LEARNING DEEP REPRESENTATION FROM COARSE TO FINE FOR FACE ALIGNMENT

Zhiwen Shao, Shouhong Ding, Yiru Zhao, Qinchuan Zhang, and Lizhuang Ma

Department of Computer Science and Engineering, Shanghai Jiao Tong University, China
{shaozhiwen, feiben, yiru.zhao, qinchuan.zhang}@sjtu.edu.cn, ma-lz@cs.sjtu.edu.cn

ABSTRACT

In this paper, we propose a novel face alignment method that trains deep convolutional network from coarse to fine. It divides given landmarks into principal subset and elaborate subset. We firstly keep a large weight for principal subset to make our network primarily predict their locations while slightly take elaborate subset into account. Next the weight of principal subset is gradually decreased until two subsets have equivalent weights. This process contributes to learn a good initial model and search the optimal model smoothly to avoid missing fairly good intermediate models in subsequent procedures. On the challenging COFW dataset [1], our method achieves 6.33% mean error with a reduction of 21.37% compared with the best previous result [2].

***Index Terms*—** Deep convolutional network, coarse-to-fine, smooth search

1. INTRODUCTION

Face alignment aims to locate facial landmarks such as eyes and noses automatically. It is a preprocessing stage for many facial analysis tasks like face verification [3] and facial attributes analysis [4]. Though great success has been achieved in this field recently, robust facial landmark detection remains a challenging problem in the presence of severe occlusion and large pose variations. Most conventional methods [1, 5, 6, 7] are based on low-level features and have limited capacity to represent highly complex faces. Thus we use deep convolutional network which is effective in extracting features and robust to occlusions [3].

We discover that there are a few key landmarks which can coarsely determine face shape including brow corners, eye corners, nose tip, mouth corners and chin tip. We call a subset consists of these landmarks as principal subset, and the elaborate subset is made up of remaining landmarks. In the first step, we set a very large weight for principal subset and so the weight of elaborate subset is tiny. In this way, our network mainly predicts the location of principal subset while locates the elaborate subset roughly. During subsequent procedures, we gradually decrease the weight of principal subset, which helps to search the optimal model steadily without missing fairly good models. And entire landmarks are accurately located with the finally learned model.

Inspired by [3], we enhance the supervision by adding supervisory signal to each of the four max-pooling layers rather than only supervising the last max-pooling layer. And we employ an effective data augmentation strategy to overcome the lack of training images.

The remainder of this paper is organized as follows. In the next section, we discuss the related works of face alignment and analyse their characteristics. In Section 3 , we elaborate our coarse-to-fine training algorithm (CFT) and illuminate the structure of our deep convolutional network, following which the implementation details is exhibited. Several comparative experiments are carried out in Section 4 to show the precision and robustness of our model. Section 5 is the conclusion of this paper.

2. RELATED WORK

Significant progress on face alignment has been achieved in recent years, including conventional methods and deep learning methods.

Conventional methods: Active appearance models (AAM) [8] reconstructs entire face using an appearance model and minimize the texture residual to estimate the shape. Supervised descent method (SDM) [6] aims at solving nonlinear least squares optimization problem, which applies non-linear SIFT [9] feature and linear regressors. Both Cao et al. [5] and Burgos-Artizzu et al. [1] use boosted ferns to regress the shape increment with pixel-difference features.

These methods mainly refine the prediction of the landmarks location iteratively from an initial estimate, which is highly relevant to the initialization. In contrast, our network takes raw faces as input without any initialization.

Deep learning methods: Sun et al. [10] estimates the positions of facial landmarks with three-level cascaded convolutional networks. Zhang et al. [11] uses successive auto-encoder networks for face alignment. Both methods use multiple deep networks to locate the landmarks in a coarse-to-fine manner. They search the optimal location of landmarks from coarse to fine for each image. On the contrary, our method

This work was sponsored by the National Natural Science Foundation of China (No. 61133009 and 61472245).

contains only one network and uses coarse-to-fine strategy during training.

Zhang et al. [2] trains a deep convolutional network with multitask learning which jointly optimizes landmark detection together with the recognition of some facial attributes. It pre-trains the network by five landmarks and then fine-tunes to predict the dense landmarks. However, our method doesn't require labeling extra attributes for training samples. Different from pre-training, we also consider predicting the location of other elaborate landmarks. Compared to the method consists of pre-training and fine-tuning, we gradually adjust the weight of principal subset and elaborate subset respectively to avoid missing good models in subsequent training procedures.

3. OUR APPROACH

We propose a novel coarse-to-fine training algorithm with a good initialization and search optimal model smoothly. And our deep convolutional network has a strong ability to extract face features and predict landmarks location precisely.

3.1. Coarse-to-fine training algorithm

Since dense landmarks are expensive to label, directly trained model is apt to overfit to small training set. Therefore, we propose an innovative coarse-to-fine training algorithm. As shown in Figure 1, given landmarks can be split into principal subset and elaborate subset. The former consists of twelve key points like eye corners, nose tip and mouth corners. Indeed pupils are also key points, but we don't choose them because many face alignment datasets such as Helen [12] and 300-W [13] don't annotate pupils.

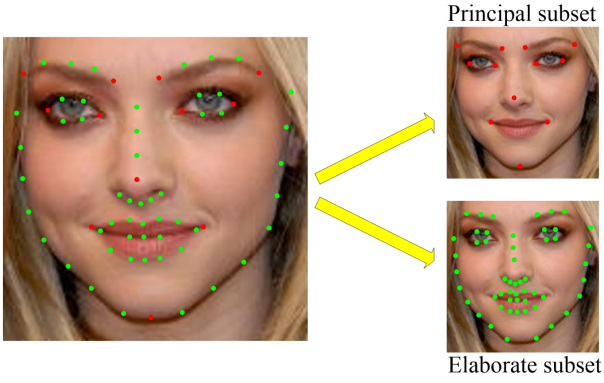

Fig. 1. Facial landmarks are divided into principal subset and elaborate subset. And the principal subset is made up of brow corners, eye corners, nose tip, mouth corners and chin tip.

Our network directly outputs the coordinate of each landmark, so we need to minimize following loss function as

$$E = \lambda E_b + (1 - \lambda) E_r, \tag{1}$$

where E_b, E_r refer to the loss of principal subset and elaborate subset respectively, which will be elaborated in Section 3.2. And λ controls the relative weight of principal subset. Our training algorithm CFT is sketched in Algorithm 1. The trainable parameters Θ are the link weights between different layers in the network.

Algorithm 1 Coarse-to-fine training algorithm.

Input: Network N with trainable initialized parameters Θ, initial control parameter λ_0, stage number $k \geq 2$.
Output: Trainable parameters Θ.

```
for i = 0 to k − 1 do
    λ = λ_0 − (λ_0 − 0.5)/(k − 1) · i;
    while not convergence do
        Training N with back propagation (BP) [14] algorithm and update Θ;
    end while
end for
```

In the beginning λ is initialized with λ_0 which is nearly equal to 1 but smaller than 1. Thus, our network mainly predicts the location of principal landmarks while remaining elaborate landmarks are still considered with the weight $1-\lambda$. This initial step lays emphasis on key points, which is beneficial for extracting essential face features. With the reduction of λ, our network searches the optimal model parameters Θ steadily until λ is equal to 0.5.

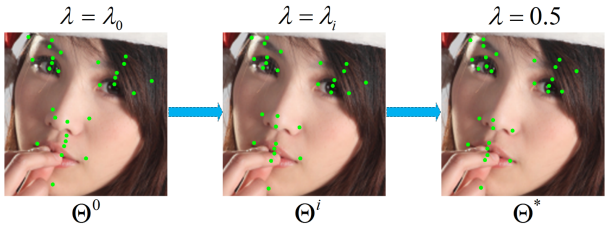

Fig. 2. Brief overview of our coarse-to-fine training algorithm. λ and Θ beside each image denote the control parameter and corresponding learned model parameters respectively in each stage. Θ^* is finally learned model parameters.

Figure 2 shows the overview of our training algorithm briefly. When finishing the first procedure, our network has been able to locate landmarks coarsely. It's clearly that the trained model is optimized stage by stage and the prediction of landmarks location using finally learned model Θ^* is very accurate.

3.2. Deep convolutional network

Our deep convolutional network mainly comprises eight convolutional layers followed by one fully-connected layer and two split fully-connected layers on behalf of principal subset and elaborate subset respectively. And every two continuous convolutional layers connect with a max-pooling layer.

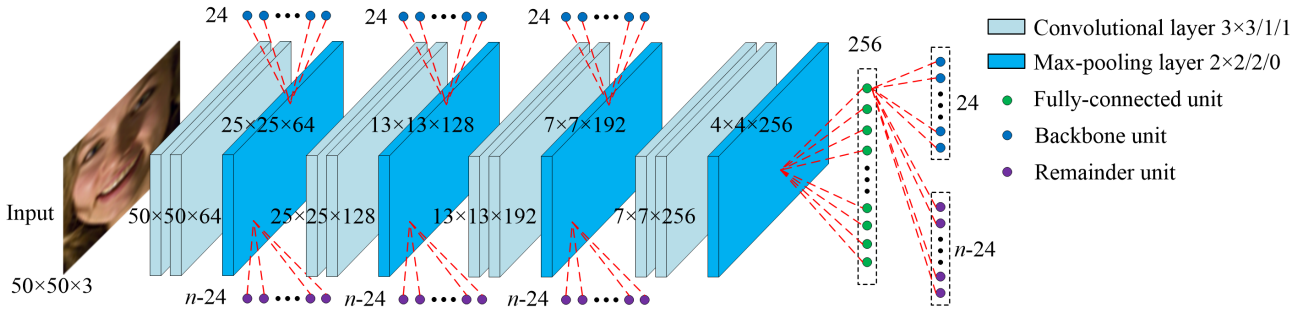

Fig. 3. The structure of our network. The equation $h \times w \times c$ beside each layer denotes that the size of map is $h \times w$ and the number of map is c. Every two continuous convolutional layers share the same equation. The equation $k_h \times k_w/s/p$ denotes that the filter size is $k_h \times k_w$, and the size of stride and padding for the filter are s and p respectively. Each convolutional layer has same filter parameters, equally applying to each max-pooling layer.

Figure 3 shows the detailed structure of our network which predicts the coordinate of each landmark in the final two split fully-connected layers. The input of our network is $50 \times 50 \times 3$ for color face patches. n is equal to double total number of landmarks, e.g. $68 \times 2 = 136$ for 300-W dataset. Besides the size of corresponding fully-connected layer for principal subset is $12 \times 2 = 24$.

Suggested by [3], we enhance the supervision by adding supervisory signal to each of the four max-pooling layers rather than only supervising the last max-pooling layer. It should be noted that the first three split fully-connected layers are connected with corresponding max-pooling layers directly without using a fully-connected layer as intermediate. The reason is that we only need to add supervisory signal to former layers, while regard final output as the prediction of landmarks location. The penultimate layer with 256 units is conducive to extract global high-level features for more accurate prediction from the final two split fully-connected layers.

In order to accelerate the training of network, we add a batch normalization layer [15] after each convolutional layer. Batch normalization is scaling and shifting the normalized input as

$$y = \gamma \hat{x} + \beta, \tag{2}$$

where $\hat{x} = \frac{x - E[x]}{\sqrt{Var[x]}}$, the expectation and variance are computed over a mini-batch from the training dataset. After normalizing each convolutional layer, ReLU nonlinearity ($y = max(0, x)$) is added to speed up convergence. We don't operate ReLU on the penultimate fully-connected layer and every two split fully-connected layers in order to preserve important information. It is worth mentioning that our network is based on VGG net [16] whose stacked multiple convolutional layers jointly form complex features.

All the four supervisory signals use same loss function defined in Equation 1. Our approach is evaluated based on alignment error measured with the distance between estimated coordinate and ground truth coordinate normalized by the inter-ocular distance. So we use normalized Euclidean distance rather than straightforward Euclidean distance to calculate loss, which is formulated as

$$E_b = \frac{\|f_b - \hat{f}_b\|_2^2}{2d^2}, \tag{3}$$

$$E_r = \frac{\|f_r - \hat{f}_r\|_2^2}{2d^2}, \tag{4}$$

where f_b and f_r are the vector concatenating the ground truth coordinate of all landmarks belonging to principal subset and elaborate subset respectively, and $\hat{f}_b$ and $\hat{f}_r$ denote predicted landmarks location correspondingly. d is the inter-ocular distance. And the coefficient $\frac{1}{2}$ makes the derivation of loss more convenient during back propagation.

3.3. Implementation details

Before starting the face alignment, we need to carry out face detection on the training images as preprocessing. Then we can acquire a face bounding box which is used for taking face patches. We conduct data augmentation since benchmark face alignment datasets such as Helen, 300-W and COFW have too small training sets.

We rotate the face image with different angles and determine new face bounding box, then slightly translate the face bounding box ensured to contain all the landmarks. It is worth mentioning that the new area contained in the face bounding box is derived from the original image rather than artificial setting which may has bad impacts on the training process.

The translation operation helps to improve the robustness of landmark detection in the condition of tiny face shift, especially in face tracking. And our model can learn to adapt complex pose variation thanks to the rotation operation. In the next steps, we horizontally flip each face patch and finally conduct JPEG compression. So our network will be trained to be robust to poor-quality images which is ubiquitous in the real case.

Table 1. Comparison of mean errors (%) with state-of-the-art methods. DT is training our network based on conventional direct training algorithm. The results of state-of-the-art methods are obtained directly from the literatures. It's worth to mention that results of some earlier methods are provided by recent papers.

Method	Helen		300-W			COFW
	194 landmarks	68 landmarks	Common Subset	Challenging Subset	Fullset	29 landmarks
ESR [5]	5.70	-	5.28	17.00	7.58	11.2
RCPR [1]	6.50	5.93	6.18	17.26	8.35	8.5
SDM [6]	5.85	5.50	5.57	15.40	7.50	11.14
LBF [7]	5.41	-	4.95	11.98	6.32	-
CFAN [11]	-	5.53	5.50	16.78	7.69	-
ERT [17]	4.90	-	-	-	6.40	-
CFSS [18]	4.74	4.63	4.73	9.98	5.76	-
TCDCN [2]	4.63	4.60	4.80	8.60	5.54	8.05
DT	5.32	5.21	5.25	10.42	6.26	6.75
CFT	4.86	4.75	4.82	10.06	5.85	**6.33**

In this way, the training face patches are increased by many times. We train our network using a deep learning framework Caffe [19], and control parameter λ_0 and k are set to be 0.995 and 3 respectively. If stage number k is too large, training our network will be time-consuming. So assigning 3 to k balances the time and accuracy suitably.

4. EXPERIMENTS

We firstly investigate the advantages and effectiveness of our coarse-to-fine training algorithm by comparing to ordinary algorithm. Then we compare our method CFT against state-of-the-art methods on three widely used benchmark datasets, Helen, 300-W and COFW. For each dataset we report the inter-ocular distance normalized error averaged over all landmarks and images, similar to most previous works.

Helen contains 2000 training images and 330 testing images, annotated densely with 194 landmarks. We also evaluate on 68 landmarks provided by [13].

300-W is created from existing datasets, including AFW [20], LFPW [21], Helen and XM2VTS [22] with annotation of 68 landmarks. In addition, it contains a challenging IBUG set. As in [7], our training set consists of AFW, the training sets of LFPW and Helen with 3148 images totally. And we perform testing with three forms: the test images from LFPW and Helen as the common subset, the IBUG as the challenging subset, and the union of them as the full set with 689 images in all.

COFW contains 1007 images annotated with 29 landmarks collected from the web. It is designed to present faces with large variations in shape and occlusions due to differences in pose, expression, use of accessories and interactions with objects. The training set consists of 845 LFPW faces and 500 COFW faces (1345 total), and the testing set contains remaining 507 COFW faces.

4.1. Algorithm discussions

The conventional training algorithm for deep convolutional network learns overall features directly, namely $\lambda = 0.5$ with only one procedure, based on randomly initialized parameters derived from a distribution. We call the conventional algorithm as direct training algorithm (DT). In contrast, our training algorithm CFT firstly concentrates on learning features of principal subset based on randomly initialized parameters. In this paper, we use standard Gaussian distribution to randomly initialize model parameters. Next, we gradually decrease the weight of principal subset and increase the weight of elaborate subset simultaneously until both are equivalent. In this way, our network can search models from coarse to fine smoothly.

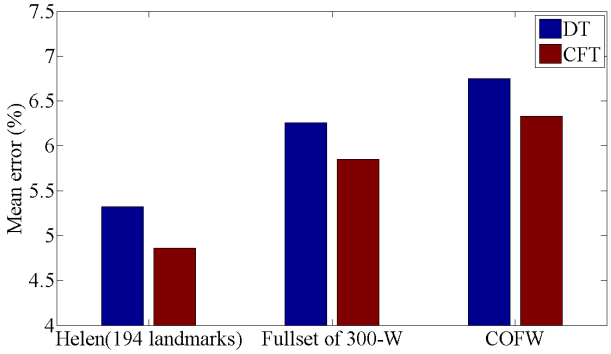

Fig. 4. Comparison between CFT and DT tested on Helen(194 landmarks), Fullset of 300-W and COFW respectively.

In order to compare CFT and DT, we also train our network using DT and test on benchmark alignment datasets. The comparison between CFT and DT is shown in Figure 4 and Table 1. Our algorithm CFT outperforms DT by a large margin, with a reduction of 8.65%, 6.55%, 6.22% on

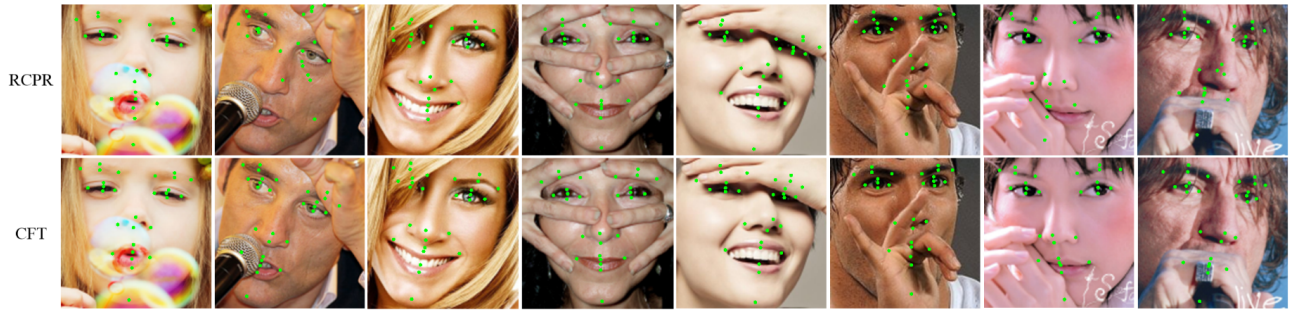

Fig. 5. The results of RCPR and CFT on several challenging images from COFW.

Helen(194 landmarks), Fullset of 300-W and COFW respectively.

4.2. Comparison with other methods

We evaluate our approach CFT on Helen, 300-W and COFW based on mean error normalised by the inter-ocular distance. And we compare with state-of-the-art methods including ESR [5], RCPR [1], SDM [6], LBF [7], CFAN [11], ERT [17], CFSS [18] and TCDCN [2] as shown in Table 1. In particular, TCDCN uses outside training images during the pre-training stage. This is unfair for other methods including ours which only use the training images provided by benchmark face alignment datasets.

It is obvious that CFT outperforms most of the state-of-the-art methods. Although the mean error of CFT tested on Helen and 300-W is slightly higher than CFSS and TCDCN, CFT performs better on challenging COFW whose faces are taken with severe occlusion. Specifically, CFT significantly reduces the error by 21.37% on the challenging COFW in comparison to the state-of-the-art TCDCN.

Sun et al. [3] prove that deep convolutional network is robust to occlusions, thus our approach can better take advantage of the superiority of deep convolutional network than TCDCN which is also based on deep convolutional network. TCDCN trains a deep convolutional network with multitask learning which jointly optimizes landmark detection together with the recognition of some facial attributes. It pre-trains the network by five landmarks and then fine-tunes to predict the dense landmarks.

Nevertheless, our method doesn't require labeling extra attributes for training samples. Different from pre-training, we also consider predicting the location of other landmarks when laying emphasis on the principal subset. Compared to the method consists of pre-training and fine-tuning, we gradually adjust the weight of principal subset and elaborate subset respectively to avoid missing optimal models in subsequent procedures.

Figure 5 shows several examples of landmark detection using RCPR and CFT respectively. It's clearly that our approach exhibits superior capability of handling complex occlusion, thanks to the good model trained by coarse-to-fine training algorithm. We also provide examples of alignment results of our approach on Helen and challenging subset IBUG of 300-W in Figure 6. It can be observed that our approach can locate landmarks accurately in the condition of complex pose, illumination and expression variations.

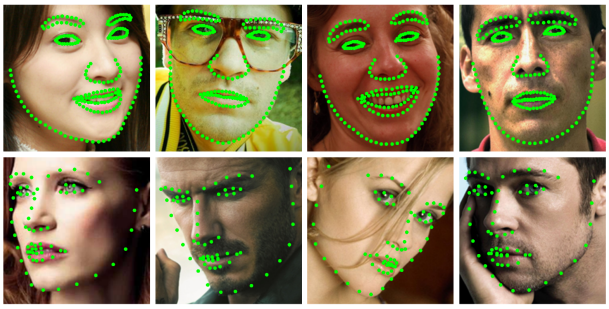

Fig. 6. Example alignment results on Helen (194 landmarks) and IBUG.

Our method takes 65 ms to process an image on a single Intel Core i7-4790 CPU. This speed is slower than 18ms of TCDCN because our network is more complicated. We will try to reduce the complexity of our network in further research.

5. CONCLUSION

We propose a novel coarse-to-fine training algorithm to train deep convolutional network for facial landmark detection. This algorithm contributes to search the optimal model steadily without missing fairly good models. Our network directly predicts the coordinates of landmarks using single network without any other additional operations, whilst significantly improves the accuracy of face alignment in the condition of severe occlusion. And we believe that the proposed training algorithm can also be applied to other problems with the use of deep convolutional network.

6. REFERENCES

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

[2] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang, "Learning deep representation for face alignment with auxiliary attributes," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, pp. 1–14, 2015.

[3] Yi Sun, Xiaogang Wang, and Xiaoou Tang, "Deeply learned face representations are sparse, selective, and robust," *arXiv preprint arXiv:1412.1265*, 2014.

[4] Ankur Datta, Rogerio Feris, and Daniel Vaqu, "Hierarchical ranking of facial attributes," in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*. IEEE, 2011, pp. 36–42.

[5] 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.

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

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

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

[9] David G Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

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

[11] Jie Zhang, Shiguang Shan, Meina Kan, and Xilin Chen, "Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment," in *Computer Vision–ECCV 2014*, pp. 1–16. Springer, 2014.

[12] Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev, and Thomas S Huang, "Interactive facial feature localization," in *Computer Vision–ECCV 2012*, pp. 679–692. Springer, 2012.

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

[14] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams, "Learning internal representations by error propagation," Tech. Rep., DTIC Document, 1985.

[15] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[16] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[17] Vahdat Kazemi and Josephine Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 1867–1874.

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

[19] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross B Girshick, Sergio Guadarrama, and Trevor Darrell, "Caffe: Convolutional architecture for fast feature embedding.," in *ACM Multimedia*, 2014, vol. 2, p. 4.

[20] Xiangxin Zhu and Deva Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2879–2886.

[21] Peter N Belhumeur, David W Jacobs, David J Kriegman, and Narendra Kumar, "Localizing parts of faces using a consensus of exemplars," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 12, pp. 2930–2940, 2013.

[22] Kieron Messer, Jiri Matas, Josef Kittler, Juergen Luettin, and Gilbert Maitre, "Xm2vtsdb: The extended m2vts database," in *Second international conference on audio and video-based biometric person authentication*. Citeseer, 1999, vol. 964, pp. 965–966.