FacePoseNet: Making a Case for Landmark-Free Face Alignment

Feng-Ju Chang[1], Anh Tuan Tran[1], Tal Hassner[2,3], Iacopo Masi[1], Ram Nevatia[1], Gerard Medioni[1]
[1] Institute for Robotics and Intelligent Systems, USC, CA, USA
[2] Information Sciences Institute, USC, CA, USA
[3] The Open University of Israel, Israel
{fengjuch,anhttran,iacopoma,nevatia,medioni}@usc.edu, hassner@isi.edu

Abstract

We show how a simple convolutional neural network (CNN) can be trained to accurately and robustly regress 6 degrees of freedom (6DoF) 3D head pose, directly from image intensities. We further explain how this FacePoseNet (FPN) can be used to align faces in 2D and 3D as an alternative to explicit facial landmark detection for these tasks. We claim that in many cases the standard means of measuring landmark detector accuracy can be misleading when comparing different face alignments. Instead, we compare our FPN with existing methods by evaluating how they affect face recognition accuracy on the IJB-A and IJB-B benchmarks: using the same recognition pipeline, but varying the face alignment method. Our results show that (a) better landmark detection accuracy measured on the 300W benchmark does not necessarily imply better face recognition accuracy. (b) Our FPN provides superior 2D and 3D face alignment on both benchmarks. Finally, (c), FPN aligns faces at a small fraction of the computational cost of comparably accurate landmark detectors. For many purposes, FPN is thus a far faster and far more accurate face alignment method than using facial landmark detectors.

1. Introduction

Facial landmark detection is rarely, if ever, an application in its own right. Instead, it is typically a means to an end: It is one component out of many in pipelines designed for other face understanding and processing tasks, often providing effective means for aligning face photos and making them easier to process. Most facial landmark detectors, however, are developed without measuring their impact on these applications but rather using standard facial landmark detection benchmarks such as the popular AFW [53], LFPW [5], HELEN [26], and IBUG [41]. These benchmarks contain face images with manually labeled *ground truth* landmarks. Better detection accuracy on these benchmarks equals better prediction of these manual positions.

Figure 1. *The problem with manually labeled ground truth facial landmarks.* Images and annotations from the AFW [53] (left two columns) and iBug [41] benchmarks. One of each pair shows manually labeled ground truth landmarks; the other, a high-error prediction of our FPN, which does not account for facial expression or 3D shape. Which is which?[1] Clearly, detection accuracy, as measured by standard benchmarks, does not necessarily reflect the quality of the landmark detection.

This raises an important question: *Does better approximation of such human labeled landmarks imply better face alignment and consequently better face understanding?*

Why would higher accuracy on landmark detection benchmarks *not* imply better alignment? The many landmark detection benchmarks used by the community to measure detection accuracy typically offer 5, 49 or 68 landmarks painstakingly labeled on hundreds or thousands of unconstrained face images, reflecting wide viewpoint, resolution and noise variations. On low resolution images, however, even expert human operators can find it hard to accurately pinpoint landmark positions. More importantly, many landmark locations are not well defined even in high resolution (e.g., points along the jawline or behind occlu-

[1] Images one, three, and five are ground truth.

sions). Thus, improved landmark detection accuracy may actually reflect better estimation of uncertain human labels rather than better face alignment (Fig. 1).

An additional concern relates to how landmarks are used for face alignment. Face alignment often implies using a global 2D or 3D transformation to *warp* faces to ideal, reference frames: Detected landmarks are matched with their corresponding landmarks in the reference coordinates and a 2D or 3D transformation is then computed by robust estimation methods. To our knowledge, the effects landmark detection noise, changing expressions or face shapes have on these estimated transformations were never fully explored.

Responding to these concerns, we offer several contributions. (1) We propose comparing landmark detection methods by evaluating bottom line face recognition accuracy on faces aligned with these methods. (2) As an alternative to existing facial landmark detectors, we further present a robust and accurate, landmark-free method for face alignment: our deep FacePoseNet (FPN). We show it to excel at global, 3D face alignment even under the most challenging viewing conditions. Finally, (3), we test our FPN extensively and report that better landmark detection accuracy on the widely used 300W benchmark [40] does *not* imply better alignment and recognition on the highly challenging IJB-A [22] and IJB-B benchmarks [44]. In particular, recognition results on images aligned with our FPN surpass those on images aligned with state-of-the-art detectors.

Some applications require landmark estimation. Our FPN provides a more accurate and far faster face alignment technique in the many cases where global alignment, rather than specific landmark positions, is needed. To support our claims, we make our code publicly available [1].

2. Related work

Applications of facial landmark detectors. Facial landmark detection is big business, as reflected by the numerous citation to relevant papers, the many facial landmark detection benchmarks [5, 23, 26, 40, 53], and popular international events dedicated to this problem. With all this effort, a rigorous survey of the many applications of facial landmarks is outside the scope of this paper. In lieu of such a survey, and to get some idea of why this problem attracts so much attention, we offer the following cursory study.

We consider two of the most widely cited face landmark detector papers of the last decade, the tree based approach [53] and supervised descent method [48]. At the time of writing, based on Google Scholar, the latter accumulated nearly a thousand citations and the former well over a thousand. We found 23 application names appearing frequently (more than ten times) in the titles of the papers that cite these two and counted the number of times these applications were mentioned. The relative frequencies of these applications are reported in Fig. 2.

Of course, this simple survey is by no means accurate: the same term is counted twice if the paper using it in its title cites both [53] and [48] and many paper titles do not clearly state the application they describe (e.g., [14] describes a method for face alignment in 3D but does not mention "alignment" in the title). Nevertheless, with around two thousand papers included in this survey, the result is quite clear: Alignment, face recognition and pose estimation – also considered alignment – are *overwhelmingly more popular* than any other application. This, of course, excluding other landmark detection papers.

What does it mean to align a face? The term *alignment* almost always appears in the titles of papers which present facial landmark detection methods [1, 7, 38] (and most others) implying that the two terms are used interchangeability. This reflects an interpretation of alignment as forming correspondences between particular spatial locations in one face image and another. A different interpretation of *alignment*, and the one used here, refers not only to establishing these correspondences but also to *warping* the two face images, thus making them easier to compare and match. Face warping with estimated 2D (in-plane) or 3D transformations is well known to have a profound impact on the performance of face recognition systems [16, 18].

Although sometimes alignments involve non-parametric or part-based warps [14, 15], often, global 2D or 3D (parametric) transformation are all that is required for this purpose. Such aligned faces are then further processed in systems for face recognition [8, 11, 52], emotion recognition [19, 29], age and gender estimation [12, 13, 28], and more. In fact, it was recently claimed that a global alignment is both more robust and far faster to warp than non-parametric transformations [31, 33]. This paper focuses on such global transformations, showing how they can be estimated quickly and accurately using a deep neural network.

Deep pose estimation. This work describes a deep network trained to estimate the 6DoF of 3D faces viewed in single images. Deep learning is increasingly used for similar purposes, though typically focusing on general object classes [4, 35, 42]. Some recently addressed faces in particular, though their methods are designed to estimate 2D landmarks along with 3D face shapes [20, 25, 51]. Unlike our proposed pose estimation, they regress poses by using iterative methods which involve computationally costly face rendering. We regress 6DoF directly from image intensities without such rendering steps.

In all these cases, absence of training data was cited as a major obstacle for training effective models. In response, some turned to larger 3D object data sets [47, 46] or using synthetically generated examples [39]. We propose a far simpler alternative and show it to result in robust and

[1] https://github.com/fengju514/Face-Pose-Net

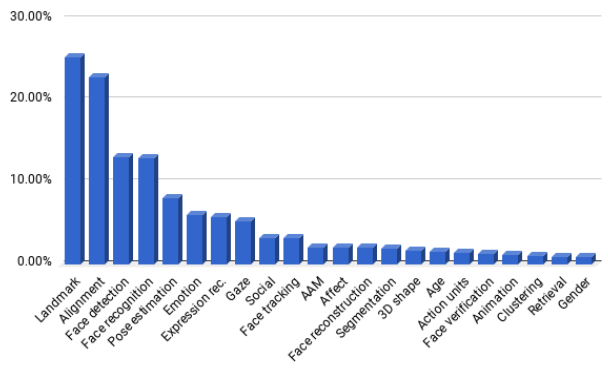


Figure 2. *Applications of facial landmarks.* Illustrating the frequency of various task and application names in paper titles citing two of the most popular landmark detectors [53] and [48].

accurate face alignment.

3. A critique of facial landmark detection

Before using an existing state-of-the-art facial landmark detector in a face processing system, the following points should be considered.

Landmark detection accuracy measures. Facial landmark detection accuracy is typically measured by considering the distances between estimated landmarks and ground truth (reference) landmarks, normalized by the reference inter-ocular distance of the face [10]:

$$e(\mathbf{L}, \hat{\mathbf{L}}) = \frac{1}{m\|\hat{\mathbf{p}}_l - \hat{\mathbf{p}}_r\|_2} \sum_{i=1}^{m} \|\mathbf{p_i} - \hat{\mathbf{p}}_\mathbf{i}\|_2, \qquad (1)$$

Here, $\mathbf{L} = \{\mathbf{p_i}\}$ is the set of m 2D facial landmark coordinates, $\hat{\mathbf{L}} = \{\hat{\mathbf{p}}_\mathbf{i}\}$ their ground truth locations, and $\hat{\mathbf{p}}_l, \hat{\mathbf{p}}_r$ the reference left and right eye outer corner positions. These errors are then translated to a number of standard quantities, including the mean error rate (MER), the percentage of landmarks detected under certain error thresholds (e.g., below 5% or 10% error rates) or the area under the accumulative error curve (AUC).

There are two key problems with this method of evaluating landmark errors. First, the ground truth compared against is manually specified, often by mechanical turk workers. These manual annotations can be noisy, they are ill-defined when images are low resolution, the landmarks are occluded (in case of large out-of-plane head rotations, facial hair and other obstructions), or located in featureless facial regions (e.g., along the jawline). Accurate facial landmark detection, as measured on these benchmarks, thus implies better matching human labels but not necessarily better detection. These problems are demonstrated in Fig. 1.

A second potential problem lies in the error measure itself: Normalizing detection errors by inter-ocular distances biases against images of faces appearing at non-frontal views. When faces are near profile, perspective projection of the 3D face onto the image plane shrinks the distances between the eyes thereby naturally inflating the errors computed for such images.

Landmark detection speed. Some facial landmark detection methods emphasize impressive speeds [21, 38]. Measured on standard landmark detection benchmarks, however, these methods do not necessarily claim state-of-the-art accuracy, falling behind more sophisticated, yet far slower detectors [50]. Moreover, aside from [51], no existing landmark detector is designed to take advantage of GPU hardware, a standard feature in commodity computer systems and most, including [51], apply iterative optimizations which may be hard to convert to parallel processing.

Effects of facial expression and shape on alignment. It was recently shown that 3D alignment and warping of faces to frontal viewpoints (i.e. *frontalization*) is effective regardless of the precise 3D face shape used for this purpose [16]. Facial expressions and 3D shapes in particular, appear to have little impact on the warped result as evident by the improved face recognition accuracy reported by that method. Moreover, it was recently demonstrated that by using such a generic 3D face shape, rendering faces from new viewpoints can be accelerated to the same speed as simple 2D image warping [31].

Interestingly, they and many others used facial landmark detectors to compute parametric transformations – projection matrix [16] or 2D affine or similarity transforms [12, 18] – by applying robust estimators to corresponding detected facial landmarks [14, 27]. Variations in landmark locations due to expressions and face shapes essentially *contribute noise* to this estimation process. The effects these variations have on the quality of the alignment were, as far as we know, never truly studied.

4. Deep, direct head pose regression

Rather than align faces using landmark detection, we refer to alignment as a global, 6DoF 3D face pose, and propose to infer it directly from image intensities, using a simple deep network architecture. We next describe the network and the novel method used to train it.

4.1. Head pose representation

We define face alignment as the 3D head pose $\mathbf{h}$, expressed using 6DoF: three for rotations, $\mathbf{r} = (r_x, r_y, r_z)^T$, and three for translations, $\mathbf{t} = (t_x, t_y, t_z)^T$:

$$\mathbf{h} = (r_x, r_y, r_z, t_x, t_y, t_z)^T \qquad (2)$$

where (r_x, r_y, r_z) are represented as Euler angles (pitch, yaw, and roll). Given m 2D facial landmark coordinates on an input image, $\mathbf{p}_{2\times m}$, and their corresponding, reference 3D coordinates, $\mathbf{P}_{3\times m}$ – selected on a fixed, generic

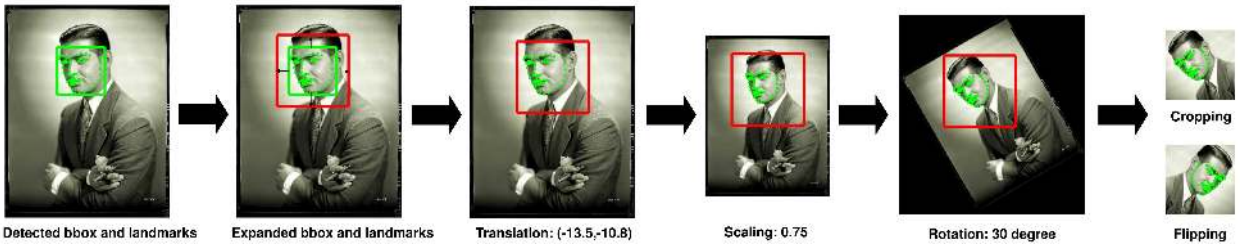


Figure 3. *Augmenting appearances of images from the VGG face dataset [34].* After detecting the face bounding box and landmarks we augment its appearance by applying a number of simple planar transformations, including translation, scaling, rotation, and flipping. The same transformations are applied to the landmarks, thereby producing example landmarks for images which may be too challenging for existing landmark detectors to process.

3D face model – we can obtain a 3D to 2D projection of the 3D landmarks onto the 2D image by solving the following equation for the standard pinhole model:

$$[\mathbf{p}, \mathbf{1}]^T = \mathbf{A}[\mathbf{R}, \mathbf{t}][\mathbf{P}, \mathbf{1}]^T, \tag{3}$$

where $\mathbf{A}$ and $\mathbf{R}$ are the camera matrix and rotation matrix respectively and $\mathbf{1}$ is a constant vector of 1. We then extract a rotation vector $\mathbf{r} = (r_x, r_y, r_z)^T$ from $\mathbf{R}$ using the Rodrigues rotation formula:

$$\mathbf{R} = \cos\theta\mathbf{I} + (1-\cos\theta)\mathbf{r}\mathbf{r}^T + \sin\theta \begin{pmatrix} 0 & -r_z & r_y \\ r_z & 0 & -r_x \\ -r_y & r_x & 0 \end{pmatrix},$$

where we define $\theta = ||\mathbf{r}||_2$.

Obtaining enough training examples. Although our network architecture is not very deep compared to deep networks used today for other tasks, training it still requires large quantities of labeled training data. We found the numbers of facial landmark annotated faces in standard data sets to be too small for this purpose. A key problem is therefore obtaining a large enough training set.

We produce our training set by synthesizing 6D, ground truth pose labels by running an existing facial landmark detector [3] on a large image set: the 2.6 million images in the VGG face dataset [34]. The detected landmarks were then used to compute the 6DoF labels for the images in this set. A potential danger in using an existing method to produce our training labels, is that our CNN will not improve beyond the accuracy of its training labels. As we show in Sec. 5, this is not necessarily the case.

To further improve the robustness of our CNN, we apply a number of face augmentation techniques to the images in the VGG face set, substantially enriching the appearance variations it provides. Fig. 3 illustrates this augmentation process. Specifically, following face detection [49] and landmark detection [3], we transform detected bounding boxes and their detected facial landmarks using a number of simple in-plane transformations. The parameters for these transformations are selected randomly from fixed distributions (Table. 1). The transformed faces are then used for training, along with their horizontally mirrored versions, to provide yaw rotation invariance. Ground truth labels are, of course, computed using the transformed landmarks.

Table 1. *Summary of augmentation transformation parameters used to train our FPN.* Where $\mathcal{U}(a, b)$ samples from a uniform distribution ranging from a to b and $\mathcal{N}(\mu, \sigma^2)$ samples from a normal distribution with mean μ and variance σ^2. $width$ and $height$ are the face detection bounding box dimensions.

Transformation	Range
Horizontal translation	$\mathcal{U}(-0.1, 0.1) \times width$
Vertical translation	$\mathcal{U}(-0.1, 0.1) \times height$
Scaling	$\mathcal{U}(0.75, 1.25)$
Rotation (degrees)	$30 \times \mathcal{N}(0, 1)$

Some example augmented faces are provided in Fig. 4. Note that augmented images would often be *too challenging for existing landmark detectors*, due to extreme rotations or scaling. This, of course, *does not affect the accuracy of the ground truth labels* which were obtained from the original images. It does, however, force our CNN to learn to estimate poses even on such challenging images.

FPN training. For our FPN we use an AlexNet architecture [24] with its initialized weights provided by [32]. The only difference is that here the output regresses 6D floating point values rather than predicts one-hot encoded, multi class labels. Note that during training each dimension of the head pose labels is normalized by the corresponding mean and standard deviation of the training set, compensating for the large value differences among dimensions. The same normalization parameters are used at test time.

2D and 3D face alignment with FPN. Given a test image, it is processed by applying the same face detector [49], cropping the face and scaling it to the dimension of the network's input layer. The 6D network output is then converted to a projection matrix. Specifically, the projection matrix is produced by the camera matrix $\mathbf{A}$, rotation matrix $\mathbf{R}$, and the translation vector $\mathbf{t}$ in Eq. (3). With this projection matrix we can render new views of the face, aligning it across 3D views as was recently proposed by others [33, 31].

For 2D alignment, we compute the 2D similarity transform to warp the 2D projected landmarks to pre-defined

Figure 4. *Example augmented training images.* Example images from the VGG face data set [34] following data augmentation. Each triplet shows the original detected bounding box (left) and its augmented versions (mirrored across the vertical axis). Both flipped versions were used for training FPN. Note that in some cases, detecting landmarks would be highly challenging on the augmented face, due to severe rotations and scalings not normally handled by existing methods. Our FPN is trained with the original landmark positions, transformed to the augmented image coordinate frame.

landmark locations. With frontal images (absolute yaw angle $\leq 30°$), we use the eye centers, the nose tip, and the mouth corners for alignment. With profile images (absolute yaw angle $> 30°$), however, only the visible eye center and the nose tip are used.

5. Results

We provide comparisons of our FPN with the following widely used, state-of-the-art, facial landmark detection methods: Dlib [21], CLNF [2], OpenFace [3], DCLM [50], RCPR [6], and 3DDFA [51] evaluating them for their effects on face recognition vs. their landmark detection accuracy.

5.1. Effect of alignment on recognition

Sec. 3 discusses the various potential problems of comparing face alignment methods by measuring their landmark detection accuracy. As an alternative, we propose comparing methods for face alignment and landmark detection by evaluating their effect on the bottom line accuracy of a face processing pipeline. Since face recognition is arguably one of the most popular applications for face alignment, we use recognition accuracy as a performance measure. To our knowledge, this is the first time alignment methods are compared based on their effect on recognition accuracy.

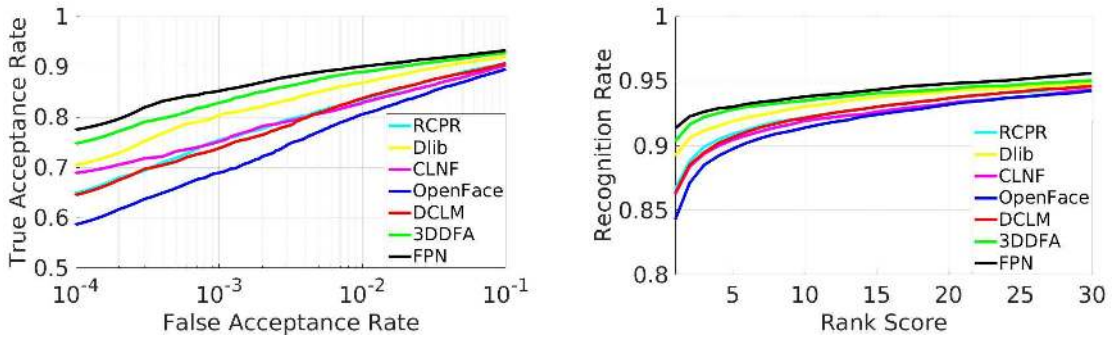
Specifically, we use two of the most recent benchmarks for face recognition: IARPA Janus Benchmark A [22] and B [44] (IJB-A and IJB-B). Importantly, these benchmarks were designed with the specific intention of elevating the difficulty of face recognition. This heightened challenge is reflected by, among other factors, an unprecedented amount of extreme out of plane rotated faces including many appearing in near-profile views [33]. As a consequence, these two benchmarks not only push the limits of face recognition systems, but also the alignment methods used by these systems, possibly more so than the faces in standard facial landmark detection benchmarks.

Face recognition pipeline. We employ a system similar to the one recently proposed by [31, 33], building on their publicly available ResFace101 model and related code. We chose this system, as it explicitly aligns faces to multiple viewpoints, including rendering novel views. These steps are highly dependent on the quality of alignment and so its recognition accuracy should reflect alignment accuracy. In practice, we used their 2D (similarity transform) and 3D (new view rendering) code directly, changing how the transformations are computed: our tests compare different landmark detectors used to recover the 6DoF head pose required by their warping and rendering method, with the 6DoF regressed using our FPN.

Their system uses a single Convolutional Neural Network (CNN), a ResNet-101 architecture [17], trained on both real face images and synthetic, rendered views. We

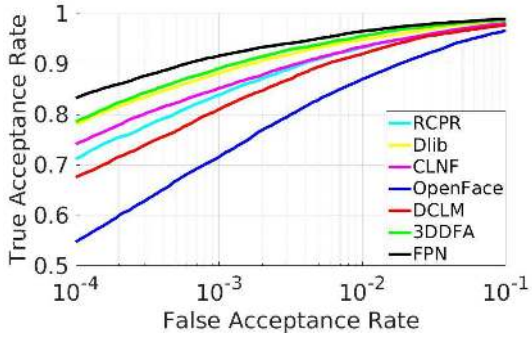
Table 2. *Verification and identification results on IJB-A and IJB-B,* comparing landmark detection based face alignment methods. Three baseline IJB-A results are also provided as reference at the top of the table. * State-of-the-art method which uses meta data seed landmarks and face bounding boxes; all others did not. ** Numbers estimated from the ROC and CMC in [44].

Method ↓	**TAR@FAR**			**Identification Rate (%)**			
Eval. →	.01%	0.1%	1.0%	Rank-1	Rank-5	Rank-10	Rank-20
IJB-A [22]							
Crosswhite et al. [9]	–	–	93.9	92.8	–	98.6	–
Ranjan et al. [37]	–	82.3	92.2	94.7	–	98.8	–
Masi et al. [31]*	56.4	75.0	88.8	92.5	96.6	97.4	98.0
RCPR [6]	64.9	75.4	83.5	86.6	90.9	92.2	93.7
Dlib [21]	70.5	80.4	86.8	89.2	91.9	93.0	94.2
CLNF [2]	68.9	75.1	82.9	86.3	90.5	91.9	93.3
OpenFace [3]	58.7	68.9	80.6	84.3	89.8	91.4	93.2
DCLM [50]	64.5	73.8	83.7	86.3	90.7	92.2	93.7
3DDFA [51]	74.8	82.8	89.0	90.3	92.8	93.5	94.4
Our FPN	**77.5**	**85.2**	**90.1**	**91.4**	**93.0**	**93.8**	**94.8**
IJB-B [44]							
GOTs [44]**	16.0	33.0	60.0	42.0	57.0	62.0	68.0
VGG face [44]**	55.0	72.0	86.0	78.0	86.0	89.0	92.0
RCPR [6]	71.2	83.8	93.3	83.6	90.9	93.2	95.0
Dlib [21]	78.1	88.2	94.8	88.0	93.2	94.9	96.3
CLNF [2]	74.1	85.2	93.4	84.5	90.9	93.0	94.8
OpenFace [3]	54.8	71.6	87.0	74.3	84.1	87.8	90.9
DCLM [50]	67.6	81.0	92.0	81.8	89.7	92.0	94.1
3DDFA [51]	78.5	89.1	95.6	89.0	94.1	95.5	96.9
Our FPN	**83.2**	**91.6**	**96.5**	**91.1**	**95.3**	**96.5**	**97.5**

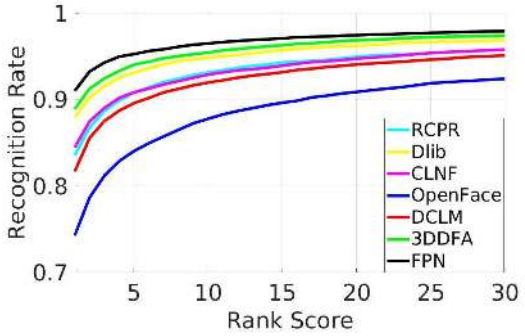


(a) ROC IJB-A

(b) CMC IJB-A



(c) ROC IJB-B



(d) CMC IJB-B

Figure 5. *Verification and identification results on IJB-A and IJB-B.* ROC and CMC curves accompanying the results reported in Table 2.

fine tune the ResFace101 CNN using $L2$-constrained Softmax Loss [36] instead of the original softmax used by Masi et al. for their publicly released model. This fine tuning is performed using the MS-Celeb face set [30] as an example set. Aside from this change, we use the same recognition pipeline from [31] and we refer to that paper for details.

Bounding box detection. We emphasize that an identical pipeline was used with the different alignment methods; different results vary only in the method used to estimate facial pose. The only other difference between recognition pipelines was in the facial bounding box detector.

Facial landmark detectors are sensitive to the face detector they are used with. We therefore report results obtained when running landmark detectors with the best bounding boxes we were able to determine. Specifically, FPN was applied to the bounding boxes returned by the detector of Yang and Nevatia [49], following expansion of its dimensions by 25%. Most detectors performed best when applied using the same face detector, without the 25% increase. Finally, 3DDFA [51] was tested with the same face detector followed by the face box expansion code provided by its authors.

Face verification and identification results. Face verification and identification results on both IJB-A and IJB-B are provided in Table 2. We report multiple recognition metrics for both verification and identification: For verification, these measure the recall (True Acceptance Rate) at three cut-off points of the False Alarm Rate (TAR-$\{1\%,0.1\%,0.01\%\}$). For identification we provide recognition rates at four ranks from the CMC (Cumulative Matching Characteristic). The overall performances in terms of ROC and CMC curves are shown in Fig. 5. The table also provides, as reference, three state-of-the-art IJB-A results [9, 31, 37] and baseline results from [44] for IJB-B (to our knowledge, we are the first to report verification and identification accuracies on IJB-B).

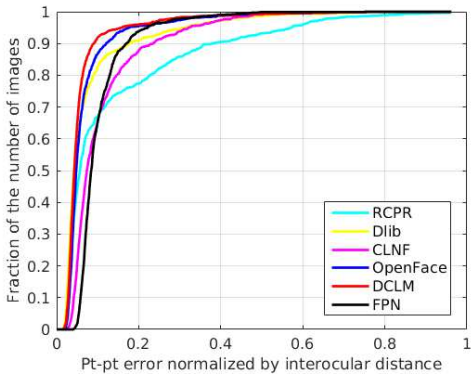
Faces aligned with our FPN offer higher recognition rates, even compared to the most recent, state-of-the-art facial landmark detection method of [50]. In fact, our verification scores on IJB-A outperform the reported those of the system used as the basis for our own [31]. Their recognition results are higher than ours, but importantly, they used ground truth annotations to initialize landmark detection search. This allowed them to correctly align faces in images where face landmark detectors would normally fail, explaining their higher recognition results. These annotations were not used by any of the other methods compared.

5.2. Landmark detection accuracy

From 6DoF pose to facial landmarks. Given a 6DoF head pose estimate, facial landmarks can then be estimated and compared with existing landmark detection methods for their accuracy on standard benchmarks. To obtain landmark predictions, 3D reference coordinates of facial landmarks

Method	$\leq 5\%$	$\leq 10\%$	$\leq 20\%$	$\geq 40\%$	MER	Sec./im.
RCPR [6]	44.44 %	66.96 %	77.39 %	9.55 %	0.1386	0.05
Dlib [21]	60.03 %	82.65 %	90.94 %	2.83 %	0.0795	2.26
CLNF [2]	20.86 %	65.11 %	87.62 %	2.63 %	0.1106	0.64
OpenFace [3]	54.39 %	86.74 %	95.42 %	1.27 %	0.0702	0.64
DCLM [50]	64.91 %	91.91 %	96.00 %	1.17 %	0.0611	16.2
3DDFA [51]	N/A	N/A	N/A	N/A	N/A	0.6
Our FPN	1.75 %	65.40 %	93.86 %	0.97 %	0.1043	0.003

(a) Quantitative results



(b) Acumulative error curves

Figure 6. *68 point detection accuracies on 300W.* (a) The percent of images with 68 landmark detection errors lower than 5%, 10%, and 20% inter-ocular distances, or greater than 40%, mean error rates (MER) and runtimes. Our FPN was tested using a GPU. On the CPU, FPN runtime was 0.07 seconds. 3DDFA used the AFW collection for training. Code provided for 3DDFA [51] did not allow testing on the GPU; in their paper, they claim GPU runtime to be 0.076 seconds. As AFW was included in our 300W test set, landmark detection accuracy results for 3DDFA were excluded from this table. (b) Accumulative error curves.

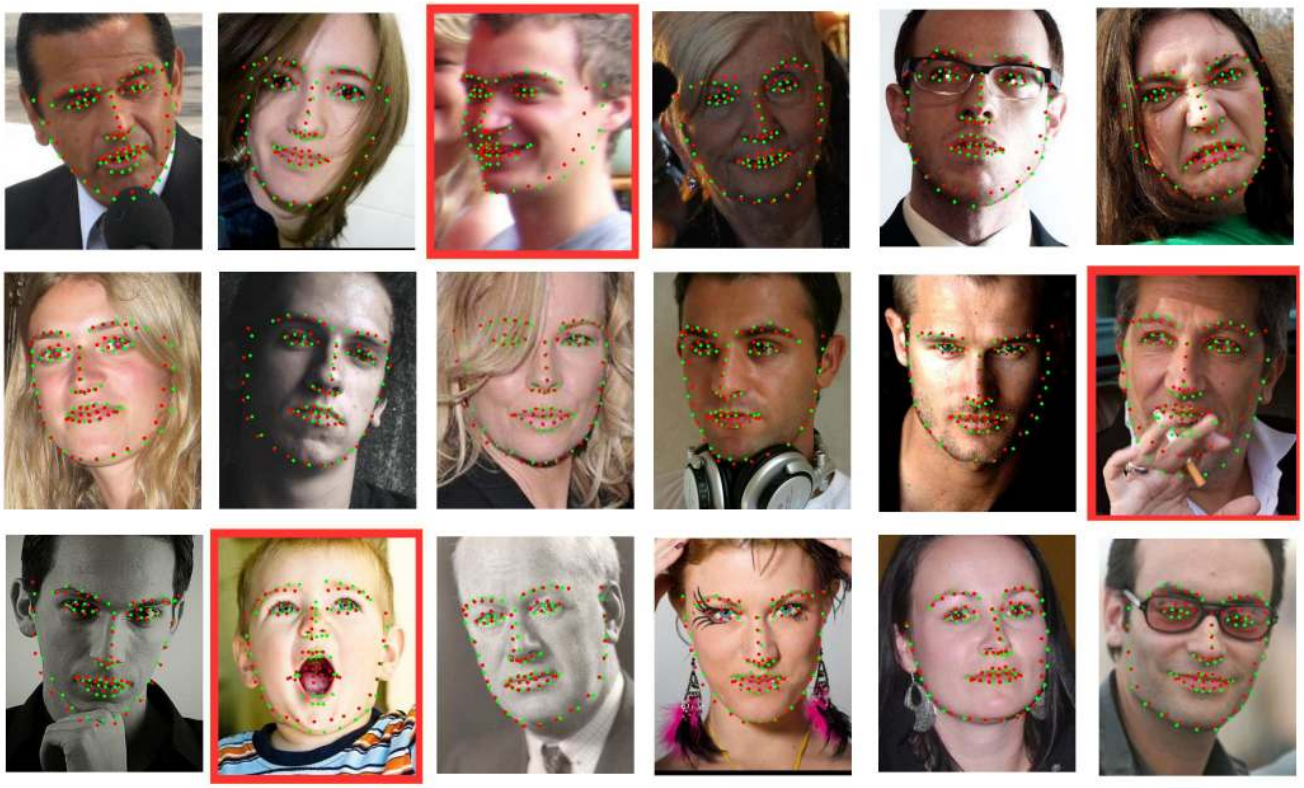


Figure 7. *Qualitative landmark detection examples.* Landmarks detected in 300W [40] images by projecting an unmodified 3D face shape, pose aligned using our FPN (red) vs. ground truth (green). The images marked by the red-margin are those which had large FPN errors ($> 10\%$ inter-ocular distance). These appear perceptually reasonable, despite these errors. The mistakes in the red-framed example on the third row was clearly a result of our FPN not representing expressions.

are selected off line once on the same *generic*, 3D face model used in [31]. Given a pose estimate, we convert it to a projection matrix and project these 3D landmarks down to the input image.

Recently, a similar process was proposed for accurate landmark detection across large poses [51]. In their work, an iterative method was used to simultaneously estimate a 3D face shape, including facial expression, and project its landmarks down to the input image. Unlike them, our tests use a single generic 3D face model, unmodified. By not iter-

ating over the face shape, our method is simpler and faster, but of course, our predicted landmarks will not reflect different 3D shapes and facial expressions. We next evaluate the effect this has on landmark detection accuracy.

Detection accuracy on the 300W benchmark. We evaluate performance on the 300W data set [40], the most challenging benchmark of its kind [45], using 68 landmarks. We note that we did not use the standard training sets used with the 300W benchmark (e.g., the HELEN [26] and LFPW [5] training sets with their manual annotations). Instead we trained FPN with the estimated landmarks, as explained in Sec. 4.1. As a test set, we used the standard union consisting of the LFPW test set (224 images), the HELEN test set (330), AFW [53] (337), and IBUG [41] (135). These 1026 images, collectively, form the 300W test set. Note that unlike others, we did not use AFW to train our method, allowing us to use it for testing.

Fig. 6 (a) reports five measures of accuracy for the various methods tested: The percent of images with 68 landmark detection errors lower than 5%, 10%, and 20% interocular distances, and the mean error rate (MER), averaging Eq. (1) over the images tested. Fig. 6 (b) additionally provides accumulative error curves for these methods.

Not surprisingly, without accounting for face shapes and expressions, our predicted landmarks are not as accurate as those predicted by methods which are influenced by these factors. Some qualitative detection examples are provided in Fig. 7 including a few errors larger than 10%. These show that mistakes can often be attributed to FPN not modeling facial expressions and shape. One way to improve this would be to use a single-view 3D face shape estimation method [15, 43] to better approximate landmark positions, though we have not tested this here.

Detection runtime. In one tested measure FPN far outperforms its alternatives: The last column of Fig. 6 (a) reports the mean, per-image runtime for landmark detection. *Our FPN is an order of magnitude faster than any other face alignment method*. This is true even compared to the GPU runtimes reported for 3DDFA in their paper [51].

All methods were tested using an NVIDIA, GeForce GTX TITAN X, 12GB RAM, and an Intel(R) Xeon(R) CPU E5-2640 v3 @ 2.60GHz, 132GB RAM. The only exception was 3DDFA [51], which required a Windows system and was tested using an Intel(R) Core(TM) i7-4820K CPU @ 3.70GHz (8 CPUs), 16GB RAM, running 8 Pro 64-bit.

5.3. Discussion

Landmarks predicted using FPN in Sec. 5.2 were less accurate than those estimated by other methods. How does that agree with the better face recognition results obtained with images aligned using FPN? As we mentioned in Sec. 3 better accuracy on a face landmark detection benchmark reflects many things which are not necessarily important when aligning faces for recognition. These include, in particular face shapes and expressions, the latter can actually cause misalignments when computing face pose and warping the face accordingly. FPN, on the other hand, ignores these factors, instead providing a 6DoF pose estimates at breakneck speeds, directly from image intensities.

An important observation is that despite being trained with labels generated by OpenFace [3], recognition results on faces aligned with FPN are *better* than those aligned with OpenFace. This can be explained in a number of ways: First, FPN was trained on appearance variations introduced by augmentation, which OpenFace was not necessarily designed to handle. Second, poses estimated by FPN were less corrupted by expressions and facial shapes, making the warped images better aligned. Third, as was recently argued by others [43], CNNs are remarkably adapt at training with label noise such as any errors in the poses predicted by OpenFace for the ground truth labels. Finally, CNNs are highly capable of domain shifts to new data, such as the extremely challenging views of the faces in IJB-A and IJB-B.

6. Conclusions

For many practical purposes, face alignment requires only global, parametric 2D or 3D transformations. This is often the case in state-of-the-art face recognition pipelines and a wide variety of other face understanding tasks. In such circumstances, accurate facial landmark detection is superfluous and its potential for introducing errors whenever facial expressions and shapes are not explicitly considered was never fully explored. In this paper we present an alternative method for aligning faces: using a simple CNN, uniquely trained to regress 6DoF face pose, directly from image intensities. We show that by using a GPU, this leads to staggering alignment speeds. Moreover, by comparing alignment methods by considering bottom line performance of a face recognition system, rather than landmark detection accuracy, we show that this simple method outperforms state-of-the-art alignment techniques in the face recognition accuracy it provides.

Acknowledgments

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA 2014-14071600011. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purpose notwithstanding any copyright annotation thereon.

References

[1] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Incremental face alignment in the wild. In *Proc. Conf. Comput. Vision Pattern Recognition*. IEEE, 2014.

[2] T. Baltrusaitis, P. Robinson, and L.-P. Morency. Constrained local neural fields for robust facial landmark detection in the wild. In *Proc. Conf. Comput. Vision Pattern Recognition Workshops*, pages 354–361. IEEE, 2013.

[3] T. Baltrušaitis, P. Robinson, and L.-P. Morency. Openface: an open source facial behavior analysis toolkit. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–10, 2016.

[4] A. Bansal, B. Russell, and A. Gupta. Marr revisited: 2D-3D alignment via surface normal prediction. In *Proc. Conf. Comput. Vision Pattern Recognition*, pages 5965–5974, 2016.

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

[6] X. P. Burgos-Artizzu, P. Perona, and P. Dollár. Robust face landmark estimation under occlusion. In *Proc. Int. Conf. Comput. Vision*, pages 1513–1520. IEEE, 2013.

[7] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. *Int. J. Comput. Vision*, 107(2):177–190, 2014.

[8] B.-C. Chen, C.-S. Chen, and W. H. Hsu. Cross-age reference coding for age-invariant face recognition and retrieval. In *European Conf. Comput. Vision*, pages 768–783. Springer, 2014.

[9] N. Crosswhite, J. Byrne, C. Stauffer, O. Parkhi, Q. Cao, and A. Zisserman. Template adaptation for face verification and identification. In *Int. Conf. on Automatic Face and Gesture Recognition*, pages 1–8. IEEE, 2017.

[10] M. Dantone, J. Gall, G. Fanelli, and L. Van Gool. Real-time facial feature detection using conditional regression forests. In *Proc. Conf. Comput. Vision Pattern Recognition*, pages 2578–2585. IEEE, 2012.

[11] C. Ding, J. Choi, D. Tao, and L. S. Davis. Multi-directional multi-level dual-cross patterns for robust face recognition. *Trans. Pattern Anal. Mach. Intell.*, 38(3):518–531, 2016.

[12] E. Eidinger, R. Enbar, and T. Hassner. Age and gender estimation of unfiltered faces. *Trans. on Inform. Forensics and Security*, 9(12), 2014.

[13] F. Gurpinar, H. Kaya, H. Dibeklioglu, and A. Salah. Kernel elm and cnn based facial age estimation. In *Proc. Conf. Comput. Vision Pattern Recognition Workshops*, pages 80–86, 2016.

[14] T. Hassner. Viewing real-world faces in 3D. In *Proc. Int. Conf. Comput. Vision*, pages 3607–3614. IEEE, 2013. Available: www.openu.ac.il/home/hassner/projects/poses.

[15] T. Hassner and R. Basri. Single view depth estimation from examples. *arXiv preprint arXiv:1304.3915*, 2013.

[16] T. Hassner, S. Harel, E. Paz, and R. Enbar. Effective face frontalization in unconstrained images. In *Proc. Conf. Comput. Vision Pattern Recognition*, 2015.

[17] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. Conf. Comput. Vision Pattern Recognition*, June 2016.

[18] G. B. Huang, V. Jain, and E. Learned-Miller. Unsupervised joint alignment of complex images. In *Proc. Int. Conf. Comput. Vision*, pages 1–8. IEEE, 2007.

[19] X. Huang, Q. He, X. Hong, G. Zhao, and M. Pietikainen. Improved spatiotemporal local monogenic binary pattern for emotion recognition in the wild. In *Int. Conf. on Multimodal Interaction*, pages 514–520. ACM, 2014.

[20] A. Jourabloo and X. Liu. Large-pose face alignment via cnn-based dense 3D model fitting. In *Proc. Conf. Comput. Vision Pattern Recognition*, 2016.

[21] D. E. King. Dlib-ml: A machine learning toolkit. *J. Mach. Learning Research*, 10(Jul):1755–1758, 2009.

[22] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, M. Burge, and A. K. Jain. Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark-A. In *Proc. Conf. Comput. Vision Pattern Recognition*, 2015.

[23] 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 *Proc. Int. Conf. Comput. Vision Workshops*. IEEE, 2011.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Neural Inform. Process. Syst.*, 2012.

[25] A. Kumar, A. Alavi, and R. Chellappa. KEPLER: keypoint and pose estimation of unconstrained faces by learning efficient H-CNN regressors. In *Automatic Face and Gesture Recognition*, pages 258–265, May 2017.

[26] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. Huang. Interactive facial feature localization. *European Conf. Comput. Vision*, pages 679–692, 2012.

[27] V. Lepetit, F. Moreno-Noguer, and P. Fua. Epnp: An accurate o (n) solution to the pnp problem. *International journal of computer vision*, 81(2):155, 2009.

[28] G. Levi and T. Hassner. Age and gender classification using convolutional neural networks. In *Proc. Conf. Comput. Vision Pattern Recognition Workshops*, June 2015.

[29] G. Levi and T. Hassner. Emotion recognition in the wild via convolutional neural networks and mapped binary patterns. In *Int. Conf. on Multimodal Interaction*, pages 503–510. ACM, 2015.

[30] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proc. Int. Conf. Comput. Vision*, 2015.

[31] I. Masi, T. Hassner, A. T. Tran, and G. Medioni. Rapid synthesis of massive face sets for improved face recognition. In *Int. Conf. on Automatic Face and Gesture Recognition*, pages 604–611. IEEE, 2017.

[32] I. Masi, S. Rawls, G. Medioni, and P. Natarajan. Pose-aware face recognition in the wild. In *Proc. Conf. Comput. Vision Pattern Recognition*, pages 4838–4846, 2016.

[33] I. Masi, A. Tran, T. Hassner, J. T. Leksut, and G. Medioni. Do We Really Need to Collect Millions of Faces for Effective Face Recognition? In

European Conf. Comput. Vision, 2016. Available www.openu.ac.il/home/hassner/projects/augmented_faces.

[34] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *Proc. British Mach. Vision Conf.*, 2015.

[35] P. Poirson, P. Ammirato, C.-Y. Fu, W. Liu, J. Kosecka, and A. C. Berg. Fast single shot detection and pose estimation. In *Int. Conf. on 3D Vision*, pages 676–684. IEEE, 2016.

[36] R. Ranjan, C. D. Castillo, and R. Chellappa. L2-constrained softmax loss for discriminative face verification. *arXiv preprint arXiv:1703.09507*, 2017.

[37] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa. An all-in-one convolutional neural network for face analysis. In *Int. Conf. on Automatic Face and Gesture Recognition*, pages 17–24. IEEE, 2017.

[38] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3000 fps via regressing local binary features. In *Proc. Conf. Comput. Vision Pattern Recognition*. IEEE, 2014.

[39] E. Richardson, M. Sela, R. Or-El, and R. Kimmel. Learning detailed face reconstruction from a single image. *arXiv preprint arXiv:1611.05053*, 2016.

[40] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: Database and results. *Image and Vision Computing*, 2015.

[41] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proc. Conf. Comput. Vision Pattern Recognition Workshops*, 2013.

[42] H. Su, C. R. Qi, Y. Li, and L. J. Guibas. Render for CNN: Viewpoint estimation in images using CNNs trained with rendered 3D model views. In *Proc. Int. Conf. Comput. Vision*, pages 2686–2694, 2015.

[43] A. Tran, T. Hassner, I. Masi, and G. Medioni. Regressing robust and discriminative 3D morphable models with a very deep neural network. In *Proc. Conf. Comput. Vision Pattern Recognition*, 2017.

[44] C. Whitelam, E. Taborsky, A. Blanton, B. Maze, J. Adams, T. Miller, N. Kalka, A. K. Jain, J. A. Duncan, K. Allen, et al. Iarpa janus benchmark-b face dataset. In *Proc. Conf. Comput. Vision Pattern Recognition Workshops*, 2017.

[45] Y. Wu, T. Hassner, K. Kim, G. Medioni, and P. Natarajan. Facial landmark detection with tweaked convolutional neural networks. *arXiv preprint arXiv:1511.04031*, 2015.

[46] Y. Xiang, W. Kim, W. Chen, J. Ji, C. Choy, H. Su, R. Mottaghi, L. Guibas, and S. Savarese. Objectnet3D: A large scale database for 3D object recognition. In *European Conf. Comput. Vision*, pages 160–176. Springer, 2016.

[47] Y. Xiang, R. Mottaghi, and S. Savarese. Beyond pascal: A benchmark for 3D object detection in the wild. In *Winter Conf. on App. of Comput. Vision*, pages 75–82. IEEE, 2014.

[48] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *Proc. Conf. Comput. Vision Pattern Recognition*. IEEE, 2013.

[49] Z. Yang and R. Nevatia. A multi-scale cascade fully convolutional network face detector. In *ICPR*, pages 633–638, 2016.

[50] A. Zadeh, T. Baltrušaitis, and L.-P. Morency. Deep constrained local models for facial landmark detection. *arXiv preprint arXiv:1611.08657*, 2016.

[51] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Li. Face alignment across large poses: A 3D solution. In *Proc. Conf. Comput. Vision Pattern Recognition*, Las Vegas, NV, June 2016.

[52] 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 *Proc. Conf. Comput. Vision Pattern Recognition*, pages 787–796, 2015.

[53] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Proc. Conf. Comput. Vision Pattern Recognition*, pages 2879–2886. IEEE, 2012.