

Pose-Invariant Face Alignment with a Single CNN

Amin Jourabloo¹, Mao Ye², Xiaoming Liu¹, and Liu Ren²

¹Department of Computer Science and Engineering, Michigan State University, MI

²Bosch Research North America, Palo Alto, CA

^{1,2} {jourablo, liuxm}@msu.edu, {mao.ye2, liu.ren}@us.bosch.com

Abstract

Face alignment has witnessed substantial progress in the last decade. One of the recent focuses has been aligning a dense 3D face shape to face images with large head poses. The dominant technology used is based on the cascade of regressors, e.g., CNNs, which has shown promising results. Nonetheless, the cascade of CNNs suffers from several drawbacks, e.g., lack of end-to-end training, hand-crafted features and slow training speed. To address these issues, we propose a new layer, named visualization layer, which can be integrated into the CNN architecture and enables joint optimization with different loss functions. Extensive evaluation of the proposed method on multiple datasets demonstrates state-of-the-art accuracy, while reducing the training time by more than half compared to the typical cascade of CNNs. In addition, we compare across multiple CNN architectures, all with the visualization layer, to further demonstrate the advantage of its utilization.

1. Introduction

Face alignment, also known as face landmark detection, is an essential process for many facial analysis tasks, such as face recognition [36], expression estimation [1] and 3D face reconstruction [20, 30]. During the last decade, face alignment technologies have been substantially improved [8–10, 32, 41]. One recent advancement in this area is to tackle challenging cases with large face poses, e.g., profile views with $\pm 90^\circ$ yaw angles [18, 19, 23, 24, 47, 50].

The dominant technology for large-pose face alignment (LPFA) utilizes a cascade of regressors which combines different types of regression designs [23, 49, 50] with feature extraction methods [18]. At each stage of this procedure, the target parameters, e.g., the 2D landmarks or the head pose, as well as the 3D face shape, are refined by regress-

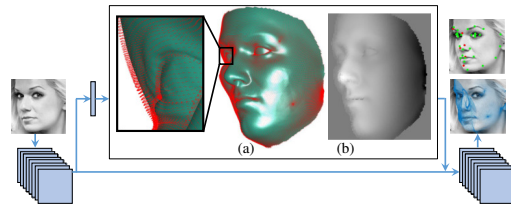


Figure 1. For the purpose of learning an end-to-end face alignment model, our novel visualization layer reconstructs the 3D face shape (a) from the estimated parameters inside the CNN and synthesizes a 2D image (b) via the surface normal vectors of visible vertices.

ing an update of these parameters. Due to the proven power of Convolutional Neural Network (CNN) in vision tasks, it is also adopted as the regressor in this framework and has achieved the state-of-the-art performance on face alignment [18, 23, 34, 50].

Despite the recent success, the cascade of CNNs, when applied to LPFA, still suffers from the following drawbacks.

Lack of end-to-end training: It is a consensus that end-to-end training is desired for CNN [6, 14]. However, in the existing methods, the CNNs are typically trained independently at each cascade stage. Sometimes even multiple CNNs are applied independently at each stage. For example, locations of different landmark sets are estimated by various CNNs and combined via a separate fusing module [33]. Therefore, these CNNs can not be jointly optimized and might lead to a sub-optimal solution.

Hand-crafted feature extraction: Since the CNNs are trained independently, feature extraction is required to utilize the result of a previous CNN and provide input to the current CNN. Simple feature extraction methods are used, e.g., extracting patches based on 2D or 3D face shapes without considering other factors including pose and expression [33, 45]. Normally, the cascade of CNNs is a collection of shallow CNNs where each one has less than five layers. Hence, this framework can not extract *deep* features by building upon the extracted features of early-stage CNNs.

Slow training speed: Training a cascade of CNNs is

This work was performed mostly during the first author’s internship at Bosch Research North America, Palo Alto, CA.

usually time-consuming for two reasons. Firstly, the CNNs are trained sequentially, one after another. Secondly, feature extraction is required between two consecutive CNNs.

To address these issues, we introduce a novel layer, as shown in Figure 1. Our CNN architecture consists of several blocks, which are called visualization blocks. This architecture can be considered as a cascade of shallow CNNs. The new layer visualizes the alignment result of a previous visualization block and utilizes it in a later visualization block. It is designed based on several guidelines. Firstly, it is derived from the surface normals of the underlying 3D face model and encodes the relative pose between the face and camera, partially inspired by the success of using surface normals for 3D face recognition [25]. Secondly, the visualization layer is differentiable, which allows the gradient to be computed analytically and enables end-to-end training. Lastly, a mask is utilized to differentiate between pixels in the middle and contour areas of a face.

Benefiting from the design of the visualization layer, our method has the following advantages and contributions:

- ◊ The proposed method allows a block in the CNN to utilize the extracted features from previous blocks and extract deeper features. Therefore, extraction of hand-crafted features is no longer necessary.
- ◊ The visualization layer is differentiable, allowing for backpropagation of an error from a later block to an earlier one. To the best of our knowledge, this is the first method for large-pose face alignment, that utilizes only one single CNN and allows end-to-end training.
- ◊ The proposed method converges faster during the training phase compared to the cascade of CNNs. Therefore, the training time is dramatically reduced.

2. Prior Work

In this section, we review related work in three topics: cascade of regressors for face alignment, convolutional recurrent neural network and visualization in deep learning.

Cascade of Regressors for Face Alignment Cascade of Regressors is a classic approach in not only conventional face alignment [44, 51], but also the large-pose face alignment [13, 17, 39, 49]. To handle large poses, many approaches go beyond 2D landmarks and also estimate 3D landmarks and 3D face shapes [18, 50]. Zhu et al. [49] use a set of local regressors to estimate the 2D shape update, and fuse their results with another regressor. The occlusion-invariant approach of RCPR [5] is applicable to large poses since self-occlusion is one type of occlusions. An iterative probabilistic method is utilized in [11, 15] for registering 3D shape to the pre-computed 2D landmarks. Tulyakov et al. [35] also use a cascade of regressors to estimate 3D landmark updates directly from a single image. Some even use

two regressors at each cascade stage. Wu et al. [39] use one regressor to estimate the 2D shape update and the other to estimate the visibility of each landmark. Similarly, Liu et al. [23] employ one regressor for 2D shape update and the other uses the 2D shape to estimate the 3D face shape.

Among methods with cascade of regressors, CNN is a popular choice of regressor due to its strong learning ability. These methods typically extract hand-crafted features between consecutive regressors. TCDCN [46] uses one CNN to estimate five landmarks, with yaw angles within $\pm 60^\circ$. A cascade of stacked autoencoder (SAE) progressively estimates 2D landmark updates from extracted patches [45]. Similarly, cascades of CNNs with global or local patches are combined at each stage, and their results are fused via averaging [33, 48]. The methods in [18, 19, 50] combine cascade of CNNs with 3D feature extraction to estimate the dense 3D face shape. All aforementioned methods lack the ability to end-to-end train the network, which is our novel contribution to large-pose face alignment.

Convolutional Recurrent Neural Network (CRNN) Methods based on CRNNs [34, 37, 40] are the first attempts to combine cascade of regressors with joint optimization, for aligning mostly frontal faces. Their convolutional part extracts features from the whole image [40] or from the patches at the landmark locations [34]. The recurrent part facilitates the joint optimization by sharing information among all regressors. The main differences between the proposed method and CRNNs are: 1) existing CRNN methods are designed for near-frontal face alignment, while ours is for LPFA; 2) the CRNN methods share the same CNN at all stages, while our CNN of each block is different which might be more suitable for estimating the coarse-to-fine mappings during the course of alignment; 3) due to our new differentiable visualization layer, our method has one additional flow of the gradient back-propagation (note the two blue arrows between consecutive blocks in Figure 2).

Visualization in Deep Learning Visualization techniques have been used in deep learning to assist in making a relative comparison among the input data and focusing on the region of interest. One category of these methods exploits the deconvolutional and upsampling layers to either expand response maps [22, 28] or represent estimated parameters [43]. Alternatively, various types of feature maps, e.g., heatmaps and Z-Buffering, can represent the current estimation of landmarks and parameters. In [4, 26, 38], 2D landmark heatmaps represent the landmarks' locations. [4] proposes a two step large pose alignment based on heatmaps to make more precise estimations. The heatmaps suffer from three drawbacks: 1) lack of the capability to represent objects in details; 2) requirement of one heatmap per landmark due to its weak representation power. 3) they cannot estimate visibility of landmarks. The Z-Buffer rendered using the estimated 3D face is also used to convey the re-

The source code of the proposed method with the trained model are released [here](#).

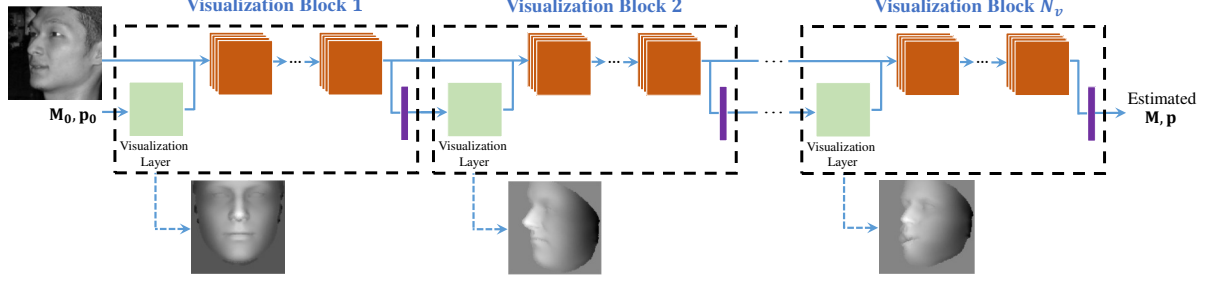


Figure 2. The proposed CNN architecture. We use green, orange, and purple to represent the visualization layer, convolutional layer, and fully connected layer, respectively. Please refer to Figure 3 for the details of the visualization block.

sults of a previous CNN to the next one [50]. However, the Z-Buffer representation is not differentiable, preventing end-to-end training. In contrast, our visualization layer is differentiable and encodes the face geometry details via surface normals. It guides the CNN to focus on the face area that incorporates both the pose and expression information.

3. Proposed Method

Given a single face image with an arbitrary pose, our goal is to estimate the 2D landmarks with their visibility labels by fitting a 3D face model. Towards this end, we propose a CNN architecture with end-to-end training for model fitting, as shown in Figure 2. In this section, we will first describe the underlying 3D face model used in this work, followed by our CNN architecture and the visualization layer.

3.1. 3D and 2D Face Shapes

We use the 3D Morphable Model (3DMM) to represent the 3D shape of a face $\mathbf{S}_p$ as a linear combination of mean shape $\mathbf{S}_0$, identity bases $\mathbf{S}^I$ and expression bases $\mathbf{S}^E$:

$$\mathbf{S}_p = \mathbf{S}_0 + \sum_k^{N_I} p_k^I \mathbf{S}_k^I + \sum_k^{N_E} p_k^E \mathbf{S}_k^E. \quad (1)$$

We use vector $\mathbf{p} = [\mathbf{p}^I, \mathbf{p}^E]$ to indicate the 3D shape parameters, where $\mathbf{p}^I = [p_0^I, \dots, p_{N_I}^I]$ are the identity parameters and $\mathbf{p}^E = [p_0^E, \dots, p_{N_E}^E]$ are the expression parameters. We use the Basel 3D face model [27], which has 199 bases, as our identity bases and the face warehouse model [7] with 29 bases as our expression bases. Each 3D face shape consists of a set of Q 3D vertices:

$$\mathbf{S}_p = \begin{pmatrix} x_1^p & x_2^p & \dots & x_Q^p \\ y_1^p & y_2^p & \dots & y_Q^p \\ z_1^p & z_2^p & \dots & z_Q^p \end{pmatrix}. \quad (2)$$

The 2D face shapes are the projection of 3D shapes. In this work, we use the weak perspective projection model with 6 degrees of freedoms, i.e., one for scale, three for rotation angles and two for translations, which projects the 3D face shape $\mathbf{S}_p$ onto 2D images to obtain the 2D shape $\mathbf{U}$:

$$\mathbf{U} = f(\mathbf{P}) = \mathbf{M} \begin{pmatrix} \mathbf{S}_p(:, \mathbf{b}) \\ \mathbf{1} \end{pmatrix}, \quad (3)$$

where

$$\mathbf{M} = \begin{bmatrix} m_1 & m_2 & m_3 & m_4 \\ m_5 & m_6 & m_7 & m_8 \end{bmatrix}, \quad (4)$$

and

$$\mathbf{U} = \begin{pmatrix} x_1^t & x_2^t & \dots & x_N^t \\ y_1^t & y_2^t & \dots & y_N^t \end{pmatrix}. \quad (5)$$

Here $\mathbf{U}$ is a set of N 2D landmarks, $\mathbf{M}$ is the camera projection matrix. With misuse of notation, we define the target parameters $\mathbf{P} = \{\mathbf{M}, \mathbf{p}\}$. The N -dim vector $\mathbf{b}$ includes 3D vertex indexes which are semantically corresponding to 2D landmarks. We denote $\mathbf{m}_1 = [m_1 \ m_2 \ m_3]$ and $\mathbf{m}_2 = [m_5 \ m_6 \ m_7]$ as the first two rows of the scaled rotation component, while m_4 and m_8 are the translations.

Equation 3 establishes the relationship, or equivalency, between 2D landmarks $\mathbf{U}$ and $\mathbf{P}$, i.e., 3D shape parameters $\mathbf{p}$ and the camera projection matrix $\mathbf{M}$. Given that almost all the training images for face alignment have only 2D labels, i.e., $\mathbf{U}$, we preform a data augmentation step similar to [18] to compute their corresponding $\mathbf{P}$. Given an input image, our goal is to estimate the parameter $\mathbf{P}$, based on which the 2D landmarks and their visibilities can be naturally derived.

3.2. Proposed CNN Architecture

Our CNN architecture resembles the cascade of CNNs, while each ‘‘shallow CNN’’ is defined as a visualization block. Inside each block, a visualization layer based on the latest parameter estimation serves as a bridge between consecutive blocks. This design enables us to address the drawbacks of typical cascade of CNNs in Section 1. We now describe the visualization block and CNN architecture, and dive into the details of the visualization layer in Section 3.3.

Visualization Block Figure 3 shows the structure of our visualization block. The visualization layer generates a feature map based on the latest parameter $\mathbf{P}$ (details in Section 3.3). Each convolutional layer is followed by a batch normalization (BN) layer and a ReLU layer. It extracts deeper features based on the features provided by the previous visualization block and the visualization layer output. Between the two fully connected layers, the first one is followed by a ReLU layer and a dropout layer, while the

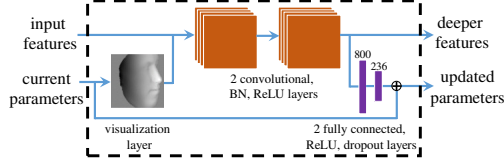


Figure 3. A visualization block consists of a visualization layer, two convolutional layers and two fully connected layers.

second one simultaneously estimates the update of $\mathbf{M}$ and $\mathbf{p}$, denoted $\Delta\mathbf{P}$. The outputs of the visualization block are deeper features and the new estimation of the parameters ($\Delta\mathbf{P} + \mathbf{P}$). As shown in Figure 3, the top part of the visualization block focuses on learning deeper features, while the bottom part utilizes those features to estimate the parameters in a ResNet-like structure [12]. During the backward pass of the training phase, the visualization block backpropagates the loss through both of its inputs to adjust the convolutional and the fully connected layers in previous blocks. This allows the block to extract better features for the next block and improve the overall parameter estimation.

CNN Architecture The proposed architecture consists of several connected visualization blocks as shown in Figure 2. The inputs include an image and an initial estimation of the parameter $\mathbf{P}^0$. The outputs is the final estimation of the parameter. Due to the joint optimization of all visualization blocks through backpropagation, the proposed architecture is able to converge with substantially fewer epochs during training, compared to the typical cascade of CNNs.

Loss Functions Two types of loss functions are employed in our CNN architecture. The first one is an Euclidean loss between the estimation and the target of the parameter update, with each parameter weighted separately:

$$\mathbf{E}_P^i = (\Delta\mathbf{P}^i - \Delta\bar{\mathbf{P}}^i)^T \mathbf{W} (\Delta\mathbf{P}^i - \Delta\bar{\mathbf{P}}^i), \quad (6)$$

where $\mathbf{E}_P^i$ is the loss, $\Delta\mathbf{P}^i$ is the estimation and $\Delta\bar{\mathbf{P}}^i$ is the target (or ground truth) at the i -th visualization block. The diagonal matrix $\mathbf{W}$ contains the weights. For each element of the shape parameter $\mathbf{p}$, its weight is the inverse of the standard deviation that was obtained from the data used in 3DMM training. To compensate the relative scale among the parameters of $\mathbf{M}$, we compute the ratio r between the average of scaled rotation parameters and average of translation parameters in the training data. We set the weights of the scaled rotation parameters of $\mathbf{M}$ to $\frac{1}{r}$ and the weights of the translation of $\mathbf{M}$ to 1. The second type of loss function is the Euclidean loss on the resultant 2D landmarks:

$$\mathbf{E}_S^i = \|f(\mathbf{P}^i + \Delta\mathbf{P}^i) - \bar{\mathbf{U}}\|^2, \quad (7)$$

where $\bar{\mathbf{U}}$ is the ground truth 2D landmarks, and $\mathbf{P}^i$ is the input parameter to the i -th block, i.e., the output of the $i-1$ -th block. $f(\cdot)$ computes 2D landmark locations using the currently updated parameters via Equation 3. For backpropagation of this loss function to the parameter $\Delta\mathbf{P}$, we use the chain rule to compute the gradient.

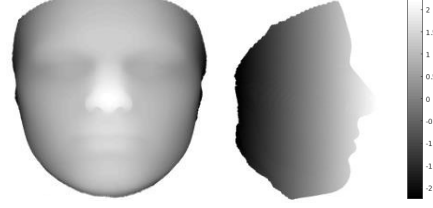


Figure 4. The frontal and side views of the mask $\mathbf{a}$ that has positive values in the middle and negative values in the contour area.

$$\frac{\partial \mathbf{E}_S^i}{\partial \Delta\mathbf{P}^i} = \frac{\partial \mathbf{E}_S^i}{\partial f} \frac{\partial f}{\partial \mathbf{P}^i}.$$

For the first three visualization blocks, the Euclidean loss on the parameter updates (Equation 6) is used, while the Euclidean loss on 2D landmarks (Equation 7) is applied to the last three blocks. The first three blocks estimate parameters to roughly align 3D shape to the face image and the last three blocks leverage the good initialization to estimate the parameters and the 2D landmark locations more precisely.

3.3. Visualization Layer

Several visualization techniques have been explored for facial analysis. In particular, Z-Buffering, which is widely used in prior works [2, 3], is a simple and fast 2D representation of the 3D shape. However, this representation is not differentiable. In contrast, our visualization is based on surface normals of the 3D face, which describes surface's orientation in a local neighbourhoods. It has been successfully utilized for different facial analysis tasks, e.g., 3D face reconstruction [30] and 3D face recognition [25].

In this work, we use the z coordinate of surface normals of each vertex, transformed with the pose. It is an indicator of “frontability” of a vertex, i.e., the amount that the surface normal is pointing towards the camera. This quantity is used to assign an intensity value at its projected 2D location to construct the visualization image. The frontability measure $\mathbf{g}$, a Q dimensional vector, can be computed as,

$$\mathbf{g} = \max \left(\mathbf{0}, \frac{(\mathbf{m}_1 \times \mathbf{m}_2)}{\|\mathbf{m}_1\| \|\mathbf{m}_2\|} \mathbf{N}_0 \right), \quad (8)$$

where $\times$ is the cross product, and $\|\cdot\|$ denotes the L_2 norm. The $3 \times Q$ matrix $\mathbf{N}_0$ is the surface normal vectors of a 3D face shape. To avoid the high computational cost of calculating the surface normals after each shape update, we approximate $\mathbf{N}_0$ with the surface normals of the mean 3D face. Note that both the face shape and pose are still continuously updated across visualization blocks, and are used to determine the projected 2D location. Hence, this approximation would only slightly affect the intensity values. To transform the surface normal based on the pose, we apply the estimation of the scaled rotation matrix ($\mathbf{m}_1$ and $\mathbf{m}_2$) to the surface normals computed from the mean face. The value is then truncated with the lower bound of 0 (Equation 8).

The pixel intensity of a visualized image $\mathbf{V}(u, v)$ is computed as the weighted average of the frontability measures within a local neighbourhood:

$$\mathbf{V}(u, v) = \frac{\sum_{q \in \mathbb{D}(u, v)} \mathbf{g}(q) \mathbf{a}(q) w(u, v, x_q^t, y_q^t)}{\sum_{q \in \mathbb{D}(u, v)} w(u, v, x_q^t, y_q^t)}, \quad (9)$$

where $\mathbb{D}(u, v)$ is the set of indexes of vertexes whose 2D projected locations are within the local neighborhood of the pixel (u, v) . (x_q^t, y_q^t) is the 2D projected location of q -th 3D vertex. The weight w is the distance metric between the pixel (u, v) and the projected location (x_q^t, y_q^t) ,

$$w(u, v, x_q^t, y_q^t) = \exp\left(-\frac{(u - x_q^t)^2 + (v - y_q^t)^2}{2\sigma^2}\right). \quad (10)$$

The Q -dim vector $\mathbf{a}$ is a mask with positive values for vertexes in the middle area of the face and negative values for vertexes around the contour area of the face:

$$\mathbf{a}(q) = \exp\left(-\frac{(x^n - x_q^p)^2 + (y^n - y_q^p)^2 + (z^n - z_q^p)^2}{2\sigma_n^2}\right), \quad (11)$$

where (x^n, y^n, z^n) is the vertex coordinate of the nose tip. $\mathbf{a}$ is pre-computed and normalized for zero-mean and unit standard deviation. The mask is utilized to discriminate between the middle and contour areas of the face. A visualization of the mask is provided in Figure 4.

Since the human face is a 3D object, visualizing it at an arbitrary view angle requires the estimation of the visibility of each 3D vertex. To avoid the computationally expensive visibility test via rendering, we adopt two strategies for approximation. Firstly, we prune the vertexes whose frontability measures $\mathbf{g}$ equal 0, i.e., the vertexes pointing against the camera. Secondly, if multiple vertexes projects to a same image pixel, we keep only the one with the smallest depth values. An illustration is provided in Figure 5.

Backpropagation To allow backpropagation of the loss functions through the visualization layers, we compute the derivative of $\mathbf{V}$ with respect to the elements of the parameters $\mathbf{M}$ and $\mathbf{p}$. Firstly, we compute the partial derivatives, $\frac{\partial \mathbf{g}}{\partial m_k}$, $\frac{\partial w(u, v, x_i^t, y_i^t)}{\partial m_k}$ and $\frac{\partial w(u, v, x_i^t, y_i^t)}{\partial p_j}$. Then the derivatives of $\frac{\partial \mathbf{V}}{\partial m_k}$ and $\frac{\partial \mathbf{V}}{\partial p_j}$ can be computed based on Equation. 9.

4. Experimental Results

We evaluate our method on two challenging LPFA datasets, AFLW and AFW, both qualitatively and quantitatively, as well as the near-frontal face dataset of 300W. Furthermore, we conduct experiments on different CNN architectures to validate our visualization layer design.

Implementation details Our implementation is built upon the Caffe toolbox [16]. In all of the experiments, we use six visualization blocks (N_v) with two convolutional layers (N_c) and fully connected layers in each block (Figure 3). Details of the network structure are provided in Table 1.

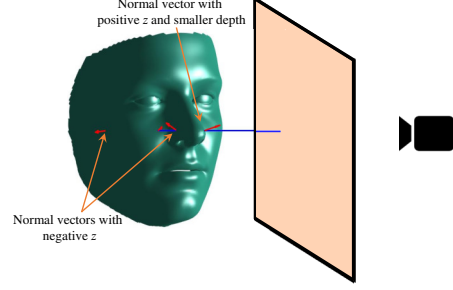


Figure 5. An example with four vertexes projected to a same pixel. Two of them have negative values in z component of their normals (red arrows). Between the other two with positive values, the one with the smaller depth (closer to the image plane) is selected.

Instead of using the sequentially pretrain strategy [42], we perform the joint end-to-end training from scratch. To better estimate the parameter update in each block and to increase the effectiveness when using visualization blocks, we set the weight of the loss function in the first visualization block to 1 and linearly increase the weights by one for each later block. This strategy helps the CNN to pay more attention to the landmark loss used in later blocks. Back-propagation of loss functions in the last blocks would have more impact in the first block, and the last block can adopt itself more quickly to the changes in the first block.

In the training phase, we set the weight decay to 0.005, the momentum to 0.99, the initial learning rate to $1e-6$. Besides, we decrease the learning rate to $5e-6$ and $1e-7$ after 20 and 29 epochs. In total, the training phase is continued for 33 epochs for all experiments.

4.1. Quantitative Evaluations on AFLW and AFW

The AFLW dataset [21] is very challenging with large-pose face images ($\pm 90^\circ$ yaw). We use the subset with 3,901 training images and 1,299 testing images released by [18]. All face images in this subset are labeled with 34 landmarks and a bounding box. The AFW dataset [51] contains 205 images with 468 faces. Each image is labeled with at most 6 landmarks with visibility labels, as well as a bounding box. AFW is used only for testing in our experiments. The bounding boxes in both datasets are used as the initialization for our algorithm, as well as the baselines. We crop the region inside the bounding box and normalize it to 114×114 . Due to the memory constraint of GPUs, we have a pooling layer in the first visualization block after the first convolutional layer to decrease the size of feature maps to half. The input to the subsequent visualization blocks is

Table 1. The number and size of convolutional filters in each visualization block. For all blocks, the two fully connected layers have the same length of 800 and 236.

Block #	1	2	3	4	5, 6
Conv.	12 (5×5)	20 (3×3)	28 (3×3)	36 (3×3)	40 (3×3)
layers	16 (5×5)	24 (3×3)	32 (3×3)	40 (3×3)	40 (3×3)

Table 2. NME (%) of four methods on the AFLW dataset.

Proposed method	LPFA [18]	PIFA	RCPR
4.45	4.72	8.04	6.26

Table 3. NME (%) of the proposed method at each visualization block on AFLW dataset. The initial NME is 25.8%.

Block #	1	2	3	4	5	6
NME	9.26	6.77	5.51	4.98	4.60	4.45

of 57×57 . To augment the training data, we generate 20 different variations for each training image by adding noise to the location, width and height of the bounding boxes.

For quantitative evaluations, we use two conventional metrics. The first one is Mean Average Pixel Error (MAPE) [44], which is the average of the pixel errors for the visible landmarks. The other one is Normalized Mean Error (NME), i.e., the average of the normalized estimation error of visible landmarks. The normalization factor is the square root of the face bounding box size [17], instead of the eye-to-eye distance in the frontal-view face alignment.

We compare our method with several state-of-the-art LPFA approaches. On AFLW, we compare with LPFA [18], PIFA [17] and RCPR [5] with the NME metric. Table 2 shows that our proposed method achieves a higher accuracy than the alternatives. The heatmap-based method, named CALE [4], reported an NME of 2.96%, but suffers from several disadvantages as discussed in Section 2. To demonstrate the capabilities of each visualization block, the NME computed using the estimated $\mathbf{P}$ after each block is shown in Table 3. If a higher alignment speed is desirable, it is possible to skip the last two visualization blocks with a reasonable NME. On the AFW dataset, comparisons are conducted with LPFA [18], PIFA [17], CDM [44] and TSPM [51] with the MAPE metric. The results in Table 4 again show the superiority of the proposed method.

Some examples of alignment results of the proposed method on AFLW and AFW datasets are shown in Figure 9. Three examples of visualization layer output at each visualization block are shown in Figure 10.

4.2. Evaluation on 300W dataset

While our main goal is LPFA, we also evaluate on the most widely used near frontal 300W dataset [31]. 300W contains 3, 148 training and 689 testing images, which are divided into common and challenging sets with 554 and 135 images, respectively. Table 5 shows the NME (normalized by the interocular distance) of the evaluated methods. The most relevant method is 3DDFA [50], which also estimates $\mathbf{M}$ and $\mathbf{p}$. Our method outperforms it on both the common and challenging sets. Methods that do not employ shape constraints, e.g., via 3DMM, generally have higher freedom

Table 4. MAPE of five methods on the AFW dataset.

Proposed method	LPFA [18]	PIFA	CDM	TSPM
6.27	7.43	8.61	9.13	11.09

Table 5. The NME of different methods on 300W dataset.

Method	Common	Challenging	Full
ESR [8]	5.28	17.00	7.58
RCPR [5]	6.18	17.26	8.35
SDM [41]	5.57	15.40	7.50
LBF [29]	4.95	11.98	6.32
CFSS [50]	4.73	9.98	5.76
RCFA [37]	4.03	9.85	5.32
RAR [40]	4.12	8.35	4.94
3DDFA [50]	6.15	10.59	7.01
3DDFA+SDM	5.53	9.56	6.31
Proposed method	5.43	9.88	6.30

and could achieve slightly better accuracy on frontal face cases. Nonetheless, they are typically less robust in more challenging cases. Another comparison is with MDM [34] via the failure rate using a threshold of 0.08. The failure rates are 16.83% (ours) versus 6.80% (MDM) with 68 landmarks, and 8.99% (ours) versus 4.20% (MDM) with 51 landmarks.

4.3. Analysis of the Visualization Layer

We perform four sets of experiments to study the properties of the visualization layer and network architectures.

Influence of visualization layers To analyze the influence of the visualization layer in the testing phase, we add 5% noise to the fully connected layer parameters of each visualization block, and compute the alignment error on the AFLW test set. The NMEs are [4.46, 4.53, 4.60, 4.66, 4.80, 5.16] when each block is modified separately. This analysis shows that the visualized images have more influence on the later blocks, since imprecise parameters of early blocks could be compensated by later blocks. In another experiment, we train the network without any visualization layer. The final NME on AFLW is 7.18% which shows the importance of visualization layers in guiding the network training.

Advantage of deeper features We train three CNN architectures shown in Figure 6 on AFLW. The inputs of the visualization block in the first architecture are the input images $\mathbf{I}$, feature maps $\mathbf{F}$ and the visualization image $\mathbf{V}$. The inputs of the second and the third architectures are $\{\mathbf{F}, \mathbf{V}\}$ and $\{\mathbf{I}, \mathbf{V}\}$, respectively. The NME of each architecture is

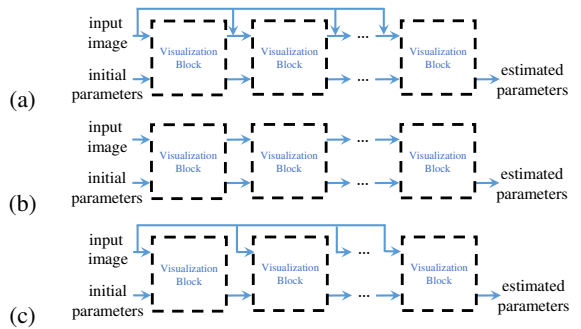


Figure 6. Architectures of three CNNs with different inputs.

Table 6. The NME (%) of three architectures with different inputs (I: Input image, V: Visualization, F: Feature maps).

Architecture a (I, F, V)	Architecture b (F, V)	Architecture c (I, V)
4.45	4.48	5.06

shown in Table 6. While the first one performs the best, the substantial lower performance of the third one demonstrates the importance of deeper features learned across blocks.

At the first convolutional layer of each visualization block, we compute the average of the filter weights, across both the kernel size and the number of maps. The averages for these three types of input features are shown in Figure 7. As observed, the weights decrease across blocks, leading to a more precise estimation of small-scale parameter updates. Considering the number of filters in Table 1, the total impact of feature maps are higher than the other two inputs in all blocks. This again shows the importance of deeper features in guiding the network to estimate parameters. Furthermore, the average of the visualization filter is higher than that of the input image filter, demonstrating stronger influence of the proposed visualization during training.

Advantage of using masks To show the advantage of using the mask in the visualization layer, we conduct an experiment with different masks. Specifically, we define another mask for comparison as shown in Figure 8. It has five positive areas, i.e., the eyes, nose tip and two lip corners. The values are normalized to zero-mean and unit standard deviation. Compared to the original mask in Fig. 4, this mask is more complicated and conveys more information about the informative facial areas to the network. Moreover, to show the necessity of using the mask, we also test using visualization layers without any mask. The NMEs of the trained networks with different masks are shown in Table 7. Comparison between the first and the third columns shows the benefit of using the mask, by differentiating the middle and contour areas of the face. By comparing the first and second columns, we can see that utilizing more complicated mask does not further improve the result, indicating the original mask provides sufficient information for its purpose.

Different numbers of blocks and layers Given the total number of 12 convolutional layers in our network, we can partition them into visualization blocks of various sizes. To compare their performance, we train two additional CNNs. The first one consists of 4 visualization blocks, with 3 con-

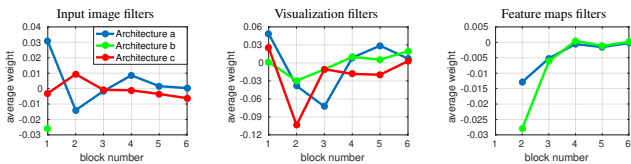


Figure 7. The average of filter weights for input image, visualization and feature maps in three architectures of Figure 6. The y -axis and x -axis show the average and the block index, respectively.

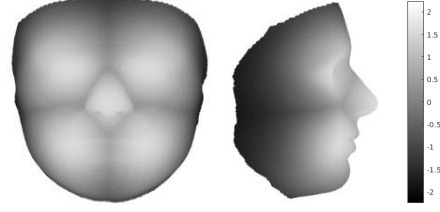


Figure 8. Mask 2, a different designed mask with five positive areas on the eyes, top of the nose and sides of the lip.

Table 7. NME (%) when different masks are used.

Mask 1	Mask 2	No Mask
4.45	4.49	5.31

Table 8. NME (%) when using different numbers of visualization blocks (N_v) and convolutional layers (N_c).

$N_v = 6, N_c = 2$	$N_v = 4, N_c = 3$	$N_v = 3, N_c = 4$
4.45	4.61	4.83

volutional layers in each. The other comes with 3 block and 4 convolutional layers per block. Hence, all three architectures have 12 convolutional layers in total. The NME of these architectures are shown in Table 8. Similar to [5], it shows that the number of regressors is important for face alignment and we can potentially achieve a higher accuracy by increasing the number of visualization blocks.

4.4. Time complexity

Compared to the cascade of CNNs, one of the main advantages of end-to-end training a single CNN is the reduced training time. The proposed method needs 33 epochs which takes around 2.5 days. With the same training and testing data sets, [18] requires 70 epochs for each CNN. With a total of six CNNs, it needs around 7 days. Similarly, the method in [50] needs around 12 days to train three CNNs, each one with 20 epochs, despite using different training data. Compared to [18], our method reduces the training time by more than half. The testing speed of proposed method is 4.3 FPS on a Titan X GPU. It is much faster than the 0.6 FPS speed of [18] and is similar to the 4 FPS speed of [40].

5. Conclusions

We propose a large-pose face alignment method with end-to-end training in a single CNN. The key is a differentiable visualization layer, which is integrated to the network and enables joint optimization by backpropagating the error from a later visualization blocks to early ones. It allows the visualization block to utilize the extracted features from previous blocks and extract deeper features, without extracting hand-crafted features. In addition, the proposed method converges faster during the training phase compared to the cascade of CNNs. Through extensive experiments, we demonstrate the superior results of the proposed method over the state-of-the-art approaches.



Figure 9. Results of alignment on AFLW and AFW datasets, green landmarks show the estimated locations of visible landmarks and red landmarks show estimated locations of invisible landmarks. First row: provided bounding box by AFLW with initial locations of landmarks, Second: estimated 3D dense shapes, Third: estimated landmarks, Fourth to sixth: estimated landmarks for AFLW, Seventh: estimated landmarks for AFW.

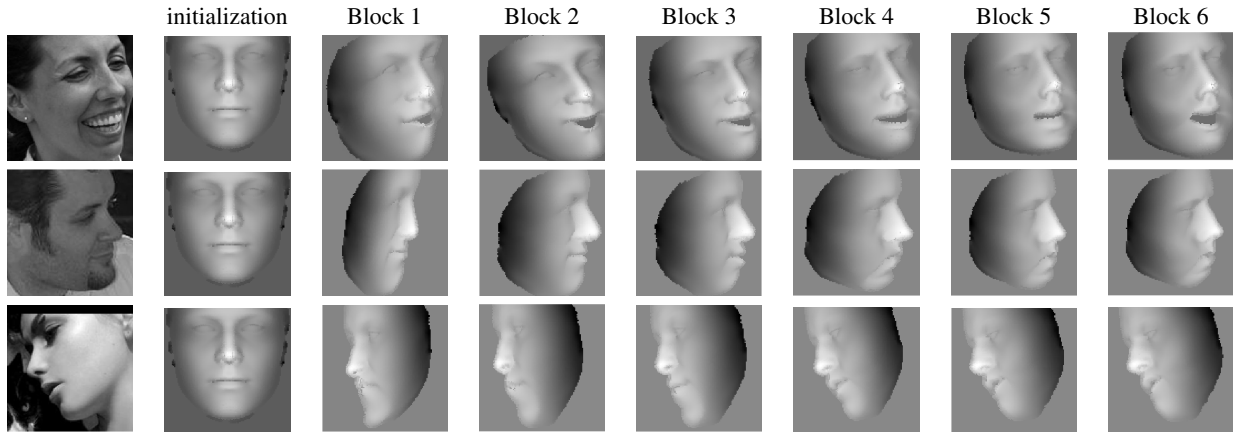


Figure 10. Three examples of outputs of visualization layer at each visualization block. The first row shows that the proposed method recovers the expression of the face gracefully, the third row shows the visualizations of a face with a more challenging pose.

References

- [1] V. Bettadapura. Face expression recognition and analysis: the state of the art. *arXiv preprint arXiv:1203.6722*, 2012. 1
- [2] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. In *ACM SIGGRAPH*, pages 187–194, 1999. 4
- [3] V. Blanz and T. Vetter. Face recognition based on fitting a 3D morphable model. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(9):1063–1074, 2003. 4
- [4] A. Bulat and G. Tzimiropoulos. Convolutional aggregation of local evidence for large pose face alignment. In *BMVC*, 2016. 2, 6
- [5] X. P. Burgos-Artizzu, P. Perona, and P. Dollár. Robust face landmark estimation under occlusion. In *CVPR*, pages 1513–1520, 2013. 2, 6, 7
- [6] H. Caesar, J. Uijlings, and V. Ferrari. Region-based semantic segmentation with end-to-end training. In *ECCV*, pages 381–397, 2016. 1
- [7] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. Face-warehouse: A 3D facial expression database for visual computing. *IEEE Trans. Vis. Comput. Graphics*, 20(3):413–425, 2014. 3
- [8] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. *Int. J. Comput. Vision*, 107(2):177–190, 2014. 1, 6
- [9] T. F. Cootes, G. J. Edwards, C. J. Taylor, et al. Active appearance models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(6):681–685, 2001. 1
- [10] D. Cristinacce and T. F. Cootes. Boosted regression active shape models. In *BMVC*, volume 1, page 7, 2007. 1
- [11] L. Gu and T. Kanade. 3D alignment of face in a single image. In *CVPR*, volume 1, pages 1305–1312, 2006. 2
- [12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4
- [13] G.-S. Hsu, K.-H. Chang, and S.-C. Huang. Regressive tree structured model for facial landmark localization. In *CVPR*, pages 3855–3861, 2015. 2
- [14] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *NIPS*, pages 2017–2025, 2015. 1
- [15] L. A. Jeni, J. F. Cohn, and T. Kanade. Dense 3D face alignment from 2D videos in real-time. In *FG*, volume 1, pages 1–8, 2015. 2
- [16] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM MM*, pages 675–678, 2014. 5
- [17] A. Jourabloo and X. Liu. Pose-invariant 3D face alignment. In *ICCV*, pages 3694–3702, 2015. 2, 6
- [18] A. Jourabloo and X. Liu. Large-pose face alignment via CNN-based dense 3D model fitting. In *CVPR*, 2016. 1, 2, 3, 5, 6, 7
- [19] A. Jourabloo and X. Liu. Pose-invariant face alignment via CNN-based dense 3D model fitting. *Int. J. Comput. Vision*, pages 1–17, 2017. 1, 2
- [20] I. Kemelmacher-Shlizerman and S. M. Seitz. Face reconstruction in the wild. In *ICCV*, pages 1746–1753, 2011. 1
- [21] 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 *ICCVW*, pages 2144–2151, 2011. 5
- [22] Y. Li, B. Sun, T. Wu, Y. Wang, and W. Gao. Face detection with end-to-end integration of a convnet and a 3D model. In *ECCV*, 2016. 2
- [23] F. Liu, D. Zeng, Q. Zhao, and X. Liu. Joint face alignment and 3D face reconstruction. In *ECCV*, pages 545–560, 2016. 1, 2
- [24] J. McDonagh and G. Tzimiropoulos. Joint face detection and alignment with a deformable hough transform model. In *ECCV*, pages 569–580, 2016. 1
- [25] H. Mohammadzade and D. Hatzinakos. Iterative closest normal point for 3D face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(2):381–397, 2013. 2, 4
- [26] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, pages 483–499, 2016. 2
- [27] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3D face model for pose and illumination invariant face recognition. In *AVSS*, pages 296–301, 2009. 3
- [28] X. Peng, R. S. Feris, X. Wang, and D. N. Metaxas. A recurrent encoder-decoder network for sequential face alignment. In *ECCV*, pages 38–56. Springer, 2016. 2
- [29] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3000 fps via regressing local binary features. In *CVPR*, pages 1685–1692, 2014. 6
- [30] J. Roth, Y. Tong, and X. Liu. Adaptive 3D face reconstruction from unconstrained photo collections. In *CVPR*, 2016. 1, 4
- [31] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *ICCVW*, pages 397–403, 2013. 6
- [32] J. M. Saragih, S. Lucey, and J. F. Cohn. Face alignment through subspace constrained mean-shifts. In *ICCV*, pages 1034–1041, 2009. 1
- [33] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *CVPR*, pages 3476–3483, 2013. 1, 2
- [34] G. Trigeorgis, P. Snape, M. A. Nicolaou, E. Antonakos, and S. Zafeiriou. Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In *CVPR*, 2016. 1, 2, 6
- [35] S. Tulyakov and N. Sebe. Regressing a 3D face shape from a single image. In *ICCV*, pages 3748–3755, 2015. 2
- [36] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, H. Mobahi, and Y. Ma. Toward a practical face recognition system: Robust alignment and illumination by sparse representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(2):372–386, 2012. 1
- [37] W. Wang, S. Tulyakov, and N. Sebe. Recurrent convolutional face alignment. In *ACCV*, 2016. 2, 6
- [38] J. Wu, T. Xue, J. J. Lim, Y. Tian, J. B. Tenenbaum, A. Torralba, and W. T. Freeman. Single image 3D interpreter network. In *ECCV*, 2016. 2
- [39] Y. Wu and Q. Ji. Robust facial landmark detection under significant head poses and occlusion. In *CVPR*, pages 3658–3666, 2015. 2

- [40] S. Xiao, J. Feng, J. Xing, H. Lai, S. Yan, and A. Kasim. Robust facial landmark detection via recurrent attentive-refinement networks. In *ECCV*, pages 57–72, 2016. 2, 6, 7
- [41] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *CVPR*, pages 532–539, 2013. 1, 6
- [42] Z. Yan, H. Zhang, R. Piramuthu, V. Jagadeesh, D. DeCoste, W. Di, and Y. Yu. HD-CNN: hierarchical deep convolutional neural networks for large scale visual recognition. In *ICCV*, pages 2740–2748, 2015. 5
- [43] J. Yang, S. E. Reed, M.-H. Yang, and H. Lee. Weakly-supervised disentangling with recurrent transformations for 3D view synthesis. In *NIPS*, pages 1099–1107, 2015. 2
- [44] 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 *ICCV*, pages 1944–1951, 2013. 2, 6
- [45] J. Zhang, S. Shan, M. Kan, and X. Chen. Coarse-to-Fine Auto-encoder Networks (cfan) for real-time face alignment. In *ECCV*, pages 1–16, 2014. 1, 2
- [46] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In *ECCV*, pages 94–108, 2014. 2
- [47] R. Zhao, Y. Wang, C. F. Benitez-Quiroz, Y. Liu, and A. M. Martinez. Fast and precise face alignment and 3D shape reconstruction from a single 2D image. In *ECCV*, pages 590–603, 2016. 1
- [48] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin. Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In *ICCVW*, pages 386–391, 2013. 2
- [49] S. Zhu, C. Li, C. C. Loy, and X. Tang. Unconstrained face alignment via cascaded compositional learning. In *CVPR*, 2016. 1, 2
- [50] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3D solution. In *CVPR*, 2016. 1, 2, 3, 6, 7
- [51] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, pages 2879–2886, 2012. 2, 5, 6