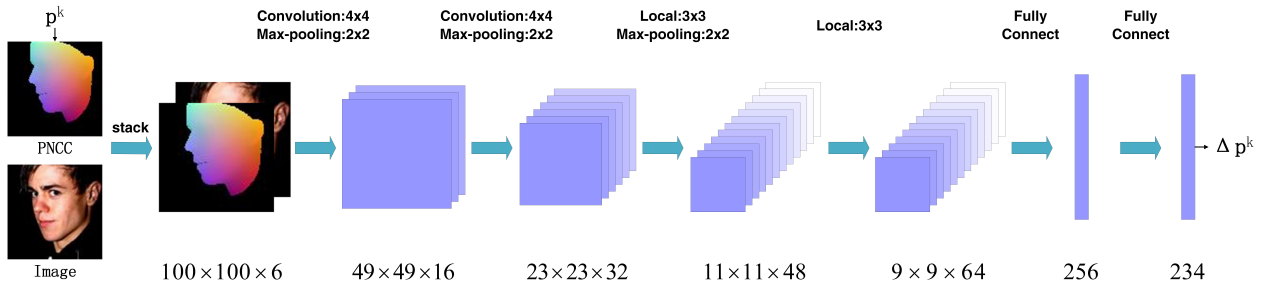


Figure 2. An overview of 3DDFA. At kth iteration, Net^k takes a medium parameter $\mathbf{p}^k$ as input, constructs the projected normalized coordinate code (PNCC), stacks it with the input image and sends it into CNN to predict the parameter update $\Delta\mathbf{p}^k$.

PNCC is stacked with the input image and transferred to CNN. Regarding the three properties, PNCC fulfills the feedback property since in Equ. 5, $\mathbf{p}$ is the output of CNN and NCC is a constant. Secondly, PNCC provides the 2D locations of visible 3D vertexes on the image plane. When CNN detects that each NCC superposes its corresponding image pattern during testing, the cascade will converge. PNCC fulfills the convergence property. Note that the invisible region is automatically ignored by Z-Buffer. Finally, PNCC is smooth in 2D space, the convolution indicates the linear combination of NCCs on a local patch. It fulfills the convolvable property.

3.4. Cost Function

The performance of CNN can be greatly impacted by the cost function, which is difficult to design in 3DDFA since each dimension of the CNN output (model parameter) has different influence on the 3DDFA result (fitted 3D face). In this work, we discuss two baselines and propose a novel cost function. Since the parameter range varies significantly, we conduct z-score normalization before training.

3.4.1 Parameter Distance Cost (PDC)

Take the first iteration as an example. The purpose of CNN is predicting the parameter update $\Delta\mathbf{p}$ to move the initial parameter $\mathbf{p}^0$ closer to the ground truth $\mathbf{p}^g$. Intuitively, we can minimize the distance between the ground truth and the current parameter with the Parameter Distance Cost (PDC):

$$E_{pdc} = \|\Delta\mathbf{p} - (\mathbf{p}^g - \mathbf{p}^0)\|^2. \tag{6}$$

Even though PDC has been used in 3D face alignment [57], there is a problem that each dimension in $\mathbf{p}$ has different influence on the resultant 3D face. For example, with the same deviation, the yaw angle will bring a larger alignment error than a shape PCA coefficient, while PDC optimizes them equally.

3.4.2 Vertex Distance Cost (VDC)

Since 3DDFA aims to morph the 3DMM to the ground truth 3D face, we can optimize $\Delta\mathbf{p}$ by minimizing the vertex distances between the fitted and the ground truth 3D face:

$$E_{vdc} = \|V(\mathbf{p}^0 + \Delta\mathbf{p}) - V(\mathbf{p}^g)\|^2, \tag{7}$$

where $V(\cdot)$ is the face construction and weak perspective projection as Equ. 2. This cost is called the Vertex Distance Cost (VDC) and the derivative is provided in supplemental material. Compared with PDC, VDC better models the fitting error by explicitly considering the semantics of each parameter. However, we observe that VDC exhibits pathological curvature [29]. The directions of pose parameters always exhibit much higher curvatures than the PCA coefficients. As a result, optimizing VDC with gradient descend converges very slowly due to the "zig-zagging" problem. Second-order optimizations are preferred but they are expensive and hard to be implemented on GPU.

3.4.3 Weighted Parameter Distance Cost (WPDC)

In this work, we propose a simple but effective cost function Weighted Parameter Distance Cost (WPDC). The basic idea is explicitly modeling the importance of each parameter:

$$\begin{aligned} E_{wpdc} &= (\Delta\mathbf{p} - (\mathbf{p}^g - \mathbf{p}^0))^T\mathbf{W}(\Delta\mathbf{p} - (\mathbf{p}^g - \mathbf{p}^0)) \\ where \quad \mathbf{W} &= diag(w_1, w_2, ..., w_n) \\ w_i &= \|V(\mathbf{p}^d(i)) - V(\mathbf{p}^g)\| / \sum w_i \\ \mathbf{p}^d(i)_i &= (\mathbf{p}^0 + \Delta\mathbf{p})_i \\ \mathbf{p}^d(i)_j &= \mathbf{p}^g_j, \quad j \in \{1, \ldots, i-1, i+1, \ldots, n\}, \end{aligned} \tag{8}$$

where $\mathbf{W}$ is the importance matrix whose diagonal is the weight of each parameter, $\mathbf{p}^d(i)$ is the i-deteriorated parameter whose ith component comes from the predicted parameter $(\mathbf{p}^0 + \Delta\mathbf{p})$ and the others come from the ground truth parameter $\mathbf{p}^g$, $\|V(\mathbf{p}^d(i)) - V(\mathbf{p}^g)\|$ models the alignment error brought by miss-predicting the ith model parameter, which is indicative of its importance. For simplicity, $\mathbf{W}$ is considered as a constant when computing the derivative. In

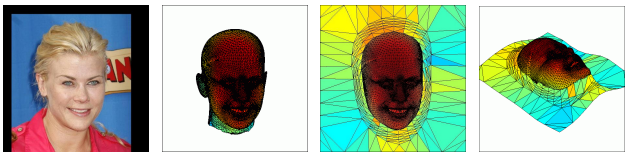
the training process, CNN firstly concentrate on the parameters with larger $\|V(\mathbf{p}^d(i)) - V(\mathbf{p}^g)\|$ such as scale, rotation and translation. As $\mathbf{p}^d(i)$ is closer to $\mathbf{p}^g$, the weights of these parameters begin to shrink and CNN will optimize less important parameters but at the same time keep the high-priority parameters sufficiently good. Compared with VDC, the WPDC remedies the pathological curvature issue and is easier to implement without the derivative of $V(\cdot)$.

4. Face Profiling

All the discriminative models rely on the training data, especially for CNN which has thousands of parameters to train. Therefore, massive labelled faces across large poses are crucial for 3DDFA. However, few of released face alignment database contains large-pose samples [56, 22, 26, 35] since labelling standardized landmarks on profile is very challenging. In this section, we demonstrate that labelled profile faces can be well simulated from existing training samples with the help of 3D information. Inspired by the recent breakthrough in face frontalization [55, 21] which generates the frontal view of faces, we propose to invert this process to generate the profile view of faces from medium-pose samples, which is called face profiling. The basic idea is predicting the depth of face image and generating the profile views with 3D rotation.

4.1. 3D Image Meshing

The depth estimation of a face image can be conducted on the face region and external region respectively, with different requirements of accuracy. On the face region, we fit a 3DMM through the Multi-Features Framework [33] (MFF), see Fig. 4(b). With the ground truth landmarks as a solid constraint throughout the fitting process, the MFF can always converge to a very good result. Few failed samples can be easily adjusted manually. On the external region, we follow the 3D meshing method proposed by Zhu et al. [55] to mark some anchors beyond the face region and estimate their depth, see Fig. 4(c). Afterwards the whole image is tuned into a 3D object through triangulation, see Fig. 4(c)4(d).

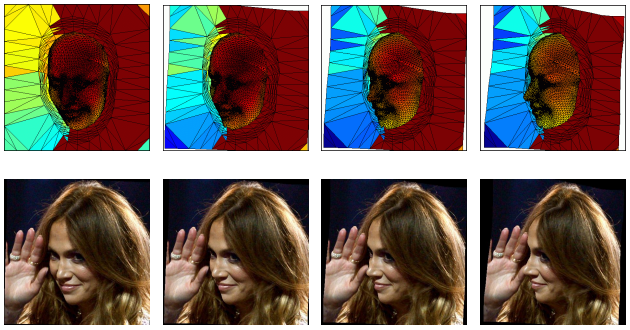


(a) (b) (c) (d)

Figure 4. 3D Image Meshing. (a) The input image. (b) The fitted 3D face through MFF. (c) The depth image from 3D meshing. (d) A different view of the depth image.

4.2. 3D Image Rotation

When the depth information is estimated, the face image can be rotated in 3D space to generate the appearances in larger poses (Fig. 5). It can be seen that the external face region is necessary for a realistic profile image. Different from face frontalization, with larger rotation angles the self-occluded region can only be expanded. As a result, we avoid the troubling invisible region filling which may produce large artifacts [55].



(a) (b) (c) (d)

Figure 5. 2D and 3D view of the image rotation. (a) The original yaw angle yaw_0. (b) yaw_0+20°. (c) yaw_0+30°. (d) yaw_0+40°.

In this work, we enlarge the yaw of the depth image at the step of 5° until 90°. Through face profiling, we not only obtain the face appearances in large poses and but also augment the dataset to a large scale, which means the CNN can be well trained even given a small database.

5. Implementation Details

5.1. Initialization Regeneration

With a huge number of parameters, CNN tends to overfit the training set and the networks at deeper cascade might receive training samples with almost zero errors. Therefore we cannot directly adopt the cascade framework as in 2D face alignment. Asthana et al. [3] demonstrates that the initializations at each iteration can be well simulated with statistics. In this paper, we also regenerate the $\mathbf{p}^k$ but with a more sophisticated method. We observe that the fitting error highly depends on the ground truth face posture (FP). For example, the error of a profile face is mostly caused by a small yaw angle and the error of an open-mouth face is always caused by a close-mouth expression parameter. As a result, it is reasonable to model the perturbation of a training sample with a set of similar-FP samples. In this paper, we define the face posture as the ground truth 2D landmarks without scale and translation:

$$\text{FP} = \mathbf{Pr} * \mathbf{R}^g * (\overline{\mathbf{S}} + \mathbf{A}_{id}\boldsymbol{\alpha}_{id}^g + \mathbf{A}_{exp}\boldsymbol{\alpha}_{exp}^g)_{landmark}, \quad (9)$$

where $\mathbf{R}^g, \boldsymbol{\alpha}_{id}^g, \boldsymbol{\alpha}_{exp}^g$ represent the ground truth pose, shape and expression respectively and the subscript $landmark$

means only landmark points are selected. Before training, we select two folds of samples as the validation set. For each training sample, we construct a validation subset $\{v_1, ..., v_m\}$ whose members share similar FP with the training sample. At iteration k, we regenerate the initial parameter by:

$$\mathbf{p}^k = \mathbf{p}^g - (\mathbf{p}^g_{v_i} - \mathbf{p}^k_{v_i}), \tag{10}$$

where $\mathbf{p}^k$ and $\mathbf{p}^g$ are the initial and ground truth parameter of a training sample, $\mathbf{p}^k_{v_i}$ and $\mathbf{p}^g_{v_i}$ come from a validation sample v_i which is randomly chosen from the corresponding validation subset. Note that v_i is never used in training.

5.2. Landmark Refinement

Dense face alignment method fits all the vertexes of the face model by estimating the model parameters. If we are only interested in a sparse set of points such as landmarks, the error can be further reduced by relaxing the PCA constraint. In the 2D face alignment task, after 3DDFA we extract HOG features at landmarks and train a linear regressor to refine the landmark locations. In fact, 3DDFA can team with any 2D face alignment methods. In the experiment, we also report the results refined by SDM [45].

6. Experiments

In this section, we evaluate the performance of 3DDFA in three common face alignment tasks in the wild, i.e., medium-pose face alignment, large-pose face alignment and 3D face alignment. Due to the space constraint, qualitative alignment results are shown in supplemental material.

6.1. Datasets

Evaluations are conducted with three databases, 300W [34], AFLW [25] and a specifically constructed AFLW2000-3D database.

300W-LP: 300W [34] standardises multiple alignment databases with 68 landmarks, including AFW [56], LFPW [4], HELEN [52], IBUG [34] and XM2VTS [30]. With 300W, we adopt the proposed face profiling to generate 61,225 samples across large poses (1,786 from IBUG, 5,207 from AFW, 16,556 from LFPW and 37,676 from HELEN, XM2VTS is not used), which is further expanded to 122,450 samples with flipping. We call the database as the 300W across Large Poses (300W-LP)

AFLW: AFLW [25] contains 21,080 in-the-wild faces with large-pose variations (yaw from $-90°$ to $90°$). Each image is annotated with up to 21 visible landmarks. The dataset is very suitable for evaluating face alignment performance across large poses.

AFLW2000-3D: Evaluating 3D face alignment in the wild is difficult due to the lack of pairs of 2D image and 3D model in unconstrained environment. Considering the recent achievements in 3D face reconstruction which can construct a 3D face from 2D landmarks [1, 55], we assume that a 3D model can be accurately fitted if sufficient 2D landmarks are provided. Therefore 3D evaluation can be degraded to 2D evaluation which also makes it possible to compare 3DDFA with other 2D face alignment methods. However, AFLW is not suitable for evaluating this task since only visible landmarks lead to serious ambiguity in 3D shape, as reflected by the fake good alignment phenomenon in Fig. 6. In this work, we construct a database called AFLW2000-3D for 3D face alignment evaluation, which contains the ground truth 3D faces and the corresponding 68 landmarks of the first 2,000 AFLW samples. Construction details are provided in supplemental material.

Figure 6. Fake good alignment in AFLW. For each sample, the first shows the visible 21 landmarks and the second shows all the 68 landmarks. The Normalized Mean Error (NME) reflects their accuracy. It can be seen that only evaluating visible landmarks cannot well reflect the fitting accuracy.

6.2. Performance Analysis

Error Reduction in Cascade: To analyze the error reduction process in cascade and evaluate the effect of initialization regeneration. We divide 300W-LP into 97,967 samples for training and 24,483 samples for testing, without identity overlapping. Fig. 7 shows the training and testing errors at each iteration, with and without initialization regeneration. As observed, the testing error is reduced due to

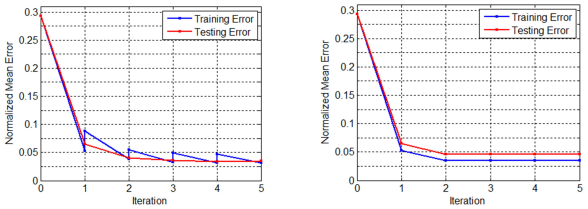


Figure 7. The training and testing errors with (a) and without (b) initialization regeneration.

initialization regeneration. In the generic cascade process the training and testing errors converge fast after 2 iterations. While with initialization regeneration, the training error is updated at the beginning of each iteration and the testing error continues to descend.

During testing, 3DDFA takes 25.24ms for each iteration, 17.49ms for PNCC construction on 3.40GHZ CPU and 7.75ms for CNN on GTX TITAN Black GPU. Note that the computing time of PNCC can be greatly reduced if Z-Buffer is conducted on GPU. Considering both effectiveness and efficiency we choose 3 iterations in 3DDFA.

Performance with Different Costs: In this experiment, we demonstrate the performance with different costs including PDC, VDC and WPDC. Fig. 8 demonstrates the testing errors at each iteration. All the networks are trained until convergence. It is shown that PDC cannot well model the

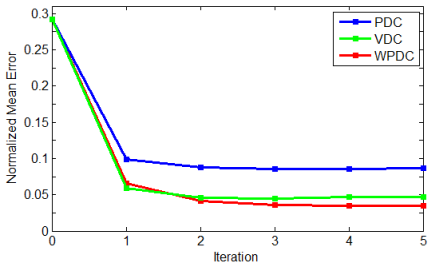


Figure 8. The testing errors with different cost function.

fitting error and converges to an unsatisfied result. VDC is better than PDC, but the pathological curvature problem makes it only concentrate on a small set of parameters, which limits its performance. WPDC explicitly models the priority of each parameter and adaptively optimizes them with the parameter weights, leading to the best result.

6.3. Comparison Experiments

In this paper, we test the performance of 3DDFA on three different tasks, including the large-pose face alignment on AFLW, 3D face alignment on AFLW2000-3D and medium-pose face alignment on 300W.

6.3.1 Large Pose Face Alignment in AFLW

Protocol: In this experiment, we regard 300W and 300W-LP as the training set respectively and the whole AFLW as the testing set. The bounding boxes provided by AFLW are used for initialization (which are not the ground truth). During training, for 2D methods we use the projected 3D landmarks as the ground truth and for 3DDFA we directly regress the 3DMM parameters. 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. The alignment accuracy is evaluated by the Normalized Mean Error (NME), which is the average of visible landmark error normalised by the bounding box size [24, 49]. Note that the metric only considers visible landmarks and is normalized by the bounding box size instead of the common inter-pupil distance. Besides, we also report the standard deviation across testing subsets which is a good measure of pose robustness.

Methods: Since little experiment has been conducted on AFLW, we choose some baseline methods with released codes, including CDM [49], RCPR [7], ESR [10] and SDM [47]. Among them ESR and SDM are popular face alignment methods in recent years. CDM is the first one claimed to perform pose-free face alignment. RCPR is a occlusion-robust method with the potential to deal with self-occlusion and we train it with landmark visibility labels computed by [21]. Table. 1 demonstrates the comparison results and Fig. 9 shows the corresponding CED curves. Each method is trained on 300W and 300W-LP respectively to demonstrate the boost from face profiling. If a trained model is provided in the code, we also demonstrate its performance. Since CDM only contains testing code, we just report its performance with the provided alignment model. For 3DDFA which depends on large scales of data, we only report its performance trained on 300W-LP, with the network structure in Fig. 2.

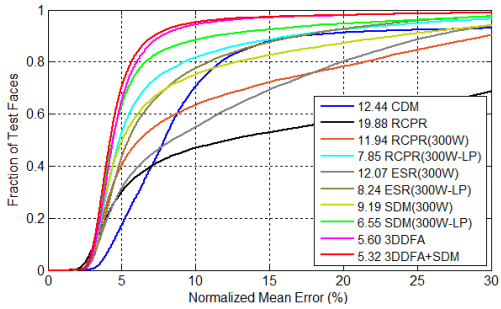


Figure 9. Comparisons of cumulative errors distribution (CED) curves on AFLW. To balance the pose distribution, we plot the CED curves with a subset of 12,081 samples whose absolute yaw angles within $[0°, 30°]$, $[30°, 60°]$ and $[60°, 90°]$ are 1/3 each.

Results: Firstly, the results indicate that all the methods benefits substantially from face profiling when dealing with large poses. The improvements in $[60°, 90°]$ are 44.06% for RCPR, 40.36% for ESR and 42.10% for SDM. This is especially impressive since the alignment models are trained on the synthesized data and tested on real samples. Thus the fidelity of the face profiling method can be well demonstrated. Secondly, 3DDFA reaches the state of the art above all the 2D methods especially beyond medium poses. The minimum standard deviation of 3DDFA also demonstrates its robustness to pose variations. Finally, the performance of 3DDFA can be further improved with the SDM landmark refinement in Section 5.2.

6.3.2 3D Face Alignment in AFLW2000-3D

As described in Section 6.1, 3D face alignment evaluation can be degraded to all-landmark evaluation considering both visible and invisible ones. Using AFLW2000-3D as

Table 1. The NME(%) of face alignment results on AFLW and AFLW2000-3D with the first and the second best results highlighted. The bracket shows the training set. The results of provided alignment models are marked with their references.

	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
CDM [49]	8.15	13.02	16.17	12.44	4.04	-	-	-	-	-
RCPR [7]	6.16	18.67	34.82	19.88	14.36	-	-	-	-	-
RCPR(300W)	5.40	9.80	20.61	11.94	7.83	4.16	9.88	22.58	12.21	9.43
RCPR(300W-LP)	5.43	6.58	11.53	7.85	3.24	4.26	5.96	13.18	7.80	4.74
ESR(300W)	5.58	10.62	20.02	12.07	7.33	4.38	10.47	20.31	11.72	8.04
ESR(300W-LP)	5.66	7.12	11.94	8.24	3.29	4.60	6.70	12.67	7.99	4.19
SDM(300W)	**4.67**	6.78	16.13	9.19	6.10	**3.56**	7.08	17.48	9.37	7.23
SDM(300W-LP)	**4.75**	5.55	9.34	6.55	2.45	3.67	4.94	9.76	6.12	3.21
3DDFA	5.00	**5.06**	**6.74**	**5.60**	**0.99**	3.78	**4.54**	**7.93**	**5.42**	**2.21**
3DDFA+SDM	**4.75**	**4.83**	**6.38**	**5.32**	**0.92**	**3.43**	**4.24**	**7.17**	**4.94**	**1.97**

the testing set, this experiment follows the same protocol as AFLW, except 1) Instead of the visible 21 landmarks, all the MultiPIE-68 landmarks [34] in AFLW2000-3D are used for evaluation. 2) With the ground truth 3D models, the ground truth bounding boxes enclosing all the landmarks are provided for initialization. There are 1,306 samples in $[0°, 30°]$, 462 samples in $[30°, 60°]$ and 232 samples in $[60°, 90°]$. The results are demonstrates in Table. 1 and the CED curves are plot in Fig. 10. We do not report the performance of provided CDM and RCPR models since they do not detect invisible landmarks.

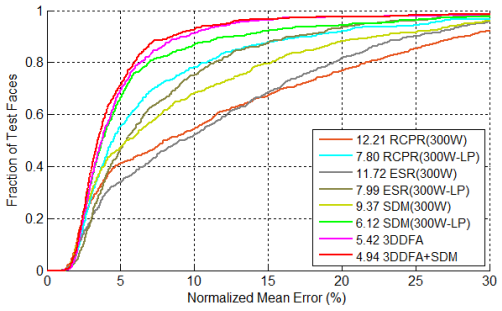


Figure 10. Comparisons of cumulative errors distribution (CED) curves on AFLW2000. To balance the pose distribution, we plot the CED curves with a subset of 696 samples whose absolute yaw angles within $[0°, 30°]$, $[30°, 60°]$ and $[60°, 90°]$ are 1/3 each.

Compared with the results in AFLW, we can see the defect of barely evaluating visible landmarks. For all the methods, despite with ground truth bounding boxes the performance in $[60°, 90°]$ and the standard deviation are obviously reduced when considering all the landmarks. We think for 3D face alignment which depends on both visible and invisible landmarks [1, 23], evaluating all the landmarks are necessary.

6.3.3 Medium Pose Face Alignment

Even though not aimed at advancing face alignment in medium poses, we are also interested in the performance of 3DDFA in this popular task. The experiments are conducted on 300W following the common protocol in [54], where we use the training part of LFPW, HELEN and the whole AFW for training (3,148 images and 50,521 after augmentation), and perform testing on three parts: the test samples from LFPW and HELEN as the common subset, the 135-image IBUG as the challenging subset, and the union of them as the full set (689 images in total). The alignment accuracy are evaluated by standard landmark mean error normalised by the inter-pupil distance (NME). It can be seen in Tabel. 2

Table 2. The NME(%) of face alignment results on 300W, with the first and the second best results highlighted.

Method	Common	Challenging	Full
TSPM [56]	8.22	18.33	10.20
ESR [10]	5.28	17.00	7.58
RCPR [7]	6.18	17.26	8.35
SDM [45]	5.57	15.40	7.50
LBF [32]	**4.95**	11.98	6.32
CFSS [54]	**4.73**	**9.98**	**5.76**
3DDFA	6.15	10.59	7.01
3DDFA+SDM	5.53	**9.56**	**6.31**

that even as a generic face alignment algorithm, 3DDFA still demonstrates competitive performance on the common set and state-of-the-art performance on the challenging set.

7. Conclusions

In this paper, we propose a novel method, 3D Dense Face Alignment (3DDFA), which well solves the problem of face alignment across large poses. Different from the traditional

landmark detection framework, 3DDFA fits a dense 3D morphable model with cascaded CNN to solve the self-occlusion in modelling and the high nonlinearity in fitting in large poses. We also propose a face profiling algorithm to synthesize face appearances in profile view, which can provide abundant samples for training. Experiments show the state-of-the-art performance in AFLW, AFLW2000-3D and 300W. In future work, we believe that 3DDFA can be further improved with more complicated network architecture, like larger input size and deeper network.

References

[1] O. Aldrian and W. A. Smith. Inverse rendering of faces with a 3D morphable model. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(5):1080–1093, 2013. 1, 2, 6, 8

[2] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Robust discriminative response map fitting with constrained local models. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3444–3451. IEEE, 2013. 2

[3] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Incremental face alignment in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1859–1866. IEEE, 2014. 5

[4] P. N. Belhumeur, D. W. Jacobs, D. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 545–552. IEEE, 2011. 6

[5] V. Bettadapura. Face expression recognition and analysis: the state of the art. *arXiv preprint arXiv:1203.6722*, 2012. 1

[6] V. Blanz and T. Vetter. Face recognition based on fitting a 3D morphable model. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(9):1063–1074, 2003. 2, 3

[7] X. P. Burgos-Artizzu, P. Perona, and P. Dollár. Robust face landmark estimation under occlusion. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 1513–1520. IEEE, 2013. 7, 8

[8] C. Cao, Q. Hou, and K. Zhou. Displaced dynamic expression regression for real-time facial tracking and animation. *ACM Transactions on Graphics (TOG)*, 33(4):43, 2014. 2

[9] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. Facewarehouse: a 3d facial expression database for visual computing. *Visualization and Computer Graphics, IEEE Transactions on*, 20(3):413–425, 2014. 3

[10] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2887–2894. IEEE, 2012. 1, 2, 7, 8

[11] T. F. Cootes, G. J. Edwards, and C. J. Taylor. A comparative evaluation of active appearance model algorithms. In *BMVC*, volume 98, pages 680–689, 1998. 1

[12] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(6):681–685, 2001. 1, 2

[13] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59, 1995. 1

[14] T. F. Cootes, G. V. Wheeler, K. N. Walker, and C. J. Taylor. View-based active appearance models. *Image and vision computing*, 20(9):657–664, 2002. 2

[15] D. Cristinacce and T. Cootes. Automatic feature localisation with constrained local models. *Pattern Recognition*, 41(10):3054–3067, 2008. 1

[16] D. Cristinacce and T. F. Cootes. Feature detection and tracking with constrained local models. In *BMVC*, volume 17, pages 929–938, 2006. 2

[17] P. Dollár, P. Welinder, and P. Perona. Cascaded pose regression. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1078–1085. IEEE, 2010. 1, 2

[18] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1627–1645, 2010. 2

[19] L. Gu and T. Kanade. 3D alignment of face in a single image. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 1305–1312. IEEE, 2006. 2

[20] T. Hassner. Viewing real-world faces in 3D. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 3607–3614. IEEE, 2013. 2

[21] T. Hassner, S. Harel, E. Paz, and R. Enbar. Effective face frontalization in unconstrained images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 1, 5, 7

[22] S. Jaiswal, T. R. Almaev, and M. F. Valstar. Guided unsupervised learning of mode specific models for facial point detection in the wild. In *Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on*, pages 370–377. IEEE, 2013. 2, 5

[23] L. A. Jeni, J. F. Cohn, and T. Kanade. Dense 3D face alignment from 2D videos in real-time. In *Automatic Face & Gesture Recognition, 2015. FG'15. 11th IEEE International Conference on*. IEEE, 2015. 2, 8

[24] A. Jourabloo and X. Liu. Pose-invariant 3D face alignment. In *Computer Vision (ICCV), 2015 IEEE International Conference on*. IEEE, 2015. 2, 7

[25] M. Köstinger, 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 *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 2144–2151. IEEE, 2011. 2, 6

[26] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang. Interactive facial feature localization. In *Computer Vision–ECCV 2012*, pages 679–692. Springer, 2012. 2, 5

[27] D. Lee, H. Park, and C. D. Yoo. Face alignment using cascade gaussian process regression trees. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4204–4212, 2015. 1, 2

[28] Z. Liang, S. Ding, and L. Lin. Unconstrained facial landmark localization with backbone-branches fully-

convolutional networks. *arXiv preprint arXiv:1507.03409*, 2015. 2, 3

[29] J. Martens. Deep learning via Hessian-free optimization. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 735–742, 2010. 4

[30] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. XM2VTSDB: The extended M2VTS database. In *Second international conference on audio and video-based biometric person authentication*, volume 964, pages 965–966. Citeseer, 1999. 6

[31] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3D face model for pose and illumination invariant face recognition. In *Advanced Video and Signal Based Surveillance, 2009. AVSS'09. Sixth IEEE International Conference on*, pages 296–301. IEEE, 2009. 3

[32] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3000 FPS via regressing local binary features. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1685–1692. IEEE, 2014. 8

[33] S. Romdhani and T. Vetter. Estimating 3D shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In *Computer Vision and Pattern Recognition (CVPR), 2005 IEEE Conference on*, volume 2, pages 986–993. IEEE, 2005. 2, 5

[34] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on*, pages 397–403. IEEE, 2013. 6, 8

[35] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. A semi-automatic methodology for facial landmark annotation. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*, pages 896–903. IEEE, 2013. 2, 5

[36] J. Saragih and R. Goecke. A nonlinear discriminative approach to AAM fitting. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007. 2

[37] J. M. Saragih, S. Lucey, and J. F. Cohn. Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision*, 91(2):200–215, 2011. 2

[38] B. M. Smith, J. Brandt, Z. Lin, and L. Zhang. Nonparametric context modeling of local appearance for pose-and expression-robust facial landmark localization. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1741–1748. IEEE, 2014. 2

[39] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3476–3483. IEEE, 2013. 2, 3

[40] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 2

[41] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, 2013. 1

[42] G. Tzimiropoulos and M. Pantic. Optimization problems for fast AAM fitting in-the-wild. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 593–600. IEEE, 2013. 1, 2

[43] M. Valstar, B. Martinez, X. Binefa, and M. Pantic. Facial point detection using boosted regression and graph models. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2729–2736. IEEE, 2010. 2

[44] J. Xiao, S. Baker, I. Matthews, and T. Kanade. Real-time combined 2D+3D active appearance models. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 535–542, 2004. 2

[45] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 532–539. IEEE, 2013. 1, 2, 6, 8

[46] X. Xiong and F. De la Torre. Global supervised descent method. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2664–2673, 2015. 2

[47] J. Yan, Z. Lei, D. Yi, and S. Z. Li. Learn to combine multiple hypotheses for accurate face alignment. In *Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on*, pages 392–396. IEEE, 2013. 7

[48] J. Yan, Y. Yu, X. Zhu, Z. Lei, and S. Z. Li. Object detection by labeling superpixels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5107–5116, 2015. 2

[49] X. Yu, J. Huang, S. Zhang, W. Yan, and D. N. Metaxas. Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 1944–1951. IEEE, 2013. 2, 7, 8

[50] J. Zhang, S. Shan, M. Kan, and X. Chen. Coarse-to-fine auto-encoder networks (CFAN) for real-time face alignment. In *Computer Vision–ECCV 2014*, pages 1–16. Springer, 2014. 1, 2

[51] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In *Computer Vision–ECCV 2014*, pages 94–108. Springer, 2014. 1, 2

[52] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin. Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In *Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on*, pages 386–391. IEEE, 2013. 6

[53] Y. Zhou, W. Zhang, X. Tang, and H. Shum. A Bayesian mixture model for multi-view face alignment. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 741–746. IEEE, 2005. 1, 2

[54] S. Zhu, C. Li, C. C. Loy, and X. Tang. Face alignment by coarse-to-fine shape searching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4998–5006, 2015. 1, 2, 8

[55] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li. High-fidelity pose and expression normalization for face recognition in the

wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 787–796, 2015. 1, 2, 5, 6

[56] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2879–2886. IEEE, 2012. 2, 5, 6, 8

[57] X. Zhu, J. Yan, D. Yi, Z. Lei, and S. Z. Li. Discriminative 3D morphable model fitting. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, pages 1–8. IEEE, 2015. 3, 4