LEARNING A MULTI-CENTER CONVOLUTIONAL NETWORK FOR UNCONSTRAINED FACE ALIGNMENT

Zhiwen Shao, Hengliang Zhu, Yangyang Hao, Min Wang, and Lizhuang Ma**

Department of Computer Science and Engineering, Shanghai Jiao Tong University, China
{shaozhiwen, hengliang_zhu, haoyangyang2014, yinger650}@sjtu.edu.cn, ma-lz@cs.sjtu.edu.cn

ABSTRACT

In this paper, we propose a novel multi-center convolutional neural network for unconstrained face alignment. To utilize structural correlations among different facial landmarks, we determine several clusters based on their spatial position. We pre-train our network to learn generic feature representations. We further fine-tune the pre-trained model to emphasize on locating a certain cluster of landmarks respectively. Fine-tuning contributes to searching an optimal solution smoothly without deviating from the pre-trained model excessively. We obtain an excellent solution by combining multiple fine-tuned models. Extensive experiments demonstrate that our method possesses superior capability of handling extreme occlusions and complex variations of pose, expression, illumination. The code for our method is available at https://github.com/ZhiwenShao/MCNet.

***Index Terms*—** multi-center convolutional neural network, unconstrained face alignment, structural correlations

1. INTRODUCTION

Face alignment refers to detecting facial landmarks such as pupil centers, nose tip and mouth corners. It is the preprocessor stage of many face analysis tasks like face recognition [1] and face animation [2]. There is a pressing need for a robust and accurate face alignment method with the development of social networks and mobile terminals. Such requirement is still challenging in unconstrained scenarios, owing to severe occlusions and large face variations. Our goal is to develop an efficient face alignment method to handle unconstrained faces.

Due to the outstanding representation power, deep convolutional networks have achieved great success in various computer vision tasks. Face alignment can be regarded as a nonlinear regression problem, which transforms appearance to shape. We design an effective deep convolutional network to model the highly nonlinear function. Motivated by the excellent performance of VGGNet [3] in representing features, the structure of our network is based on stacked convolutional layers.

(a) Chin is occluded. (b) Right contour is invisible.

Fig. 1. Examples of unconstrained face images with partial occlusion and large pose.

We believe that each facial landmark is not isolated but highly correlated with adjacent landmarks. As shown in Figure 1(a), facial landmarks along the chin are all occluded, and landmarks around the mouth are partially occluded. Figure 1(b) shows that landmarks on the right side of face are almost invisible. Therefore, landmarks in the same local face region have similar properties including occlusion and visibility. We divide facial landmarks into several clusters based on their spatial location.

We propose a novel convolutional neural network, referred to as Multi-Center Network (MCNet), to reinforce the learning for each cluster which is treated as a separate center. Each center in our MCNet is fine-tuned to emphasize on the shape prediction of a specific face region respectively. By employing shared feature representations from a pre-trained basic model and multiple center-specific feature representations, we attain an excellent model. Another interesting aspect of the MCNet architecture is that the complexity of combined model is not increased compared to the basic model.

2. RELATED WORK

Our method achieves unconstrained face alignment based on a multi-center convolutional network. We review researches

* Corresponding author.

This work is supported by the National Natural Science Foundation of China (No. 61472245, 61502220 and U1304616), and the Science and Technology Commission of Shanghai Municipality Program (No. 16511101300).

978-1-5090-6067-2/17/$31.00

from three aspects related to our method: generic face alignment, unconstrained face alignment, and face alignment via deep learning.

Generic Face Alignment: Active Appearance Model [4] employs an appearance model and minimizes the texture residual to estimate the shape. Xiong et al. [5] predicted the location of facial landmarks by solving the nonlinear least squares problem, with SIFT [6] features and linear regressors applied. ESR [7] uses cascaded fern regression to predict the shape increment with pixel-difference features. Ren et al. [8] uses a locality principle to obtain a set of local binary features jointly learning a linear regression for locating landmarks. Most of these methods give an initial shape and refine the shape in an iterative manner. The final solutions are apt to get trapped in a local optima with a poor initialisation. Unlike these methods, our network takes raw face patches as input.

Unconstrained Face Alignment: Large pose variations and severe occlusions are main challenges in unconstrained environments. Many methods utilize 3D shape models to solve large-pose face alignment. Yu et al. [9] uses a cascaded deformable shape model to locate landmarks of large-pose faces. Cao et al. [2] employs a Displaced Dynamic Expression regression to estimate the 3D face shape and 2D facial landmarks. Jourabloo et al. [10] proposed a cascaded coupled-regressor to infer parameters of 3D shapes. It can predict both location and visibility of facial landmarks. RCPR [11] detects occlusions explicitly and uses shape-indexed features to regress the shape increment. Wu et al. [12] designed a robust cascaded regressor to handle complex occlusions and large head poses. Different from these methods, our method is not based on 3D models and does not process occlusions specifically.

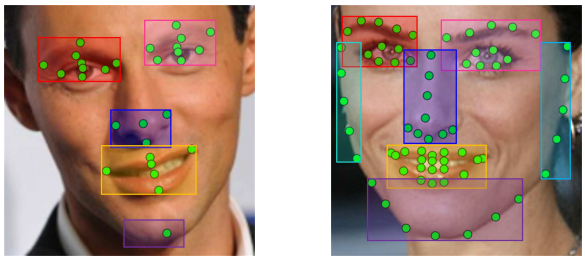
Face Alignment via Deep Learning: Cascaded CNN [13] estimates the position of five facial landmarks with cascaded convolutional networks. It uses average estimation in each level and refines the shape level by level. Zhou et al. [14] also uses multistage deep networks to detect facial landmarks from coarse to fine. CFT [15] learns the mapping from input face patch to estimated shape using a coarse-to-fine training strategy. It searches the solution smoothly by adjusting the relative weight between principal landmarks and elaborate landmarks. TCDCN [16] employs auxiliary facial attribute recognition to obtain correlative facial properties like expression and pose, which improves the performance of landmark detection. In contrast, our method uses only one network and is independent of additional facial attributes. Both CFT and TCDCN utilize fine-tuning methods to improve the effectiveness of learning process. Our method also use the fine-tuning strategy to obtain a better solution from the pre-trained model.

3. MULTI-CENTER NETWORK

In this section, we describe the structure of our MCNet and the learning algorithm. Our network reinforces the learning for landmarks of each local facial part.

3.1. Network Architecture

We propose an effective multi-center convolutional neural network to learn a mapping from appearance to shape. We analyse the facial structure and partition facial landmarks into seven clusters, as shown in Figure 2. The seven clusters are left eye, right eye, nose, mouth, left contour, right contour and chin.



(a) Partition of 29 landmarks. (b) Partition of 68 landmarks.

Fig. 2. Partition of facial landmarks.

Our network consists of shared layers and multiple center-specific shape prediction layers, as illustrated in Figure 3. We initialize shared layers and each center-specific layer with a pre-trained basic model which has only one shape prediction layer. There are m branches of center-specific layers at the end of our network. The value of m is 5 and 7 for 29 and 68 facial landmarks respectively. Each center-specific layer estimates x and y coordinates of all n facial landmarks, while focusing on the shape estimation of a specific face region. We obtain a new shape prediction layer by combining estimation units from corresponding center-specific layers. Shared layers and combined shape prediction layer compose the combined model whose complexity is as same as the basic model.

In our network, eight convolutional layers and one fully-connected layer are used for learning generic feature representations. We perform the batch normalization [17] and Rectified Linear Unit [18] activation after each convolution, to accelerate the convergence of our network. Each max-pooling layer follows a stack of two convolutional layers proposed by VGGNet [3]. We use inter-ocular distance normalized Euclidean loss [15] to measure the performance of estimation. It should be noted that the inter-ocular distance is the Euclidean distance between the two pupil centers.

In order to increase the diversity of training data, we employ a similar data augmentation method to [19] with four steps: rotation, translation, horizontal flip, and JPEG compression. This is beneficial for avoiding overfitting and improving the robustness of learned models. During the pre-training process, due to the large initial loss, we employ a small base learning rate to avoid divergence. According to the principle of Adaptive Learning Rate (ALR) [19] algorithm,

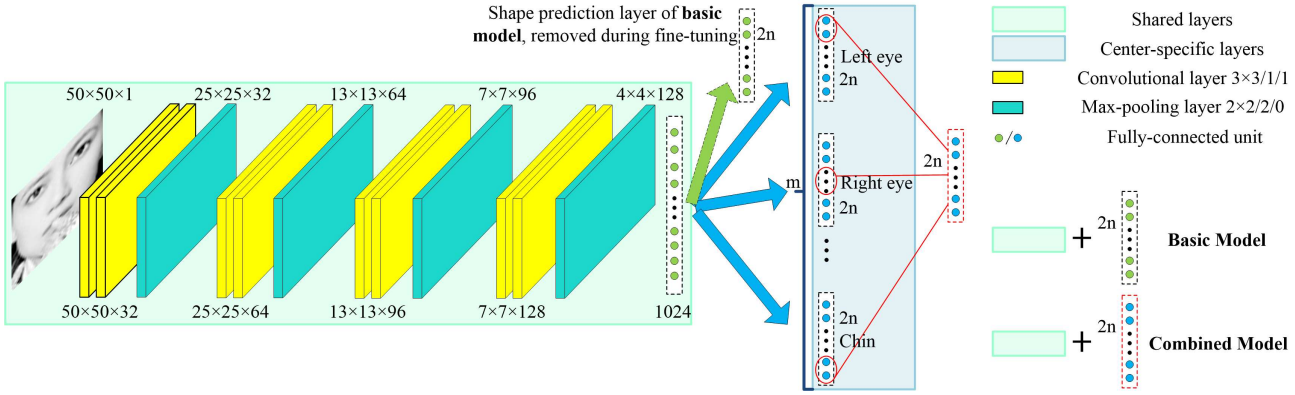


Fig. 3. The structure of our MCNet. It finally obtains a combined model fine-tuned from a pre-trained basic model. The equation attached to each layer signifies the height, width and channel respectively. Every stack of two convolutional layers possesses the same equation. The equation $k_1 \times k_2/k_3/k_4$ symbolizes the height, width, stride and padding of filters respectively. The same type of layers use identical filters.

we increase the learning rate when the loss is reduced significantly.

Compared to other typical convolutional networks like VGGNet [3], our network is substantially smaller and shallower. We believe that such a concise structure is efficient for estimating the location of facial landmarks. Firstly, face alignment aims to regress coordinates of fewer than 100 facial landmarks generally, which demands much smaller model complexity than visual recognition problems. Secondly, a very deep network may fail to work well for landmark detection owing to reduction of spatial information layer by layer. Finally, a simple network is not easy to overfit given a small amount of training data.

3.2. Learning Algorithm

Algorithm 1 is the overview of our learning algorithm. The basic model and combined model both have only one branch C. Θ is the set of weights and biases in our network, which is updated using Stochastic Gradient Descent algorithm at each iteration. Ω and Φ are used for training and model selection respectively. We represent shared layers and the i-th center-specific layer of our network with S and C^i respectively. $\hat{f}_{2j-1}$ and $\hat{f}_{2j}$ denote predicted x coordinate and y coordinate of the j-th facial landmark respectively, and f signifies ground truth coordinates. w_j is the weight of the j-th landmark, whose value is 1 during pre-training. d denotes the ground truth inter-ocular distance. Θ^S signifies the corresponding part of shared layers in Θ.

We first pre-trains a basic model, and further fine-tunes each center-specific layer to search a better solution from a good initial point respectively. After fine-tuning all the center-specific layers, we replace these layers with a single branch and combine their corresponding parameters. The final combined model improves the location performance of each facial landmark by exploiting the advantages of every center-specific solutions.

Algorithm 1 Multi-Center Learning Algorithm

Input: A multi-center network N with initialized parameter set Θ, a training set Ω, a validation set Φ.

Output: Θ.

1: Pre-train S and C of N using ALR [19] on Ω until convergence;
2: **for** $i = 1$ to m **do**
3: Use the loss $E = \sum_{j=1}^{n} w_j[(f_{2j-1} - \hat{f}_{2j-1})^2 + (f_{2j} - \hat{f}_{2j})^2]/(2d^2)$;
4: Fine-tune C^i from C with the parameters of S fixed until convergence;
5: Save the corresponding part of center-specific landmarks in Θ as $\Theta^{i(c)}$;
6: **end for**
7: $\Theta = \Theta^S \cup \Theta^{1(c)} \cup \cdots \cup \Theta^{m(c)}$;
8: Return Θ.

When fine-tuning a center-specific layer, we give a much larger weight to the corresponding cluster of facial landmarks than other landmarks. Since landmarks from the same cluster have similar properties, they share an identical weight. For the i-th fine-tuning step, $w^{i(c)}$ and $w^{i(m)}$ denote the weight of center-specific landmarks and remaining minor landmarks respectively. Different fine-tune steps have different center-specific and minor facial landmarks. If the j-th landmark is center-specific, then $w_j = w^{i(c)}$; If the j-th landmark is minor, then $w_j = w^{i(m)}$. We assume there is a multiple relationship between two weights as

$$w^{i(c)} = \eta w^{i(m)}, \tag{1}$$

where $\eta \gg 1$ is an amplification factor. $s^{i(c)}$ refers to the number of center-specific facial landmarks. To be consisten-

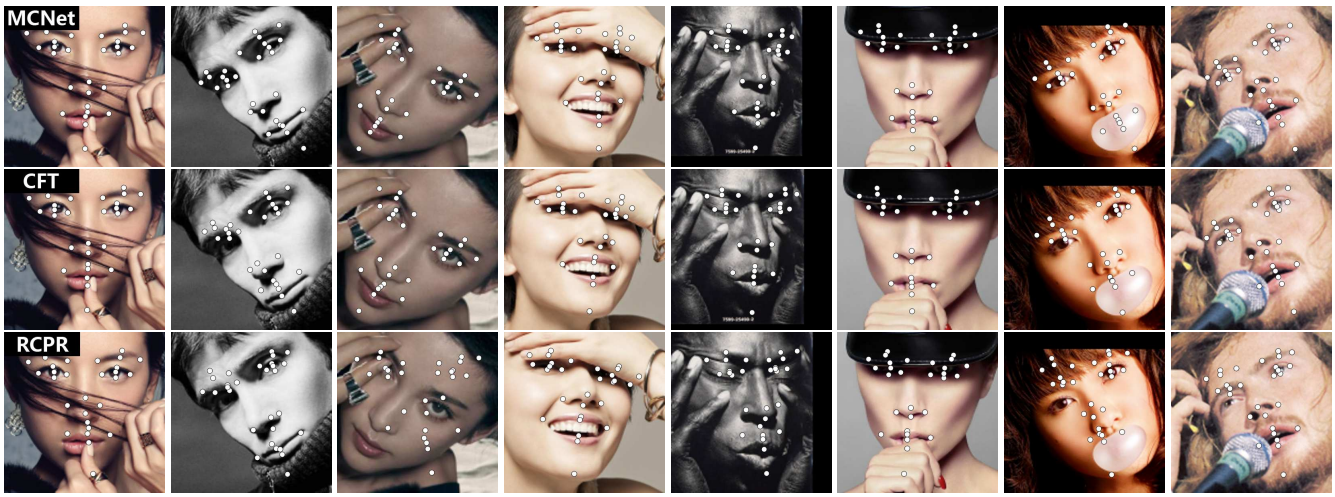


Fig. 4. Several images from COFW where our method indicates higher accuracy than RCPR and CFT in details. These examples are suffered from extreme occlusions.

t with the basic model, we keep weights conforming to the following formula

$$w^{i(c)}s^{i(c)} + w^{i(m)}(n - s^{i(c)}) = n. \tag{2}$$

By solving above two equations, we obtain the respective weights

$$\begin{aligned} w^{i(c)} &= \eta n/[(\eta - 1)s^{i(c)} + n], \\ w^{i(m)} &= n/[(\eta - 1)s^{i(c)} + n]. \end{aligned} \tag{3}$$

We train our MCNet using an open source deep learning framework Caffe [20]. In our experiments, $\eta = 125$, and the base learning rate of pre-training and each fine-tuning step are 0.02 and 0.001 respectively. It is worth mentioning that the base learning rate of fine-tuning should be small to avoid deviating from the pre-trained model overly.

4. EXPERIMENTS

In this section, we demonstrate the effectiveness of multi-center learning algorithm and compare against state-of-the-art methods on two face alignment benchmarks.

4.1. Datasets and Settings

Datasets: There are two challenging benchmarks, COFW [11] and IBUG [21], for evaluating face alignment with severe occlusion and large variations of pose, expression and illumination. COFW is an occluded dataset with $1,345$ training images and 507 testing images. IBUG includes 135 testing images with large appearance variations. When performing evaluation on IBUG, we use 3148 images from 300-W [21] for training. We employ the provided face bounding boxes to crop face patches.

Evaluation Metric: Similar to previous methods [7, 13, 16], we report the mean of inter-ocular distance normalized error, and treat the mean error larger than 10% as a failure. To obtain a more comprehensive comparison, we also plot the cumulative errors distribution (CED) curves.

4.2. Validation of Multi-Center Learning Algorithm

We validate the multi-center learning algorithm by comparing the basic model with the combined model. The results of mean error and failure rate for two models are shown in Table 1.

Table 1. Comparison of mean error (%) and failure rate (%) for the basic model and combined model.

Method	COFW		IBUG	
	Mean	Failure	Mean	Failure
Basic	6.26	3.16	9.23	33.33
Combined	6.08	2.96	8.87	25.93

It is demonstrated that the combined model has smaller mean error and failure rate than the basic model in both datasets. It is noteworthy that the basic method has already achieved a good performance, which verifies the effectiveness of our network. Our multi-center learning algorithm exploits the representation power of the network by reinforce the learning for each local face region. We can conclude that the algorithm improves the accuracy and robustness of face alignment remarkably.

4.3. Comparison with Other Methods

We develop an effective unconstrained face alignment method to compare against state-of-the-art methods including ESR

Fig. 5. Example images from IBUG where our method MCNet outperforms LBF and CFSS. These cases are challenging due to large variations of pose, expression and illumination.

[7], SDM [5], RCPR [11], LBF [8], CFSS [22], TCDCN [16], CFT [15] and Wu et al. [12]. Our method and other methods except TCDCN all learn models using given training images from the benchmark. In addition to provided images, TCDCN uses outside training data labeled with facial attributes.

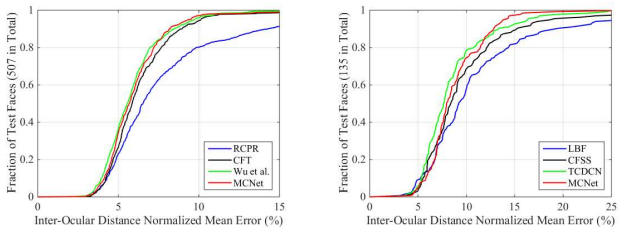
Table 2. Comparison of mean error (%) with state-of-the-art methods. Several methods did not share their results on the benchmarks, so we use results from [16] marked with "*".

Method	COFW	IBUG
ESR [7]	11.2*	17.00*
SDM [5]	11.14*	15.40*
RCPR [11]	8.5	17.26*
LBF [8]	-	11.98
CFSS [22]	-	9.98
TCDCN [16]	8.05	**8.60**
CFT [15]	6.33	10.06
Wu et al. [12]	**5.93**	-
MCNet	6.08	8.87

We report the results of our method MCNet and previous works in Table 2. We can see that our method outperforms most of the state-of-the-art methods. It is worth noting that TCDCN obtains better performance than our method on IBUG partly owing to their larger training data. Although occlusions are not detected explicitly, we achieve an outstanding performance on par with Wu et al. on COFW. Benefiting from utilizing structural correlations among different facial parts, our method is robust to severe occlusions.

We plot the CED curves for our method and several state-of-the-art methods in Figure 6. It is observed that our method achieves competitive performance on both two benchmarks, especially for high-level normalized mean error. Therefore, our method is strongly robust to unconstrained environments.

We compare with other methods on several challenging images from COFW and IBUG, as shown in Figure 4 and 5 respectively. It is obvious that our method demonstrates superior capability of handling severe occlusions and complex variations of pose, expression, illumination.



(a) CED for COFW. (b) CED for IBUG.

Fig. 6. Comparisons of CED curves with previous methods.

Our method only takes 18 ms on average to process one face on a single Intel Core i5-6200U CPU, profiting from low model complexity and computational cost of our network. We believe that our method can be extended to real-time facial landmark tracking in unconstrained scenarios.

5. CONCLUSION

We propose an effective multi-center convolutional neural network for unconstrained face alignment. Our method exhibits superior ability of handling large variations of pose, expression, illumination, and occlusion. The multi-center network is also promising for being applied in relevant research areas such as facial attribute recognition. Furthermore, it is worth exploring the multi-center learning strategy in other fields of machine learning.

6. REFERENCES

[1] Zhenyao Zhu, Ping Luo, Xiaogang Wang, and Xiaoou Tang, "Deep learning identity-preserving face space," in *IEEE International Conference on Computer Vision*. IEEE, 2013, pp. 113–120.

[2] Chen Cao, Qiming Hou, and Kun Zhou, "Displaced dynamic expression regression for real-time facial tracking and animation," *ACM Transactions on Graphics*, vol. 33, no. 4, pp. 43, 2014.

[3] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[4] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, 2001.

[5] Xuehan Xiong and Fernando De la Torre, "Supervised descent method and its applications to face alignment," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2013, pp. 532–539.

[6] David G Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[7] Xudong Cao, Yichen Wei, Fang Wen, and Jian Sun, "Face alignment by explicit shape regression," *International Journal of Computer Vision*, vol. 107, no. 2, pp. 177–190, 2014.

[8] Shaoqing Ren, Xudong Cao, Yichen Wei, and Jian Sun, "Face alignment at 3000 fps via regressing local binary features," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2014, pp. 1685–1692.

[9] Xiang Yu, Junzhou Huang, Shaoting Zhang, Wang Yan, and Dimitris N Metaxas, "Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model," in *IEEE International Conference on Computer Vision*, 2013, pp. 1944–1951.

[10] Amin Jourabloo and Xiaoming Liu, "Pose-invariant 3d face alignment," in *IEEE International Conference on Computer Vision*. IEEE, 2015, pp. 3694–3702.

[11] Xavier P Burgos-Artizzu, Pietro Perona, and Piotr Dollár, "Robust face landmark estimation under occlusion," in *IEEE International Conference on Computer Vision*. IEEE, 2013, pp. 1513–1520.

[12] Yue Wu and Qiang Ji, "Robust facial landmark detection under significant head poses and occlusion," in *IEEE International Conference on Computer Vision*, 2015, pp. 3658–3666.

[13] Yi Sun, Xiaogang Wang, and Xiaoou Tang, "Deep convolutional network cascade for facial point detection," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2013, pp. 3476–3483.

[14] Erjin Zhou, Haoqiang Fan, Zhimin Cao, Yuning Jiang, and Qi Yin, "Extensive facial landmark localization with coarse-to-fine convolutional network cascade," in *IEEE International Conference on Computer Vision Workshops*, 2013, pp. 386–391.

[15] Zhiwen Shao, Shouhong Ding, Yiru Zhao, Qinchuan Zhang, and Lizhuang Ma, "Learning deep representation from coarse to fine for face alignment," in *IEEE International Conference on Multimedia and Expo*. IEEE, 2016, pp. 1–6.

[16] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang, "Learning deep representation for face alignment with auxiliary attributes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 5, pp. 918–930, 2016.

[17] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, 2015, pp. 448–456.

[18] Vinod Nair and Geoffrey E Hinton, "Rectified linear units improve restricted boltzmann machines," in *International Conference on Machine Learning*, 2010, pp. 807–814.

[19] Zhiwen Shao, Shouhong Ding, Hengliang Zhu, Chengjie Wang, and Lizhuang Ma, "Face alignment by deep convolutional network with adaptive learning rate," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2016, pp. 1283–1287.

[20] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *ACM International Conference on Multimedia*. ACM, 2014, pp. 675–678.

[21] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge," in *IEEE International Conference on Computer Vision Workshops*. IEEE, 2013, pp. 397–403.

[22] Shizhan Zhu, Cheng Li, Chen Change Loy, and Xiaoou Tang, "Face alignment by coarse-to-fine shape searching," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4998–5006.