LEFV: A Lightweight and Efficient System for Face Verification with Deep Convolution Neural Networks

Ming Liu
State Key Laboratory of Mathematical Engineering and Advanced Computing
Zhengzhou China
lm_puree@outlook.com

Ping Zhang
State Key Laboratory of Mathematical Engineering and Advanced Computing
Zhengzhou China
Zhangping@163.com

Qingbao Li
State Key Laboratory of Mathematical Engineering and Advanced Computing
Zhengzhou China
Qingbao_Li@163.com

Jinjin Liu
State Key Laboratory of Mathematical Engineering and Advanced Computing
Zhengzhou China
liujinjin0809@zzti.edu.cn

Zhifeng Chen
State Key Laboratory of Mathematical Engineering and Advanced Computing
Zhengzhou, China
xiaohouzi06@163.com

ABSTRACT

The emergence of deep learning has made great progress in face recognition. With the popularization of embedded devices, deploying the deep model on embedded devices has become a trend. Most high-precision models require lots of computation costs. Therefore, developing a lightweight deep face recognition system running on embedded devices is a hot topic in current research. To achieve high-accuracy real-time performance of an embedded device, we now present a simple and effective face recognition system LEFV, including face detection, face normalization and face recognition. The quantitative experiments on two large-scale challenging datasets, WIDER FACE dataset and IJB-A dataset, show competitive performances on both runtime and accuracy.

CCS Concepts

•Computing methodologies → Object recognition

Keywords

Deep Convolution Neural Networks; Face Detection; Face Verification; Lightweight Model

1. INTRODUCTION

Over the past two decades, face recognition has become one of the research hotspots in many fields. Although these previous works have achieved high accurate and satisfactory results in some tasks, these also have high computational costs. Despite significant progress, these models have not been adequate for industrial application. Usually, most tested images and videos are collected in unconstrained situations. These photos contain the influence of gestures, expressions, lighting, occlusion and other actors. These factors make the algorithm performance deteriorating rapidly. In many practical face recognition scenarios, face images are difficult to accurately align because of complex appearance variations or low-resolutions faces. There are two main ways to deal with complex faces: the recover category and the aggregation category.

Figure 1. Examples from test dataset. The top row depicts some normal face. The bottom row shows some challenging face image.

The first methods can generate local and global features to reconstruct frontal faces from profile faces caught in the wild. Several research [4, 12, 16, 19] attempt to use Generative Adversarial Network (GAN) generating new gallery images to improve the accuracy of recognition models. TP-GAN [4] adopts a double pathway generator to synthesize local face block and generate the structural information of global. To make some changes of standard GAN, the generator of DA-GAN [6] exactly recovers realistic face image meanwhile preserves identity information with double agents. Yu et al. [24] firstly propose an FF-GAN to deal with extreme head pose expression and self-occlusion. It involving the 3D Morphable Models(3DMM) with GAN can repair extreme yaw angle faces. The experiment results of these works vastly enhance the performance in following face recognition.

Different from the recover-based models, the aggregation models directly extract discriminative features from original face images and then assign a certain weight to aggregate a comprehensive

ICVIP 2019, December 20–23, 2019, Shanghai, China
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-7682-2/19/12…$15.00
https://doi.org/10.1145/3376067.3376077

feature representation. The Aggregation Module of NAN [11] is an innovation work, which can produce a comprehensive robust vector representing face image set of same people. MPNet [3] introduces a Dense Subgraph (DSG) to learn multiple prototype set-level representations implicitly.

Based on the unconstrained face recognition, this paper proposes a lightweight and efficient recognition system that is more consistent with practical application. Thus, we proposed an automatic face validation system consists of three separate deep convolution neural network model. Firstly, a real-time face detection framework is designed, running on a fast speed while achieving high accuracy. After that, we input the face area to the face normalization model to finally generate a frontal, expression-free face image for face recognition. Inspired by lightweight networks [10, 13, 14, 20], we designed a lightweight yet powerful CNN model to extract discriminative facial features. Finally, we compared our methods with other methods in the IJB-A databases. Although IJB-A data have a lot of challenges in posture, light, expression, resolution, we also achieved a competitive result. Compared with previous work, our main contributions are concluded as follows:

- First, combined with network pruning and optimizing strategy, we design a light detector network which runs at a high speed on an edge device
- Second, face normalization network introduces a serious of effective loss for reconstructing photorealistic faces from the normal face set,
- Third, to extract sufficiently discriminative features, we add a Multi-Scale Feature Fusion (MSFF) module, combining features at different levels in a novel way. Qualitative and quantitative experiments compared with other light networks and large-scale network demonstrate our system feasible and efficient.

2. Related Work

This paper aims at designing an automatic face verification system running on mobile devices. We briefly review the previously mentioned works from four aspects, including face detection, model compression, generative adversarial network and face recognition models.

2.1 Face Detection

Face detection is finding all the face positions in the image and returns the coordinates of the face feature points. It usually marked with a rectangular box. Early works about face detection mainly depend on manual annotation and classifiers. A recent study [18] cascaded the CNN model to construct multi-task learning network. Its main structure is similar with cascade. An end-to-end face detection framework [5, 7] with a scale invariability, can handle face image of any size just once. Without cascade structure [20], the full connection layer is replaced by full connection convolution layer, leading to the size of the input image was not restricted. Face tracking is an important module in the face recognition. Bai et al. [2] find out the importance of peripheral information to detect small targets. Therefore, they trained separate multi-task detectors to detect tiny face refined. Nevertheless, the works are still not appropriate to mobile devices in terms of the model sizes and computation.

2.2 Model Compression

Researchers are paying rising attention to the lightweight deep convolutional neural framework design. The previous work is mainly divided into: design compression model, weight sharing, kernel thinning, network pruning, binary network, quantization network, etc. We pay more attention on the first two. MobileNet [10] proposed a deep separable convolutional neural network to reduce the parameters of convolution layers. VGG-16[1] and MobileNet [10] have the same accuracy in ImageNet, but the parameters of the former are 32 times than the latter. Although MobileNet, ResNet and other networks can significantly reduce the complexity significantly and the performance does not deteriorate greatly. A large number of 1×1 convolution operation still cost vast resources. For example, in ResNet and MobileNet, 1×1 convolution operation account for 93% and 94.86% respectively. [20] the author proposes an operation that adds a channel shuffle to the middle feature map layer, so that each group can accept the features of different groups at the previous layer. They solve the problems mentioned above.

2.3 Generative Adversarial Network

Inspired by gaming theory, GAN has received extensive attention from deep learning and computer vision. Chen et al. [16] maximize the mutual information between the generated image and the input image. Explicable features can be obtained without supervised learning and additional computational costs. Shrivastava et al. [19] propose simulated and unsupervised learning method to calculate synthetic images and remove artifact. It significantly reduces the cost of training. Huang et al. [4] propose a GAN network which can consider whole and local information like human. When transforming a non-frontal face into a gallery image, identity information can be preserving perfectly. The accuracy of face recognition in a large angle can be greatly improved. The success of these works [4, 12, 16, 19] inspired us to use GAN to correct human faces in difficult and unrestricted circumstances and in cases of mismatches.

2.4 Deep Face Recognition

Compared with general machine learning methods, convolutional neural network has deeper layers and learns deeper related advanced features of human face, which is obviously superior to artificial features in face recognition. The best face recognition methods often rely on the advanced architecture of deep CNN. At the end of 2014, Sun et al. [15] proposed to use both authentication signal and identification signal to train the network, so as to boost the discriminability of the model. Schroff et al. [8] design FaceNet in 2015, which adopted the triplet loss to train the network. Parkhi et al. [10] also trained the VGG network based on minimizing triple losses. In contrast, Sun et al [21]. employed a larger training network and stronger supervisory training (i.e., applying both authentication and identification supervision to each layer of network characteristics). Under such strong supervision, Sun et al. found that the features formed by the trained network in the highest hidden layer began to have moderate sparsity, robustness to occlusion and selectivity to higher-order attributes.

3. Proposed System

The system serves as an integrated process of automatic validation process in which face will be detected in every image or video frame, followed by passing face areas to face normalization and compute the similarity of images/videos. Figure 1 shows the system. The details of each component will be introduced in the following sections.

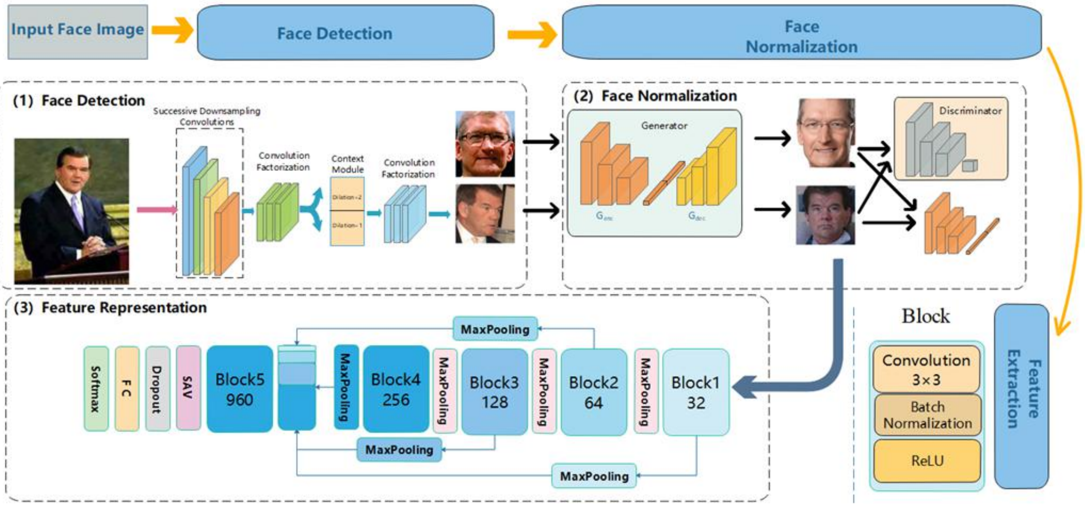

Figure 2. Architecture of our system. There are mainly three part of it. The face detection shown in (1), a face in the wild fed into detector. Then, we pass the face area to face normalization in (2) and perform face validation to calculate the similarity between a pair of images/video in (3).

3.1 Face Detection

3.1.1 Depthwise Separable Convolutions

The convolution factor decomposition， the primary strategy for reducing the complexity of face detection network, converts each standard convolution layer into a depth-wise convolution layer and a point-wise convolution layer. The depth-wise convolution is firstly proposed in Xception [14]. Using this method, we replace each standard convolution in the backbone with depthwise separable convolution. Convolution factorization has two advantages:(1) The network parameters are greatly reduced. (2) The computational complexity is also reduced.

Table 1. Complexity comparison

Convolutions	**Parameter**	**Computation**
Standard	n × n× N_{in} × N_{out}	H × W × n × n× N_{in} × N_{out}
Depthwise Separable	n × n× N_{in} + 1 × 1× N_{in} × N_{out}	H × W × n × n× N_{in} + 1 × 1× N_{in} × N_{out} × H × W

3.1.2 Successive Downsampling Convolutions

With limited resources, a compact network should maximize the input information to the output layer and avoid excessive computing costs. Some scaling methods can increase the information of the image so that the quality of the scaled image exceeds the quality of the original picture. Assuming the size of a picture is W × H, s times lower sampling can be carried out to generate the resolution image of (W/s × H/s) size. Outlining the details in the scene serves as its most outstanding function of it. We take four successive subsamples of the input picture. During the initial stage of feature embedding, the lower sampling step can be used continuously to avoid feature mapping in large space dimension. With the increasing depth of each stage, it becomes more focused on specific features. Due to the low computing power of embedded devices, we use fewer layers to keep the low computational complexity of backbone.

3.1.3 Context Module

Context module is one of the most effective ways to improve detection accuracy without additional computational cost. Different from the general object detection task, the scene in the whole picture can provide clues for the detection of the target. Cars often appear in the street and airplanes often appear in the blue sky, then faces can appear in many scenes. Therefore, we use a smaller dilation convolution parameter to restrict the context area. This reduces the number of channels in each branch. For example, due to the structure of the human body.

Moreover, Leaky ReLU was selected as the activation function instead of the commonly used ReLU in the experiment.

$$\text{ReLU} = \begin{cases} x & if\ x \geq 0 \\ 0 & if\ x < 0 \end{cases} \tag{1}$$

$$\text{LeakyReLU} = \begin{cases} x & if\ x \geq 0 \\ \lambda x & if\ x < 0 \end{cases} \tag{2}$$

The framework of the face detection is shown in Figure 2. Although leaky ReLU adds a small amount of computational cost compared to ReLU. But this increase is trivial compared to the total convolution. We experimentally demonstrated the improvement in performance using leaky ReLU.

3.2 Face normalization

Real-time face recognition systems still face many challenges. These challenges come mainly from the various kinds of unpredictable changes that may exist in face images, such as expression, posture, light, resolution, and obscuring. These changes may lead to serious data migration between the training image and the identified image. Due to the influence of expression, illumination and occlusion, it will affect the recognition results if feature directly extracted from the original image. Therefore, we will regularize the face region obtained in 3.1. The purpose of face normalization is to recover a normalized view from a profile face. We have carefully designed the face regular network. Different from the previous work, our proposed normalization network inherits a lightweight face expert network, which has outstanding discrimination ability to solve the problem of complex changes and unpaired image. By introducing pixel-level loss, leading to more stable optimization, we generate a high-quality face image. As shown in Figure 2, we use a public lightweight face expert network MobileFaceNets [23] as the encoder of the generator. It has a strong prior knowledge in face recognition. In order not to add extra computations, we fixed the parameters of network training to reduce the backpropagation operation.

We use the face expert network MobileFaceNet [23] as the encoder of the generator, mapping the input image to the feature space, denoted as G_{enc}. An ideal generator will make abnormal faces converted to realistic faces while maintaining identity information. A decoder G_{dec}, is designed to further try to preserve the identity information. It is harder to extract facial identity from the unconstrained face than to generate a standardized face from the identity. We introduced a series of discriminators to distinguish real positive face images and generated positive faces. Considering the structural characteristics of human face, these several identify four regions to identify. We input the eyes, nose, mouth and whole face respectively. As you can see, we built four discriminators. Unlike the general image generation, local facial texture is more important to face synthesis. Specifically, the generator is stated as $\mathrm{G} = G_{enc} \circ G_{dec}$. Let face image from non-normal set be denoted by $\mathrm{x} \in \Re^{H \times W \times C}$ and the responding generated face image be denoted by $\bar{x} \in \Re^{H \times W \times C}$, then

$$\bar{x}=G_{dec}(G_{enc}(\mathrm{x})) \tag{3}$$

To this end, we also input normal face to G:

$$\bar{y}=G_{dec}(G_{enc}(\mathrm{y})) \tag{4}$$

where y denotes the elements of the normal face set, and $\bar{y}$ denotes generated new face image. In this way, we use pixel-wise loss to guides the optimization of GAN in condition of unpaired data.

Synthesizing face should keep the identity consistency of the input face and output face. make it look like a real face. Two basic losses are introduced to meet the requirements, denoted by L_{adv},L_{ip} as following:

$$L_{adv} = \sum_{k=1}^{4} D_k(\bar{x}_k) + \sum_{k=1}^{4} D_k(\bar{y}_k) - \sum_{k=1}^{4} D_k(y_k) \tag{5}$$

$$L_{ip} = \|G_{enc}(x) - G_{enc}(\bar{x})\|_2^2 + \|G_{enc}(y) - G_{enc}(\bar{y})\|_2^2 \tag{6}$$

k represents the number of attention discriminator. L_{adv} is an adaptive confrontation loss, which is used to add reality to the generated picture. L_{ip} is used for storing the identity information. It is still an essential part for preserving and accelerated optimization of both pose and combination information. $\|^{\mathrm{q}}\|_2^2$ denotes the square of the 2-norm.

Symmetry is an inherent feature of human face. Using the prior knowledge of biology to constrain the synthetic image can effectively increase the robustness of extreme posture. Specifically, Laplace space is used to apply symmetry to the image that recovers the front face, so as to alleviate the problem of self-occlusion:

$$L_{sym} = \frac{1}{\frac{W}{2} * H} \sum_{i}^{W/2} \sum_{j}^{H} \left|\bar{x}_{i,j} - \bar{x}_{W-(i-1),j}\right| \tag{7}$$

where W, H represent width and height of the final recovery image $\bar{x}$. i, j traverses each pixel in the image $\bar{x}$.

L_p is a pixel-wise loss, which is introduced for the consistency improvement of pose as well as the consistency detailing of synthesized face image and original image via GAN:

$$L_p = \frac{1}{W \times H \times C} \sum_{w,h,c}^{W,H,C} \left|y_{w,h,c} - \bar{y}_{w,h,c}\right| \tag{8}$$

where w, h, c traverses all pixels and channels of y and $\bar{y}$.

The overall objective function of model A2 is:

$$\begin{cases} L_{G_{dec}} =- L_{adv} + \lambda_1 L_{ip} + \lambda_2 L_p + \lambda_3 L_{sym} \\ L_{D= L_{adv}} \end{cases} \tag{9}$$

3.3 Deep Convolutional Face Representation

More and more characteristics extraction and learning methods are being applied to the identification of face recognition and greatly improved recognition performance. For inferring quickly, the network should be shallow, with few parameters, thereby reducing memory costs. Therefore, our backbone network consists of only four blocks with 32, 64, 128, and 256 convolutions filters respectively. Each block has the same structure as shown in Figure 2, the blocks are very compact. For improving the accuracy, we introduce a multi-scale feature fusion (MSFF) module. MSFF is used to fuse the multi-scale features of each block into the comprehensive representation, while chav emphasizes the important spatial information, which both improve the discrimination ability of the result features. Then we applied a Dropout layer and softmax layer with cross entropy loss to guide network training. The overall architecture is shown in Figure 2. The features of a face image are usually extremely rich, including low levels of color, brightness, texture, direction, etc., and include gestures, expressions, age, and race. The various characteristics of the image are combined in a nonlinear way, which is easy to interfere with each other, especially when the noise characteristics are mixed with the normal characteristics, and the noise characteristics may play a leading role in the classification results.

Each of the parameters of the CNN network will affect the feature map. Meanwhile, these parameters are often closely related to the current task and cannot be easily modified. So, it is difficult to learn multi-scale information. This network structure is designed to explore the characteristics of different scales at the beginning of the design, and then combine the prediction results of different scale characteristics to obtain the final output. CNN is made up of multiple convolution layers. Each layer learns different information: the lower layer captures the underlying elements, such as basic colors and edges, while the higher layer encodes abstractions and semantic cues. For better presentation, it is natural to combine features from different layers. For example, DenseNet [9] uses very dense connections to integrate these features. However, such connections are very heavy, resulting in expensive computing costs.

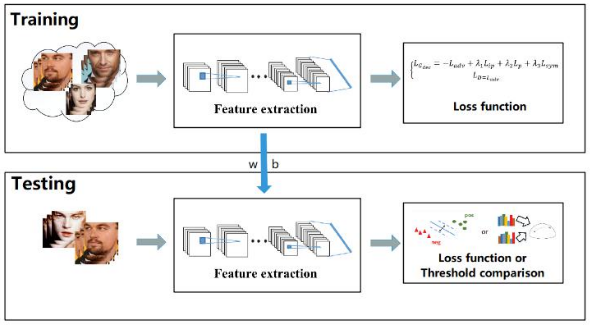

Figure 3. Extract representation for recognition

The single scale of the convolution neural network has a strict feed-forward hierarchical structure. The multi-scale network not only inputs the last features to the classifier, but also the features proposed by other convolutional layers are also fed directly to the classifier by branches. The shallow convolutional network can better represent the local detailed features. Multi-scale convolutional network can obtain composite features including lower layers of images and more local features, thus effectively improving the ability of feature expression and the final model recognition effect. In the following experimental chapter, we prove the effect of such multi-scale features can improve the recognition rate.

In particular, we extract feature mappings from the four convolution blocks corresponding to the information captured by different facial regions. Then, all feature mappings are sampled through the max-pool. We integrated feature maps of different scales with another convolution layer consisting of 960 convolution filters (Block 5), so that we could produce a more discriminative feature representation.

4. Experiments

In this section, we first investigate some important ablation tests to prove the effect of each optimizing strategy. We then show the experimental results on IJB-A and WIDER-FACE databases. All networks are trained on two NVIDIA Tesla V100 GPUs. We use Xavier method to initialized parameters. Experiment results on the Raspberry Pi1 3b+ and Jetson TX2 shows the efficiency of face detector and whole system feasibly.

4.1 Databases

IJB-A [22]: It aims to promote the development of unconstrained face detection and recognition. It includes not only the still image of the person being photographed, but also video fragments of the person being photographed, about 11.4 images and 4.2 videos per subject. All media are collected in a completely unconstrained environment. Many of the people photographed had huge changes in facial posture, dramatic changes in lighting, and different image resolutions. The subjects are also from different countries, regions and RACES of the world, with a wide range of regions. It is because the IJB-A data set has realistic application features that the data set is very suitable for practical application scenarios. Of course, it also brings great challenges.

WIDER FACE [17]: It is a large and richly annotated face detection data.32, 203 images (with 393,703 tagged faces) were selected from a number of public databases or searched by search engines. Each image usually contains multiple faces. The database focuses more on face detection at small scales, occlusion, and extreme poses. These factors are common in reality monitoring video. In each category, there were faces that differed significantly. We divide them into three categories according to difficulty

4.2 Ablation Study

We evaluate the contributions of : Convolution Factorization (CF), Successive Downsampling Convolutions (SDC) and Context Modules (CM). We train four networks: (1) the basic network with only seven stages following the VGG-style; (2) the basic network with CF; (3) the basic network with CF and SDC; and (4) the basic network with both the CF, SDC and CM. Table2 shows that with the same network structure, the performance of our model lows a little than the baseline, but the computation cost of ours is far much less than the baseline.

Table 2 Ablation study on WIDER-FACE

Contribution				**Baseline**
CF	○	○	○	
SDC		○	○	
CM			○	
Accuracy(mAP)	83.7%	82.5%	83.3%	87.4%
FLOPS	87.3M	73.4M	78.5M	440.4M

To verify the efficiency of face detector, we also make compared our method with a lightweight network MTCNN [18] and a large-scale high-precision network HR [13]. Tables 3 shows the results of the face detector ran on Pi 3b+.

Table 3 Performance comparison of face detecting on WIDER-FACE

Method	**Easy**	**Medium**	**Hard**	**Model Size**
MTCNN[18]	0.836	0.809	0.622	1.90MB
HR[2]	0.862	0.844	0.749	98.9MB
Ours	0.833	0.798	0.565	0.95MB

4.3 Face Verification on IJB-A

Face normalization model can generate high-fidelity and identity-preserved normal face on unconstrained dataset IJB-A. There are intricate face variations in unconstrained environment. These results demonstrate robustness of normalization model to large pose, lighting, occlusion and expression. It's worth noting that the difficulty of face normalization on unconstrained environment lies in not only extreme variations but mixture of these variations. With the robustness of face recognition model, our model generates normalized face from high-level semantic feature instead of image. As our objective is face normalization with unpaired data, we strictly keep the same setting on constrained and unconstrained environments. In particular, we do not use paired data and identity information under both environments, which is available in controlled environment.

As shown in Table 4, it shows our model compared to some state-of-art methods, our method performs a competitive result at FAR=0.001 and FAR=0.01. It is noteworthy that our method perform fastest speed on mobile devices.

Table4 Performance comparison term of TAR evaluation on IJB-A

Method	**1:1 verification TAR(%)**		
	FAR=0.001	FAR=0.01	FAR=0.1
NAN [11]	88.10±1.10	94.10±0.80	97.80±0.30
AvgPool	88.82±1.22	96.18±0.92	98.16±0.40
QAN [24]	89.31±3.92	94.20±1.53	98.02±0.55
VGGFace2	92.10±1.40	96.80±0.60	99.0±0.20
Ours	89.82±1.16	94.20±1.53	98.26±0.40

We also evaluate comparative experiments on Jetson TX2, and compared it with large-scale high-precision models and lightweight model. The experimental results show that the proposed method is more suitable for embedded devices with limited computing power.

Table 5 Running speed (fps) on TX2

Model	Jetson TX2	
	GPU	ARM
VGG-16	7.09	0.43
ResNet-34	8.44	0.58
Ours	25.41	7.90

5. Conclusion

In this paper, we build a lightweight, efficient face recognition system with three independent deep CNN models. We carefully designed the backbone of each model and cleverly combined the of optimization strategies. The proposed model has a high accuracy recognition rate in natural scenes and real-time capability. At the same time, it can be deployed in small embedded devices, which promotes the application of face recognition in a wider range.

REFERENCES

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

[2] Hu, Peiyun, and Deva Ramanan. Finding tiny faces. *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017. DOI= https://doi.org/10.1109/CVPR.2017.166

[3] Zhao, Jian, et al. Multi-Prototype Networks for Unconstrained Set-based Face Recognition. arXiv preprint arXiv:1902.04755 (2019).

[4] Huang, Rui, et al. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. *Proceedings of the IEEE International Conference on Computer Vision*. 2017. DOI= https://doi.org/10.1109/ICCV.2017.267

[5] Zhang, Shifeng, et al. S3FD: Single shot scale-invariant face detector. *Proceedings of the IEEE International Conference on Computer Vision*. 2017. DOI= https://doi.org/10.1109/ICCV.2017.30

[6] Zhao, Jian, et al. "3d-aided dual-agent gans for unconstrained face recognition." *IEEE transactions on pattern analysis and machine intelligence* (2018). DOI= https://doi.org/10.1109/TPAMI.2018.285881

[7] Najibi, Mahyar, et al. SSH: Single stage headless face detector. *Proceedings of the IEEE International Conference on Computer Vision. 2017.*DOI= https://doi.org/10.1109/ICCV.2017.522

[8] Schroff, Florian, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015. DOI= https://doi.org/10.1109/CVPR.2015.7298682

[9] Iandola, Forrest, et al. Densenet: Implementing efficient convnet descriptor pyramids.arXiv preprint arXiv:1404.1869(2014).

[10] Howard, Andrew G., et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications.arXiv preprint arXiv:1704.04861(2017).

[11] Yang, Jiaolong, et al. Neural aggregation network for video face recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2017. DOI= https://doi.org/10.1109/CVPR.2017.554

[12] Shu, Zhixin, et al. Neural face editing with intrinsic image disentangling. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017. DOI= https://doi.org/10.1109/CVPR.2017.578

[13] Iandola, Forrest N., et al. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size.arXiv preprint arXiv:1602.07360(2016).

[14] Chollet, François. Xception: Deep learning with depthwise separable convolutions. *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017. DOI= https://doi.org/10.1109/CVPR.2017.195

[15] Sun, Yi, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation from predicting 10,000 classes. *Proceedings of the IEEE conference on computer vision and pattern recognition. 2014*. DOI= https://doi.org/10.1109/CVPR.2014.244

[16] Chen, Xi, et al. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems*. 2016.

[17] Yang, Shuo, et al. Wider face: A face detection benchmark. *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2016. DOI= https://doi.org/10.1109/CVPR.2016.596

[18] Zhang, Kaipeng, et al. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters* 23.10 (2016): 1499-1503. DOI= https://doi.org/10.1109/LSP.2016.2603342

[19] Shrivastava, Ashish, et al. Learning from simulated and unsupervised images through adversarial training. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. DOI= https://doi.org/10.1109/CVPR.2017.241

[20] Zhang, Xiangyu, et al. Shufflenet: An extremely efficient convolutional neural network for mobile devices. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 2018. DOI= https://doi.org/10.1109/CVPR.2018.00716

[21] Sun, Yi, Xiaogang Wang, and Xiaoou Tang. Deeply learned face representations are sparse, selective, and robust. *Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.* DOI= https://doi.org/10.1109/CVPR.2015.7298907

[22] Klare, Brendan F., et al. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2015. DOI= https://doi.org/10.1109/CVPR.2015.7298803

[23] Chen, Sheng, et al. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. arXiv preprint arXiv:1804.07573

[24] Yin, Xi, et al. Towards large-pose face frontalization in the wild. Proceedings of the *IEEE International Conference on Computer Vision. 2017*. DOI= https://doi.org/10.1109/ICCV.2017.430