Toward End-to-End Face Recognition Through Alignment Learning

Yuanyi Zhong, Jiansheng Chen, *Member, IEEE*, and Bo Huang

***Abstract*—A common practice in modern face recognition methods is to specifically align the face area based on the prior knowledge of human face structure before recognition feature extraction. The face alignment is usually implemented independently, causing difficulties in the designing of end-to-end face recognition models. We study the possibility of end-to-end face recognition through alignment learning in which neither prior knowledge on facial landmarks nor artificially defined geometric transformations are required. Only human identity clues are used for driving the automatic learning of appropriate geometric transformations for the face recognition task. Trained purely on publicly available datasets, our model achieves a verification accuracy of 99.33% on the LFW dataset, which is on par with state-of-the-art single model methods.**

***Index Terms*—End-to-end (e2e) training, face alignment, face recognition, spatial transformer.**

I. Introduction

THE introduction of convolutional neural networks (CNNs) have dramatically improved the state-of-the-art performances of many computer vision tasks including face recognition. Well trained CNNs are capable of handling pose, occlusion, and illumination variations of faces considerably well [1]–[7]. However, large pose variation still remains a challenge for practical face recognition systems. Introducing an explicit face alignment procedure is a commonly used approach, which can efficiently improve the recognition performance [2], [3]. Nowadays, a typical face recognition system mainly consists of four separated parts: face detection, face alignment, recognition feature extraction, and identity decision based on feature matching. As for the face alignment, recent research reveals that under the CNN based framework, three-dimensional (3-D) alignment shows no significant advantage in terms of recognition accuracy over 2-D alignment [8]. Therefore, we only focus on the 2-D face alignment in this letter.

Most existing face alignment methods [9]–[12] rely heavily on the accurate facial landmark location, a vision problem even more difficult than face recognition itself since that the manual labeling of facial landmarks is much more laborious and expensive than simply collecting personal identity information for the training. It is true that facial landmarks can be used in other interesting applications, such as expression synthesis and face beautification. What we argue here is that using the facial landmark location as a prerequisite for the face recognition might not be necessary. Also, the geometric transformation type is usually artificially defined in the face alignment. The most widely used stratagem is to align the landmarks around eyes and mouth through the nonreflective similarity transformation. Nonetheless, it is not clear whether or not the succeeding recognition feature extraction can benefit from different kinds of 2-D transformations. Generally speaking, it is weird that the face alignment still relies so much on human face structure priors and artificially defined geometric transformations while all the other parts of face recognition have been proved to be purely data-driven trainable.

Actually, the problem of learning geometric transformations has already been studied. Jaderberg *et al.* introduced a differentiable CNN component called the *spatial transformer*, which aims at improving the robustness of CNN toward translation, scaling, rotation, and even more generic image warping [14]. Due to its differentiability, a spatial transformer can be trained to learn the optimum parameterized transformation for a particular computer vision task based on a specific feature map through backward propagation. Very recently, Chen *et al.* successfully used the spatial transformer in a supervised manner for boosting the performance of face detection [15].

Inspired by these works, we study the approach of enabling simultaneous learning of the optimum face alignment together with the recognition feature extraction. More specifically, we propose a deep learning based face recognition model in which a spatial transformer module is used to supersede the process of face alignment, so that the face alignment and recognition can be unified to an end-to-end (e2e) trainable framework. Training of the proposed model does not require any explicit knowledge about the human face structure nor artificially defined alignment principles. The model is able to automatically learn the appropriate way of aligning face images so as to best suit face recognition purposes. Fig. 1 shows the projective transformations predicted by our model on testing face images with large pose variations. The model predictions intuitively comply well with the underlying transformations. This is interesting considering that no supervision signal on the transformations was ever used during model training. A related previous work is [16], in which a neural network was used to predict the geometric face transformation. However, this network was trained in a supervised manner using artificially defined transformation parameters as the ground truth. While our method enables the automatic learning of the optimum geometric transformation for face recognition driven by personal identity clues only.

Manuscript received March 14, 2017; revised May 15, 2017 and June 1, 2017; accepted June 7, 2017. Date of publication June 14, 2017; date of current version June 28, 2017. This work was supported by the National Natural Science Foundation of China (61673234). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Sanghoon Lee. *(Corresponding author: Jiansheng Chen.)*

The authors are with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (e-mail: zhongyy13@mails.tsinghua.edu.cn; jschenthu@mail.tsinghua.edu.cn; huangb14@mails.tsinghua.edu.cn).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2017.2715076

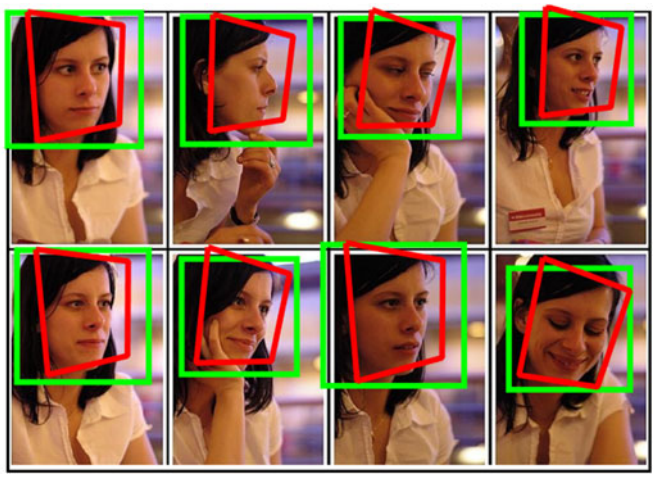

Fig. 1. Predicted transformations on testing images from the AFW [13] dataset. Green rectangles are face detection results and red rectangles visualize the model predicted projective transformations for face alignments.

II. Methodology

A. Overall Architecture

Under the traditional technological framework, heterogeneous mathematic models were adopted for different stages of face recognition. Concatenating these models for e2e training is basically intractable. Nonetheless, recent research results have already confirmed the effectiveness of CNNs on face detection [12], alignment [17], and recognition [18]. This has actually made the design and implementation of an e2e face recognition model technically possible. Ideally, such a model should be trained in an fully e2e manner so that the optimum image regions as well as the optimum transformation imposed on them can be simultaneously learned to benefit identity discrimination. However, such a model can be extremely difficult to train. To simplify the problem, we keep the face detection as an independent task and focus on the e2e implementation of the alignment and recognition tasks, as is shown in Fig. 2. Such a design is consistent with the hypothesis in cognitive neuroscience that face detection and identification might use separate dedicated resources and mechanisms in human brains [19].

For face detection, we simply stack two additional layers for the face saliency map prediction and the face bounding box regression after the third inception module of the GoogLeNet [20], and fine-tuned the fully convolutional network on the publicly available WIDER dataset [21]. This model achieves a recall rate of 86.7% at 200 false alarms on the FDDB [22] dataset. Sample face detection results are also shown in Fig. 1. We have found through experiments that face detection variations of certain degrees can be well tolerated by the proposed model. Fig. 3 shows the aligned faces corresponding to different face bounding boxes. Quantitatively, the face verification accuracy on LFW drops only 0.6% (from 99.33% to 98.73%) when the face bounding boxes are shifted by 9 pixels toward randomly selected directions.

For the alignment and recognition task, we propose an e2e network consisting of three major parts: A localization network that predicts the 2-D transformation parameters based on the input face region; a spatial transformer that warps the face image according to the predicted transformation parameters; and a deep recognition feature extraction network. Details of these parts are discussed in the following sections. It is in fact possible to implement a fully e2e face recognition system based on our proposed framework. The face detection stage can actually function as a region proposal network [23] or an attention model [24] to propose candidate face regions, so that it can be directly connected to the proposed e2e alignment and recognition network. This can be a promising future direction.

B. Localization Network

During the training stage, the detected face bounding boxes and the personal identity information are used for supervision. The input to the e2e network are cropped face regions that are resized to 128×128 pixels. The input face crops are further down-sampled to 64×64 pixels before being fed into the localization network. To decide the localization network structure, several differently structured networks are connected to a shallow recognition CNN; and e2e recognition training using human identity clues only are performed. According to the recognition performance, we adopt a neural network with three convolutional layers with kernel sizes of $5 \times 5, 3 \times 3$, and 3×3; and 24, 48, and 96 nodes, respectively. The PReLU [25] and 2×2 pooling are used after each convolution layer. After that, two fully connected layers of the size 64 are used before the regression of the geometric transformation parameters (eight parameters for projective; six for affine; four for similarity). The estimated transformation parameters are then used as input to the spatial transformer together with the 128×128 cropped face region.

C. Spatial Transformer

Theoretically, the spatial transformer can be used to implement any parameterizable transformation including translation, scaling, affine, projective, and even thin plate spline transformation [14]. In current face alignment implementations, the similarity transformation is most commonly adopted. However, Wagner *et al.* showed the effectiveness of the projective transformation for handling large pose variations [26]. Hence, we investigate three types of homogeneous transformations, namely similarity, affine, and projective, in this letter. Considering the fact that Jaderberg *et al.* only elaborated the detailed implementation of the affine transformation in [14], we hereby briefly revisit the spatial transformer by deducing the forward and backward computations for the projective transformation and the similarity transformation.

Suppose that for the ith target point $\mathbf{p_i^t} = (x_i^t, y_i^t, 1)$ in the output image, a grid generator generates its source coordinates $(x_i^s, y_i^s, 1)$ in the input image according to transformation parameters. For the projective transformation, such a process can be expressed by (1) in which A to H are eight transformation parameters and $z_i^s = Gx_i^t + Hy_i^t + 1$.

$$\begin{pmatrix} x_i^s \\ y_i^s \\ 1 \end{pmatrix} = \mathcal{T} \circ \mathbf{p_i^t} = \frac{1}{z_i^s} \begin{pmatrix} A & B & C \\ D & E & F \\ G & H & 1 \end{pmatrix} \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix}. \quad (1)$$

Then, the input image U is sampled at generated source coordinates. This is equivalent to convolving a sampling kernel k with the source image of size (W, H) as is shown in (2) in which V_i stands for the pixel value of the i_{th} point in the output

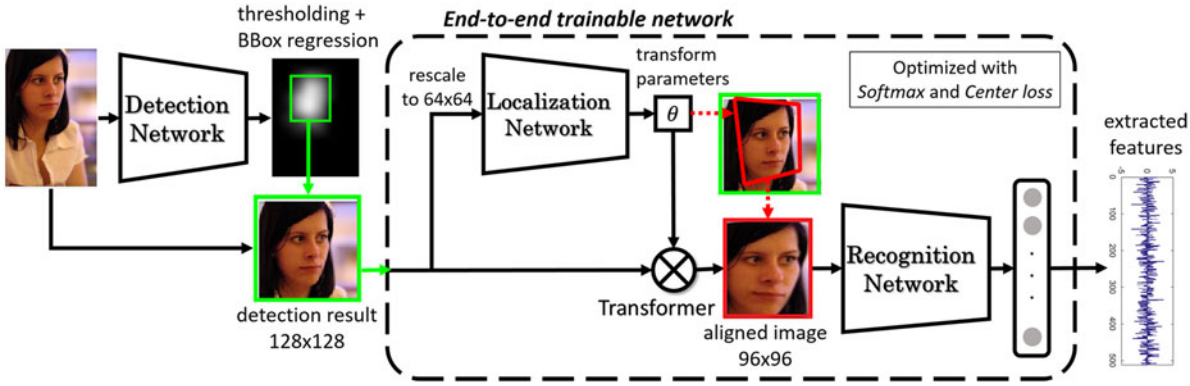

Fig. 2. Overall architecture of the proposed model. The face alignment and recognition feature extraction form an e2e trainable network.

Fig. 3. Model robustness toward face bounding box shifting. Different line styles indicate different bounding boxes and corresponding alignment results.

image. We use the bilinear kernel so that $k(w - x_i^s, h - y_i^s) = \max(0, 1 - |w - x_i^s|) \times \max(0, 1 - |h - y_i^s|)$.

$$V_i = \sum_{h=1}^{H} \sum_{w=1}^{W} U_{wh} k(w - x_i^s, h - y_i^s). \tag{2}$$

During the backward propagation, we need to calculate the gradient of V_i with respect to each of the eight transformation parameters. The function shown in (2) may not be differentiable when $w = x_i^s$ or $y = y_i^s$. However, this seldom happens since the chance that the calculated x_i^s or y_i^s are integers is very low in practice. We empirically set the gradient at these points to be 0 considering that their effect on the back propagation process are negligible. For differentiable points, the chain rule can be applied to get the gradient. An example regarding G is shown in (3), in which the gradient with respect to the source coordinates are defined in (4) and (5).

$$\begin{aligned} \frac{\partial V_i}{\partial G} &= \frac{\partial V_i}{\partial z_i^s} \frac{\partial z_i^s}{\partial G} = \left(\frac{\partial V_i}{\partial x_i^s} \frac{\partial x_i^s}{\partial z_i^s} + \frac{\partial V_i}{\partial y_i^s} \frac{\partial y_i^s}{\partial z_i^s} \right) x_i^t \\ &= -\frac{x_i^t}{z_i^s} \left(\frac{\partial V_i}{\partial x_i^s} x_i^s + \frac{\partial V_i}{\partial y_i^s} y_i^s \right) \end{aligned} \tag{3}$$

$$\begin{aligned} \frac{\partial V_i}{\partial x_i^s} &= \sum_{h=1}^{H} \sum_{w=1}^{W} U_{wh} \frac{\partial}{\partial x_i^s} k(w - x_i^s, h - y_i^s) \\ &= \sum_{h=1}^{H} \sum_{w=1}^{W} U_{wh} \max(0, 1 - |h - y_i^s|) sg(w - x_i^s) \end{aligned} \tag{4}$$

$$sg(w - x_i^s) = \begin{cases} 0, & |w - x_i^s| > 1 \\ 1, & 0 \le w - x_i^s \le 1 \\ -1, & -1 \le w - x_i^s \le 0. \end{cases} \tag{5}$$

The similarity transformation is defined in (6) in which α is the rotation angle, λ is the scaling factor, and t_1, t_2 are the horizontal and vertical translation displacements, respectively. Analogously, the gradients of V_i respected to α and λ are shown in (7) and (8), respectively.

$$\begin{pmatrix} x_i^s \\ y_i^s \\ 1 \end{pmatrix} = \mathcal{T} \circ \mathbf{p^t} = \begin{pmatrix} \lambda \cos\alpha & -\lambda \sin\alpha & t_1 \\ \lambda \sin\alpha & \lambda \cos\alpha & t_2 \end{pmatrix} \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} \tag{6}$$

$$\begin{aligned} \frac{\partial V_i}{\partial \alpha} &= \frac{\partial V_i}{\partial x_i^s} \frac{\partial x_i^s}{\partial \alpha} + \frac{\partial V_i}{\partial y_i^s} \frac{\partial y_i^s}{\partial \alpha} = \frac{\partial V_i}{\partial x_i^s} (t_2 - y_i^s) \\ &\quad + \frac{\partial V_i}{\partial y_i^s} (x_i^s - t_1) \end{aligned} \tag{7}$$

$$\begin{aligned} \frac{\partial V_i}{\partial \lambda} &= \frac{\partial V_i}{\partial x_i^s} \frac{\partial x_i^s}{\partial \lambda} + \frac{\partial V_i}{\partial y_i^s} \frac{\partial y_i^s}{\partial \lambda} = \frac{\partial V_i}{\partial x_i^s} \left(x_i^t \cos\alpha - y_i^t \sin\alpha \right) \\ &\quad + \frac{\partial V_i}{\partial x_i^s} \left(x_i^t \sin\alpha + y_i^t \cos\alpha \right). \end{aligned} \tag{8}$$

To explore the most suitable transformation type for face recognition, we will train three models with different kinds of transformations namely similarity, affine, and projective.

D. Deep Recognition Network

Inspired by [4], we use the similar deep residual network (ResNet) [27] for recognition feature extraction. The ResNet consists of 9 residual blocks with 27 convolution layers in total, producing a 512 dimensional output feature vector, which is presumably able to capture the intrapersonal variations holistically. The CenterLoss proposed in [4] is also used along with the Large Margin SoftMax loss proposed in [28] for learning discriminative features for recognition.

III. Experimental Results

Previous works have already confirmed that the face recognition accuracy can be effectively enhanced either by increasing the training set size [2] or by fusing multiple deep models in an ensemble manner [29]. However, in this letter, we mainly focus on studying the feasibility of the proposed e2e architecture as well as how different transformation types may affect the face recognition accuracy. Therefore, we used only the publicly available CASIA-Webface [30] and CelebA [31] datasets

TABLE I
FACE RECOGNITION PERFORMANCES ON LFW, YTF, AND MEGAFACE

Methods	Trainset	#Models	LFW	YTF	MegaFace
DeepFace [1]	4M	3	97.35%	91.4%	–
FaceNet [2]	200M	1	99.63%	95.1%	–
DeepID [35]	0.2M	100	97.45%	–	–
DeepID2+ [36]	0.2M	25	99.47%	93.2%	–
VGG Face [3]	2.6M	1*	99.13%	97.3%	–
ResNet(MTCNN)	0.5M	1	98.78%	94.5%	–
Center Face [4]	0.7M	1	99.28%	94.9%	65.23%
ResNet(e2e aligned)	0.7M	1	98.93%	–	–
e2e(similarity)	0.7M	1	98.65%	94.6%	–
e2e(affine)	0.7M	1	98.87%	94.7%	–
e2e(projective)	0.7M	1	99.33%	95.0%	65.16%

*The recognition feature is the average of that of 30 multiscale patches.

for training and a single deep model in recognition. Face verification experiments are performed on the LFW [32] and the YTF [33] datasets, and face identification experiments are performed on the MegaFace [34] dataset.

During training, we set the batch size to 64 images for each iteration. The center loss and the large margin softmax loss are jointly used and the coefficient of the center loss is set to 0.005, slightly smaller than that recommended in [4]. The learning rate of the recognition network is initially set to 0.01 and decay by 0.7 every 10 000 iterations. The learning rate of the localization net is 10 to 100 times smaller than that of the recognition net, as the gradient value of the transformation parameters is one to two magnitudes higher than that of the recognition net in practice. The training takes around 8 h on a TitanX GPU with about 100 000 iterations.

In the verification experiments, we adopt the 10-fold cross validation according to the standard *unrestricted with labeled outside data* protocol on both LFW and YTF datasets. We average the two feature vectors of each test image and its mirrored version as the deep feature representation for recognition. The similarity score between a pair of images is computed based on the cosine distance between the corresponding feature representations. Three e2e models using different transformation types (similarity, affine, and projective) are trained and tested, respectively. For the face identification experiments, we conduct the MegaFace Challenge 1 Set 1 [34], of which the task is to identify a person from the probe set under 1 million scale distractors. We follow the protocol of small training set by only using CASIA-WebFace for training.

Table I shows the numerical recognition performances. For identification, only the Rank-1 accuracies are listed. We trained a base-line ResNet model using CASIA-Webface [30] images aligned with MTCNN [12] and the corresponding recognition performance is shown as ResNet(MTCNN). It can be observed that our method significantly outperforms the baseline ResNet and is on par with the Center Face [4], which is also a single-model-based method trained on relatively small-sized dataset. The receiver operating characteristic curve (ROC) curves of e2e models using different transformation types are plot in Fig. 4. Comparing to the commonly used similarity transformation and the affine transformation, the projective transformation seems to be more suitable for face recognition. This is expected since the projective transformation can describe the camera imaging process more accurately.

To further reveal the advantage of the proposed e2e training, we use the trained e2e(projective) model to generate aligned

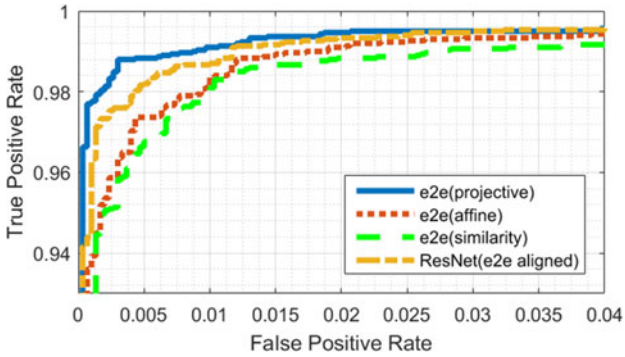

Fig. 4. ROC curves for different transformation types on LFW.

Fig. 5. Evolution of learned alignments during model training. Row 1 to row 3 correspond to similarity, affine, and projective transformations. Column 1 to column 4 correspond to iterations 1000, 7000, 14 000, and 20 000.

training face images. A ResNet recognition model with the same architecture as the one used in the e2e model is trained on the aligned images. Surprisingly, the verification performance (98.93%) of this model is significantly inferior to that of e2e(projective) model. This is probably because that at the early stage of the proposed e2e training, the inaccurate alignments automatically produce a de facto data augmentation effect on the training of the recognition network. Fig. 5 shows how the learned alignments evolve during the training process for different transformation types. To demonstrate the robustness of the proposed model toward large pose variations, we test our model on the PaSC dataset [37], [38] following the *handheld video to video* configuration. The proposed model achieves TAR=80%@FAR=1%, which significantly outperforms the best previous result of TAR=59%@FAR=1% reported on the PaSC homepage.

IV. CONCLUSION

We propose an e2e trainable framework in which face alignment and recognition are jointly trained using only the personal identities for supervision. As such, explicit knowledge about human face characteristics and artificially defined geometric transformation principles are no longer needed for face alignment for the recognition task. This indicates that despite the neuroscience evidences for the existence of special functional organizations for face perception [39], whether and how face perception differs from generic fine-grained object perception remains an open question. Our proposal actually lay a foundation for the future implementation of a fully e2e face recognition system, which can be readily applied to other new-style find-grained object recognition tasks such as the flower recognition or the animal face recognition.

REFERENCES

[1] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1701–1708.

[2] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 815–823.

[3] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis. Conf.*, 2015, pp. 41.1–41.12.

[4] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for Deep face recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 499–515.

[5] C. Ding and D. Tao, "A comprehensive survey on pose-invariant face recognition," *ACM Trans. Intell. Syst. Technol.*, vol. 7, no. 3, pp. 37:1–37:42, 2016.

[6] I. Masi, S. Rawls, G. Medioni, and P. Natarajan, "Pose-aware face recognition in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4838–4846.

[7] S. Liao, A. K. Jain, and S. Z. Li, "Partial face recognition: Alignment-free approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 5, pp. 1193–1205, May 2013.

[8] S. Banerjee *et al.*, "To frontalize or not to frontalize: Do we really need elaborate pre-processing to improve face recognition performance?" *arXiv:1610.04823*, 2016.

[9] X. Xiong and F. D. L. Torre, "Supervised descent method and its applications to face alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 532–539.

[10] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment at 3000 fps via regressing local binary features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1685–1692.

[11] D. Chen, S. Ren, Y. Wei, X. Cao, and J. Sun, "Joint cascade face detection and alignment," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 109–122.

[12] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

[13] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2879–2886.

[14] M. Jaderberg *et al.*, "Spatial transformer networks," in *Proc. Neural Inf. Process. Syst.*, 2015, pp. 2017–2025.

[15] D. Chen, G. Hua, F. Wen, and J. Sun, "Supervised transformer network for efficient face detection," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 122–138.

[16] O. Tadmor, Y. Wexler, T. Rosenwein, S. Shalev-Shwartz, and A. Shashua, "Learning a metric embedding for face recognition using the multibatch method," *arXiv:1605.07270*, 2016.

[17] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin, "Extensive facial landmark localization with coarse-to-fine convolutional network cascade," in *Proc. Int. Conf. Comput. Vis. Workshops*, 2013, pp. 386–391.

[18] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Proc. Neural Inf. Process. Syst.*, 2014, pp. 1988–1996.

[19] D. Y. Tsao and M. S. Livingstone, "Mechanisms of face perception," *Neuroscience*, vol. 31, no. 31, pp. 411–437, 2008.

[20] C. Szegedy, W. Liu, Y. Jia, and P. Sermanet, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[21] S. Yang, P. Luo, C. C. Loy, and X. Tang, "WIDER face: A face detection benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5525–5533.

[22] V. Jain and E. Learned-Miller, "FDDB: A benchmark for face detection in unconstrained settings," Univ. Massachusetts, Amherst, MA, USA, *Tech. Rep.* UM-CS-2010–009, 2010.

[23] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[24] T. Xiao, Y. Xu, K. Yang, and J. Zhang, "The application of two-level attention models in deep convolutional neural network for fine-grained image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 130–160.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1026–1034.

[26] 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.*, vol. 34, no. 2, pp. 372–386, Feb. 2012.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[28] W. Liu, Y. Wen, Z. Yu, and M. Yang, "Large-margin SoftMax loss for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 507–516.

[29] Y. Sun, D. Liang, X. Wang, and X. Tang, "DeepID3: Face recognition with very deep neural networks," *arXiv:1502.00873*, 2015.

[30] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv:1411.7923*, 2014.

[31] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. Int. Conf. Comput. Vis.*, 2015, pp. 3730–3738.

[32] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Univ. Massachusetts, Amherst, MA, USA, *Tech. Rep.* 07–49, 2007.

[33] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 529–534.

[34] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard, "The MegaFace benchmark: 1 million faces for recognition at scale," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4873–4882.

[35] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1891–1898.

[36] Y. Sun, X. Wang, and X. Tang, "Deeply learned face representations are sparse, selective, and robust," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 2892–2900.

[37] J. R. Beveridge *et al.*, "The IJCB 2014 PaSC video face and person recognition competition," in *Proc. IEEE Int. Joint Conf. Biometrics*, Clearwater, FL, USA, Sep. 29–Oct. 2, 2014, pp. 1–8.

[38] J. R. Beveridge *et al.*, "Report on the FG 2015 video person recognition evaluation," in *Proc. 11th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit.*, Slovenia, May 4–8, 2015, pp. 1–8.

[39] N. Kanwisher, J. Mcdermott, and M. M. Chun, "The fusiform face area: A module in human extrastriate cortex specialized for face perception," *J. Neurosci.*, vol. 17, no. 11, pp. 4302–4311, 1997.