IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 29, 2019, accepted November 19, 2019, date of publication November 25, 2019,
date of current version December 11, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2955757

SFA: Small Faces Attention Face Detector

SHI LUO, XIONGFEI LI, (Member, IEEE), RUI ZHU, AND XIAOLI ZHANG
Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, China
College of Computer Science and Technology, Jilin University, Changchun 130012, China

Corresponding author: Xiaoli Zhang (xiaolizhang@jlu.edu.cn)

This work was supported in part by the National Science and Technology Pillar Program of China under Grant 2012BAH48F02, in part by the National Natural Science Foundation of China under Grant 61801190, in part by the Nature Science Foundation of Jilin Province under Grant 20180101055JC, in part by the Outstanding Young Talent Foundation of Jilin Province under Grant 20180520029JH, in part by the China Postdoctoral Science Foundation under Grant 2017M611323, in part by the Industrial Technology Research and Development Funds of Jilin Province under Grant 2019C054-3, and in part by the Fundamental Research Funds for the Central Universities, JLU.

ABSTRACT Tremendous strides have been made in face detection thanks to convolutional neural network. However, the performance of previous face detectors deteriorates dramatically as the face scale shrinks. In this paper, we propose a novel scale-invariant face detector, named Small Faces Attention (SFA) face detector, for better detecting small faces. We first present multi-branch face detection architecture which pays more attention to faces with small scale. Then, feature maps of neighbouring branches is fused so that the features coming from large scale can auxiliary detect hard faces with small scale. Finally, we simultaneously adopt multi-scale training and testing to make our model robust towards various scale. Comprehensive experiments show that SFA significantly improves face detection performance, especially on small faces. Our method achieves promising detection performance on challenging face detection benchmarks, including WIDER FACE and FDDB datasets, with competitive runtime speed. Both our code and model will be available at https://github.com/shiluo1990/SFA.

INDEX TERMS Face detection, small face, convolutional neural network, deep learning.

I. INTRODUCTION

Face detection is a fundamental step of many face related applications, such as face alignment [1], [2], face recognition [3], [4], face verification [5], [6] and face expression analysis. Excellent face detectors can exactly classify and locate faces from an image. In recent years, deep learning methods especially convolutional neural networks (CNN) have achieved remarkable successes in a variety of computer vision tasks, ranging from image classification [7], [8] to object detection [9]–[12], which also inspire face detection. Unlike traditional methods of hand-crafted features, CNN-based method can extract face features automatically. Anchor-based face detectors play a dominant role in CNN-based face detectors. They detect faces by classifying and regressing a series of pre-set anchors, which are generated by regularly tiling a collection of boxes with different scale on the images.

Small faces are difficult to be detected due to its small scale. Faces with high detection difficulty are categorized as hard faces. Most of small faces belong to hard faces. However, small scale is just one of those variations making faces hard to be detected. Better tackling hard faces is helpful for detecting small faces.

The associate editor coordinating the review of this manuscript and approving it for publication was Li He.

Despite significant progress, there are still relevant open questions in face detection. Specifically, the performance of anchor-based face detectors drops dramatically as the face scale reduces. To solve this problem, some improvements are applied in our method to better detect small faces. That is our initial motivation.

In this paper, we propose the Small Faces Attention (SFA) face detector to seek out more faces with small scale. We first present multi-branch face detection network to deal with large, medium and small faces respectively. In particular, two branches in SFA focus on small faces. Then, we redesign the anchors, named small faces sensitive anchor design, by adding more anchors to match small faces. Besides, feature map fusion is applied in SFA by combining high-level features into low-level features. We fuse the feature maps of neighboring branches and employ the features coming from large scale to auxiliary detect hard faces with small scale. Note that only two branches for small faces mentioned above use feature map fusion. Finally, we adopt multi-scale training and testing to enhance the performance of face detection. Though previous face detectors are scale-invariant by design, image pyramid can also improve the performance in both training and testing phase.

SFA performs face detection in a single stage via scanning the entire image with a sliding window fashion. It detects faces directly from the early feature maps by classifying a

set of predefined anchors and regressing them at the same time. More importantly, SFA can find faces from images with arbitrary size and the runtime of our method is independent of the number of faces. This is in contrast to proposal-based two stage detectors such as Faster R-CNN [12], whose scale linearly with the number of proposals. Meanwhile, SFA is scale-invariant by design. We simultaneously detect faces with multiple scale from different layers in a single forward pass of the network. For clarity, the main contributions of this paper can be summarized as:

(1) We present multi-branch face detection architecture which pays more attention to small faces.

(2) Feature map fusion is applied by fusing the feature maps of neighbouring branches and employ the features coming from large scale to auxiliary detect hard faces with small scale.

(3) We simultaneously adopt multi-scale training and testing to make our model robust towards various scale.

(4) Our method achieves promising detection performance on challenging face detection benchmarks, including WIDER FACE and FDDB datasets, with competitive runtime speed.

The rest of the paper is organized as follows. Section II briefly reviews the related work in face detection. Section III presents the proposed SFA face detector. Section IV shows our experimental results. Section V concludes this paper.

II. RELATED WORKS

Face detection is a critical and fundamental step to all facial analysis applications, and has been extensively studied over the past few decades. The existing algorithms can be roughly divided into two categories as follows.

Traditional Approaches: The milestone work of Viola and Jones [13] used Haar-like features and AdaBoost to train a cascade of face detectors that achieved a good accuracy. After that, many approaches have been proposed based on the Viola-Jones detectors to advance the state-of-the-art in face detection. LBP [14] and its extension methods introduced local texture features for face detection. These features have been proved to be robust to illumination variation. NPDFace [12] was to address challenges in unconstrained face detection, such as arbitrary pose and heavy occlusion. All of these detectors extract hand-crafted features and optimize each component separately, which makes these traditional face detectors less optimal.

CNN-Based Approaches: In contrast to traditional face detection approaches, CNN-based face detectors greatly improve the detecting performance in recent years. These methods can train on huge and challenging face datasets and automatically extract discriminative features. Furthermore, they can be easily parallelized on GPU cores for acceleration in testing phase. CascadeCNN [16] developed a cascade architecture built on CNNs to detect face coarse to fine. Faceness [17] trained a series of CNNs for facial attribute recognition to detect partially occluded faces. MTCNN [18] proposed to jointly solve face detection and alignment using several multi-task CNNs. FaceHunter [19] proposed a new multi-task CNNs based face detector to discriminate face/non-face and regress face box.

Anchor was first proposed by Faster R-CNN, and then it was widely used in both two stage and single stage object detectors. Later, anchor-based detecting methods were applied in face detection leading to a remarkable progress. SSH [20] introduced a single stage headless face detector and modelled the context information by large filters on each prediction module. S^3FD [21] presented a scale-equitable framework to handle different scales of faces. FaceBoxes [22] introduced anchor densification to ensure different types of anchors have the same density on the image. Face R-CNN [23] employed a new multi-task loss function based on Faster R-CNN framework. CMS-RCNN [24] exploited contextual information to enhance performance. Face R-FCN [25] re-weighted embedding responses on score maps and eliminated the effect of non-uniformed contribution in each facial part.

Despite its great achievement, the main drawback of these frameworks is their poor detection performance for faces with small scale. To address this problem, great efforts have been done in this aspect. HR [26] built multi-level image pyramids to find upscaled small faces. S^3FD [21] proposed anchor matching strategy to improve the recall rate of small faces. Shrivastava *et al.* [27] introduced a novel anchor design to guarantee high overlaps between small faces and anchor boxes.

Although many face detectors are developed, the detection accuracy is still not satisfied, especially for small faces. In this paper, we are interested in developing efficient face detector to better deal with small faces. To this end, SFA face detector is proposed extending from SSH which is an elegant and efficient detection architecture.

III. PROPOSED METHOD

A. GENERAL ARCHITECTURE

The pipeline of face detection using SFA is illustrated in Fig. 1(a). The input image I with arbitrary size is resized to form an image collection $P = \{P_1, P_2, \cdots, P_i, \cdots, P_n\}$ according to scale $S = \{S_1, S_2, \cdots, S_i, \cdots, S_n\}$ in Multi-scale Testing. Each image P_i uses SFA to generate detection result D_i. We merge these detection results to get image D_f as our final detection result of input image I.

Fig. 1(b) shows the network architecture of SFA. First of all, VGG-16 [7] is deployed to extract feature maps from resized image P_i. Then, Feature Map Fusion is applied to fuse feature maps from Conv3_3, Conv4_3, and Conv5_3. Finally, we use a set of multi-branch detection modules to classify face/non-face and regress the bounding boxes. Detection module M_0, M_1, M_2, and M_3 detect faces with small, medium, and large scale respectively. We exploit NMS to generate detection result D_i of image P_i.

B. MULTI-BRANCH DETECTION ARCHITECTURE

CNN-based face detectors exploit convolution and pooling operation to extract discriminative features with different receptive fields (RF). Specifically, the size of RF enlarges

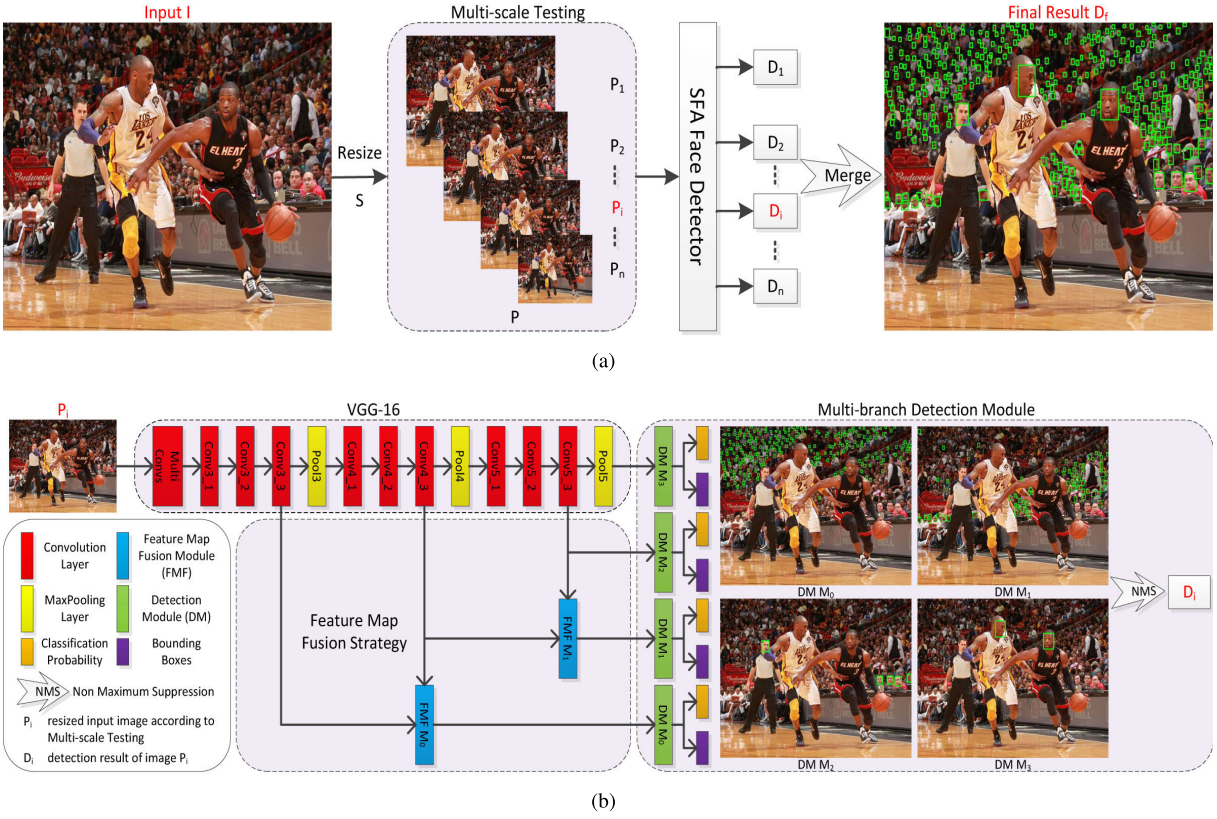


FIGURE 1. (a) The pipeline of face detection using SFA. (b) The network architecture of SFA. It consists of VGG-16, feature map fusion, Multi-branch Detection Module and Multi-scale Testing.

TABLE 1. The detection layer, receptive field and attention scale of each detection module in SFA. DL: detection layer, RF: receptive field, AS: attention scale of faces, DM: detection module.

DM	DL	RF	AS
M_0	Conv3_3	40	small
M_1	Conv4_3	92	small
M_2	Conv5_3	196	medium
M_3	Pool5	212	large

gradually as the feature maps are extracted from low-level to high-level layers as listed in Tab. 1. Thus, the size of RF challenges the scale of faces. Low-level features are lost gradually when CNN-based feature extraction method is applied. In the end, minority feature information is preserved for small faces, which leads to poor performance in detecting small faces. Therefore, it is necessary to detect small faces from early detection layers where still maintain more low-level features.

To this end, we propose a new scale-invariant face detection architecture, named multi-branch detection architecture as shown in Fig. 1(b). Inspired by the divide and conquer strategy, we detect faces from four different layers of VGG-16 using detection modules M_0, M_1, M_2, and M_3. Conv3_3, Conv4_3, Conv5_3, and Pool5 are selected to connect to the detection modules M_0, M_1, M_2, and M_3 separately. These modules have strides of 4, 8, 16, and 32. And they are designed to detect small, medium, and large faces respectively. In particular, two branches of M_0 and M_1 in SFA focus on faces with small scale.

During the training phase, each detection module is trained to detect faces from target scale. To specialize each of the four detection modules for a specific range of scale, we only back-propagate the loss for the anchors which are assigned to faces in the corresponding range. This is implemented by distributing the anchors based on their size to these four modules as discussed in Section III-C. Unlike S^3FD which merges different scale feature maps and forms a comprehensive face features, our work indicates that multi-branch detection modules in scale can be optimally learned separately. In this way, different scale of faces can be automatically divided into different detection modules. This is the divide and conquer strategy to tackle unconstrained face detection in a single detector.

During inference, the predicted boxes from the different branches are joined together followed by Non-Maximum Suppression (NMS) to form the final detection result.

C. SMALL FACES SENSITIVE ANCHOR DESIGN

Anchor-based face detection methods can be regarded as a binary classification problem, which determine if an anchor is face or not. However, few anchors in previous face detectors are offered to match small faces. For example, the size of smallest anchors in SSH [20], S^3FD [21] and Shrivastava *et al.* [27] is 16.

TABLE 2. Small faces sensitive anchor design. DM: detection module, AR: anchor rate, BS: base size, AS: attention scale.

DM	Stride	AR	BS	Anchor	AS
M_0	4	1, 2	4	4, 8	small
M_1	8	4, 8	4	16, 32	small
M_2	16	16, 32	4	64, 128	medium
M_3	32	64, 128	4	256, 512	large

To better detect small faces, we propose small faces sensitive (SFS) anchor design. We tile anchors on a wide range of size varying from 4 to 512 (i.e., 4, 8, 16, 32, 64, 128, 256, 512 in our method), which guarantees that various scale of faces have enough features for detection. More precisely, the smallest anchor in our method is 4 as listed in Tab. 2. And the anchors of 4, 8, 16, and 32 are applied for faces with small scale. Benefit from the multi-branch detection architecture as discussed in Section III-B, SFA reasonably arranges small faces sensitive anchors into these detection modules and forms our SFS anchor design, which improves the robustness to face scale.

For implementation, we use anchor ratio (AR) and base size (BS) to form anchor design. AR multiple BS is the size of anchor. The AR of $\{1, 2\}$ in M_0, $\{4, 8\}$ in M_1, $\{16, 32\}$ in M_2, and $\{64, 128\}$ in M_3 is denoted as 4-branch AR. As listed in Tab. 2, we form the SFS anchor design using 4-branch AR with the BS of 4. Thus, plenty of small anchors are densely tiled on the image. However, these small anchors inevitably lead to a sharp increase in the number of negative anchors on the background. Thanks to OHEM [28], SFA can balance the positive and negative anchors with a ratio of 1:3 in each mini-batch. Mining hard samples in training is critical to strengthen the power of detector.

D. FEATURE MAP FUSION

Small faces are difficult to be detected not only because of their small scale. Atypical pose, heavy occlusion, extreme illumination, low resolution and other variations in unconstrained scenarios always make CNN-based feature extraction hard to obtain sufficient and complete features for detecting small faces. Therefore, most of small faces become hard faces.

To further improve the ability of detecting hard faces with small scale, we use the Feature Map Fusion (FMF) strategy. FMF is applied in SFA by combining high-level features into low-level features. We fuse the feature maps of neighboring branches and apply the features coming from large scale to auxiliary detect small faces according to a bold guess that faces with neighboring scale have similar features. We use the FMF strategy in branch M_0 and M_1 as seen in Fig. 1(b), which receive the early extracted feature maps from Conv3_3 and Conv4_3. Fig. 2 shows the architecture of FMF module. More precisely, feature maps F_{i+1} are upsampled and summed up with feature maps F_i where $i \in \{0, 1\}$, followed by a 3×3 convolutional layer. We used bilinear upsampling in the fusion process.

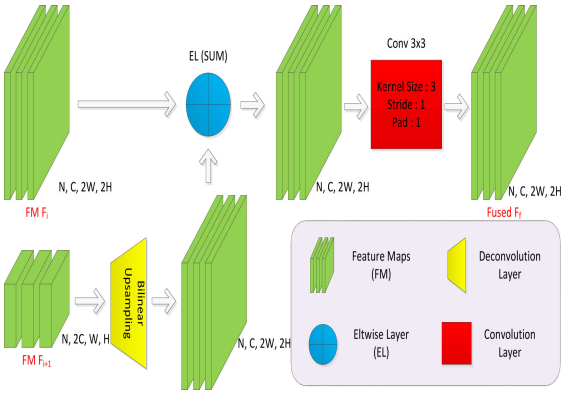


FIGURE 2. The architecture of feature map fusion module.

By using FMF strategy, SFA is robust to different kinds of variations for small faces to some extent, including occlusion, illumination, low resolution, blur, etc. Benefit from the feature maps coming from neighboring branch with large scale, SFA can also detect small faces well even though the feature maps in current branch are insufficient and incomplete due to different kinds of variations. From the results of ablation study in Section IV-C3, we can see that FMF strategy significantly improves the detection performance on the hard set of WIDER FACE [29] dataset which includes a lot of small faces.

In fact, medium scale faces can achieve sufficient and complete features extracted by CNN-based face detector. Therefore, there is no need to fuse the feature maps between medium and large faces. Ablation study in Section IV-C3 also shows that FMF strategy is not fit for medium faces.

E. MULTI-SCALE TRAINING AND TESTING

Instead of using a fixed scale in both training and testing phase, we perform Multi-scale Training (MS-Training) and Multi-scale testing (MS-Testing) strategy to learn more features across a wide range of scale, which makes our model more robust towards different scale and significantly improves the detection performance.

In the training phase, we first resize the shortest side of the input image I up to S_i ($S_i \in S$) while keeping the largest side below Max Size (1600 in our method). Then, we scale the image according to S in MS-Training. For example, when the scale S of MS-Training is set to 500, 800, 1200, and 1500, denoted as 4-scale, the input image I is first resized to 1200×1600, then we scale the resized image with the size of 500, 800, 1200, and 1500 in the pyramid. In the testing phase, MS-Testing is performed accordingly. We build an image pyramid with a wider range of scale for each test image. Limited to the capacity of GPU memory, the scale of 500, 600, 700, 800, 900, 1000, 1100, 1200, and 1600, denoted as wide-scale, are applied in multi-scale testing phase. Each scale in the pyramid is independently tested. The detection results from various scale are eventually merged together as the final result D_f of the input image I as shown in Fig. 1(a).

MS-Training makes parameters of four detection modules (detection module M_0, M_1, M_2, and M_3) in SFA robust to detect faces with various scale as illustrated in Tab. 2. Different detection modules focus on its own attention scale of faces. As MS-Testing is used in the testing phase, each face of input image I will be rescaled accordingly. These rescaled faces may be detected by SFA from different detection modules whose attention scale match with the size of rescaled faces. When at least one rescaled face is found by certain detection module, the original face in input image I is successfully detected.

Benefit from MS-Training and MS-Testing, SFA enlarges small faces and easily detect them in medium and large anchors. Fig. 3 shows an example of using MS-Testing. The table in Fig. 3 lists different detection result D_i of rescaled P_i. These detection results are merged to generate the left image as its final detection result D_f. We denote $f = \{f_1, f_2, f_3, f_4, f_5, f_6\}$ as the face collection of final detection result D_f. Faces f_3 and f_4 are small faces while they can be detected from rescaled image P_4 by using detection module M_2 whose anchors attention faces with medium scale. At the same time, SFA shrinks large faces and better detects them in small and medium anchors as well as rescales and finds medium faces with the help of small and large anchors to some extent. As seen in Fig. 3, face f_5 is medium face but they can be detected from rescaled image P_4 by using detection module M_3 whose anchors attention faces with large scale.

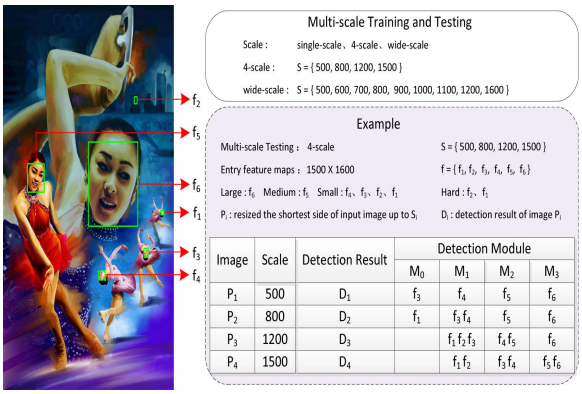


FIGURE 3. An example of using multi-scale testing.

Though SFA is scale-invariant by design, image pyramid can also improve the performance in both training and testing phase. Ablation study in Section IV-C3 shows that MS-Training can enhance the detection performance on all subsets, especially on the hard set. Surprisingly, the runtime of SFA will not increase if we adopt MS-Training. Hence, we denoted it as our real-time SFA face detector which adopts MS-Training strategy only. Besides, MS-Testing can improve the detection performance on all subsets by a large margin. Therefore, we deploy both MS-Training and MS-Testing strategy in our final SFA face detector model.

F. LOSS FUNCTION

During the training phase, SFA uses a multi-task loss function [9], [12], [20], [21]. This loss function Eq. (1) can be formulated as follows:

$$L(\{p_i\}, \{t_i\}) = \sum_k \frac{1}{N_k^{cls}} \sum_{i \epsilon A_k} L_{cls}(p_i, p_i^*) + \lambda \sum_k \frac{1}{N_k^{reg}} \sum_{i \epsilon A_k} p_i^* L_{reg}(b_i, b_i^*) \quad (1)$$

where index k goes over the detection modules $\{M_k\}_1^K$ (e.g., $K = 4$ in SFA with 4-branch detection modules) and i is the index of an anchor in detection module M_k. A_k represents the set of anchors defined in detection module M_k. The classification loss L_{cls} is softmax loss over two classes (face vs. background). p_i is the predicted probability that anchor i is a face. The ground-truth label p_i^* is 1 if the anchor is positive, and 0 if the anchor is negative. N_k^{cls} is the number of anchors in detection module M_k which participate in the classification loss computation. The regression loss L_{reg} is the smooth L_1 loss defined in [9]. It can be formulated as seen in Eq. (2) and Eq. (3):

$$L_{reg}(x, y) = smooth_{L_1}(x - y) \quad (2)$$

$$smooth_{L_1}(x) = \begin{cases} 0.5x^2 & if\ |x| < 1 \\ |x| - 0.5 & otherwise \end{cases} \quad (3)$$

b_i is a vector representing the 4 parameterized coordinates of the predicted bounding box, and b_i^* is that of the ground-truth box associated with a positive anchor. $p_i^* L_{reg}$ means the regression loss is activated only for positive anchors and disabled otherwise, and $N_k^{reg} = \sum_{i \epsilon A_k} p_i^*$. Besides, λ is used to balance these two loss terms.

IV. EXPERIMENTS

In this section, we firstly analyze the effectiveness of our proposed strategies with comprehensive ablative experiments. Then, we evaluate the final optimal model and achieve promising results on common face detection benchmarks. The inference time is finally presented.

A. EXPERIMENTAL SETUP

The parameters of SFA network are initialized from a pre-trained ImageNet classification model. Our method fine-tunes the resulting model using stochastic gradient descent (SGD) with 0.9 momentum and 0.0005 weight decay. The maximum number of iterations is 54k and stepsize is 18k. The learning rate is firstly set to 0.004 and decreases by a factor of 0.1. Anchors with IoU greater than 0.45 are assigned to positive class and anchors which have IoU less than 0.35 with all ground-truth faces are assigned to the negative class while the rest are ignored. For anchor generation, we use AR of {1, 2} in M_0, {4, 8} in M_1, {16, 32} in M_2, and {64, 128} in M_3 with a BS of 4. All anchors have aspect ratio of one. Each training image uses horizontal flipping with probability of 0.5 as our data augmentation strategy. We employ the multi-task loss as our objective function. Besides, online negative and positive mining (OHEM) [28] is applied to balance the positive and negative training examples with a ratio of 1:3. During training, 256 detections per module

TABLE 3. The results of comprehensive ablation studies.

Experiment	Description	AP			Contribution
		Easy	Medium	Hard	
Baseline	3-branch	0.926	0.914	0.819	Multi-branch
I	4-branch	0.927	0.915	0.822	Architecture
II	(I)+4-branch AR with BS = 16	0.902	0.852	0.391	
III	(I)+4-branch AR with BS = 8	0.922	0.901	0.798	SFS
IV	(I)+4-branch AR with BS = 4	0.925	0.914	0.821	
V	(IV)+FMF M_0	0.927	0.914	0.819	
VI	(IV)+FMF M_1	0.925	0.915	0.821	FMF
VII	(IV)+FMF M_0M_1	0.927	0.915	0.830	
VIII	(IV)+FMF $M_0M_1M_2$	0.928	0.914	0.812	
IX	(VII)+MS-Training(4-scale)	0.930	0.916	0.839	
X	(Baseline)+MS-Testing(4-scale)	0.940	0.928	0.849	MS-Training
XI	(VII)+MS-Testing(4-scale)	0.943	0.931	0.855	and
XII	(VII)+MS-Training(4-scale)+MS-Testing(4-scale)	0.946	0.934	0.862	MS-Testing
XIII	(VII)+MS-Training(4-scale)+MS-Testing(wide-scale)	**0.949**	**0.936**	**0.866**	

are selected for each image. During inference, each module outputs 1000 best scoring anchors as detections and NMS with a threshold of 0.3 is performed on the outputs of all modules together. Our method is implemented in Caffe [30] and all the experiments are trained on 2 NVIDIA GeForce GTX 1080Ti GPUs in parallel.

B. DATASETS

WIDER FACE dataset [29]: This dataset contains 32,203 images with 393,703 labeled faces with a high degree of variability in scale, pose and occlusion. It is organized based on 61 event classes, which have much more diversities and are closer to the real-world scenarios. The images in this dataset are split into training (40% and 12880 images), validation (10% and 3226 images), and testing (50% and 16097 images) set. Thus, 158989 labeled faces are in the training set, while 39496 in the validation set and the rest in the testing set. Faces in this dataset are classified into Easy, Medium, and Hard subsets according to the difficulties of detection. The hard subset includes a lot of small faces. Average precision (AP) score is used as the evaluation metric. Plotting scripts for generating the precision-recall (PR) curves are provided to evaluate the performance on the validation set online. While evaluating on the testing set, the results are needed to be sent to the dataset server for receiving the PR curves. We train all models on the training set of the WIDER FACE dataset and evaluate on its validation and test sets. Ablation studies are also performed on the validation set.

FDDB dataset [31]: It contains the annotations for 5171 faces in a set of 2845 images taken from news articles on Yahoo websites. Most of the images in the FDDB dataset have less than 3 faces that are clear or slightly occluded. The faces generally have large sizes and high resolutions compared to WIDER FACE. Instead of rectangle bounding boxes, faces in FDDB are represented by bounding ellipses. We use the same model of Experiment XIII presented in Section IV-C which trained on WIDER FACE training set to perform the evaluation on the FDDB dataset.

C. ABLATION STUDY

We conduct ablation experiments to examine how each of these proposed strategies affects the final performance. The detailed experimental results of the ablation studies are listed in Tab. 3.

1) BASELINE SETUP

Our baseline detector consists of 3-branch detection architecture (branch M_1, M_2, and M_3) and 3-branch AR ($\{1, 2\}$ in M_1, $\{4, 8\}$ in M_2, and $\{16, 32\}$ in M_3) with a BS of 16 into three detection modules as listed in Tab. 3.

2) ABLATION SETTING

First of all, to better understand the impact of multi-branch detection architecture, we add a new branch M_0 on the baseline to form 4-branch detection architecture and denote it as Experiment I. To be fair, detection module M_0 use the same AR as detection module M_1 does (e.g., $\{1, 2\}$ in both M_0 and M_1). All other factors are the same.

Second, we evaluate the effect of SFS anchor design. For anchor generation, we use 4-branch AR (e.g., $\{1, 2\}$ in M_0, $\{4, 8\}$ in M_1, $\{16, 32\}$ in M_2, and $\{64, 128\}$ in M_3) but with different BS of 16, 8, and 4 in Experiment II, III, and IV separately. All of these experiments are based on 4-branch detection architecture like Experiment I. Other parameters remain the same.

Third, by further examining the impact of FMF strategy, we add the FMF module M_0, M_1, M_0M_1, and $M_0M_1M_2$ in experiment V, VI, VII, and VIII respectively. All of these experiments are based on the detection architecture of Experiment IV.

Fourth, we evaluate the influence of MS-Training and MS-Testing. At first, we exploit MS-Training in Experiment IX, which is based on Experiment VII. Similar to SSH, 4-scale (e.g., 500, 800, 1200, and 1500) is used in MS-Training. Next, we apply 4-scale MS-Testing in Experiment X and XI based on Baseline and Experiment VII. Then, both MS-Training and MS-Testing are deployed in Experiment XII, also based on experiment VII, with the same 4-scale mentioned above. Finally, compared to Experiment XII, a wider range of scale is used in Experiment XIII for MS-Testing. Limited to the capacity of GPU memory, wide-scale (e.g., 500, 600, 700, 800, 900, 1000, 1100, 1200, and 1600) is selected.

3) ABLATION RESULTS

a: MULTI-BRANCH DETECTION ARCHITECTURE IS BETTER

Compared to the 3-branch baseline in Tab. 3, 4-branch detection architecture in Experiment I slightly improves the detection performance on the hard set (rising by 0.3%). The result of Baseline and Experiment I show that 4-branch detection architecture is better for improving the detection performance, especially on the hard set. Therefore, the following ablation studies will adopt the 4-branch detection architecture.

b: SMALL FACES SENSITIVE ANCHOR DESIGN IS CRUCIAL FOR DETECTING SMALL FACES

The comparison among the result of Experiment II, III, and IV in Tab. 3 indicates that the detection performance gradually improves on the easy, medium, and hard set as the decrease of BS in anchor design. Besides, compared to Experiment I, the result in Experiment IV is slightly lower on all validation sets. Though the same AR of $\{1, 2\}$ in detection module M_0, BS is 16 in Experiment I but 4 in Experiment IV, leading to different anchor sizes. Smaller anchors make it possible to find some more small faces at the cost of the rising of false positive. Thanks to the FMF strategy mentioned in Section III-D above, we can decrease the rate of false positive in the following ablation studies. In order to achieve an elegant anchor design, we will adopt SFS anchor design like Experiment IV in the following ablation studies.

c: FEATURE MAP FUSION STRATEGY IS PROMISING FOR DETECTING HARD FACES

From the results of Experiment VII in Tab. 3, we can see that the detection performance has a great improvement, especially on hard set (about 0.9% compared to Experiment IV), by using FMF module M_0M_1 simultaneously. Surprisingly, the detector with FMF is robust to different kinds of variations to some extent, including occlusion, illumination, blur, etc. When FMF module $M_0M_1M_2$ are used in Experiment VIII, the detection performance sharply drops on the hard set. Compared to Experiment IV without feature maps fusion, the detection performance in Experiment VIII is worse on the hard set (about 0.9%). Therefore, we will use FMF module M_0M_1 in the following ablation studies.

d: MULTI-SCALE TRAINING AND TESTING CAN SIGNIFICANTLY IMPROVE THE DETECTING PERFORMANCE

The result of Experiment IX shows that MS-Training is helpful for enhancing the detection performance, especially on the hard set. We denoted it as our real-time SFA face detector which adopts MS-Training strategy. Benefit from MS-Testing, the detection performance of Experiment X and XI have a great improvement on all validation sets compared to Baseline and Experiment VII. Later, Experiment XII adopts both MS-Training and MS-Testing with the same 4-scale simultaneously and further improves the detection performance. Finally, wide-scale is used in Experiment XIII for MS-Testing. Compared to Experiment VII, the result of Experiment XIII increases 2.2%, 2.1%, and 3.6% on the easy, medium, and hard set separately, which demonstrates that MS-Training and MS-Testing can significantly improve the detecting performance.

Combining all the above strategies achieve the best detection performance (as shown in Experiment XIII) and denote it as our final SFA detector model.

D. EVALUATION ON BENCHMARK

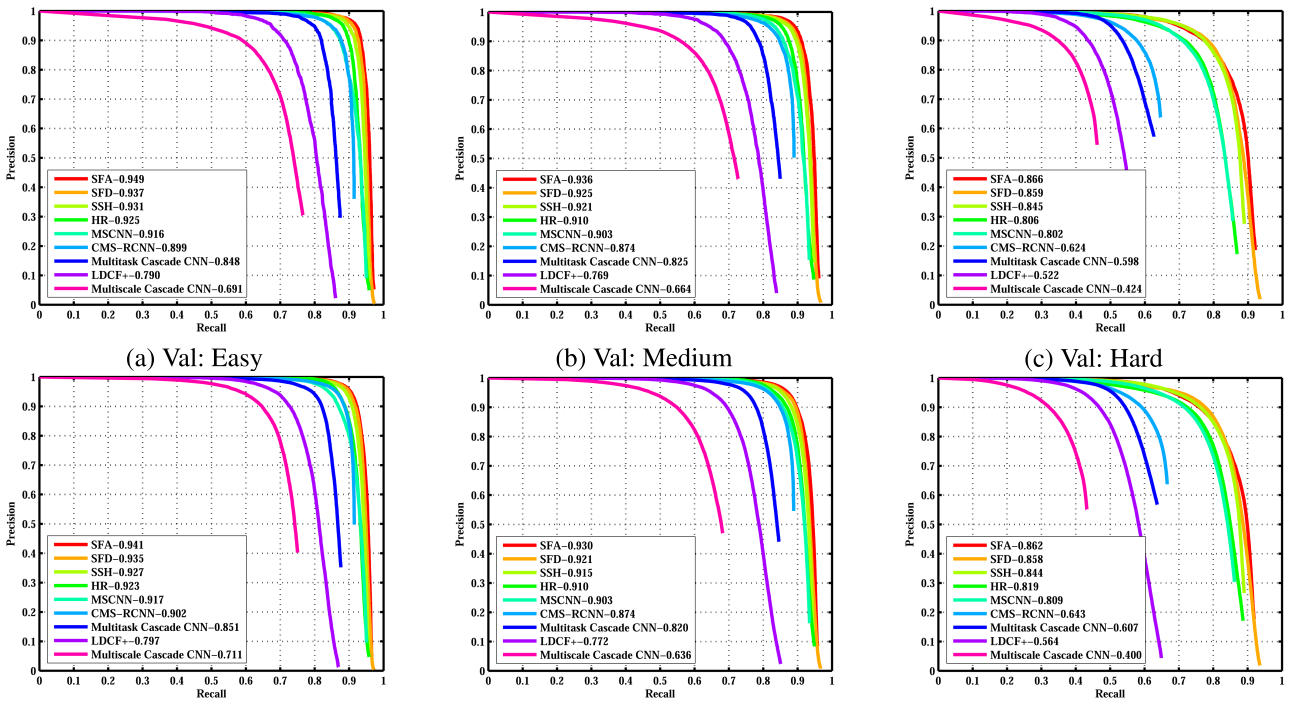
We evaluate our proposed method against state-of-the-art methods on two public face detection benchmarks (i.e. WIDER FACE [29] and FDDB [31]).

1) WIDER FACE DATASET

Our method is trained on the training set of the WIDER FACE dataset and evaluate on its validation and testing set against the recently published state-of-the-art face detection methods including S^3FD [21], SSH [20], HR [26], MSCNN [32], CMS-RCNN [24], Multitask Cascade CNN [18], LDCF+ [33] and Multiscale Cascade CNN [29]. The precision-recall curves and AP values on WIDER FACE validation and testing sets are presented in Fig. 4. As can be seen, the proposed SFA approach consistently achieves the impressive performance across all the three subsets, especially on the hard subset which mainly contains small faces. It achieves the promising average precision in all level faces, i.e. 0.949 (Easy), 0.936 (Medium), and 0.866 (Hard) for validation set, and 0.941 (Easy), 0.930 (Medium), and 0.862 (Hard) for testing set. The result in Fig. 4 not only demonstrates the effectiveness of the proposed method but also strongly shows the superiority of the proposed model in detecting small and hard faces.

2) FDDB DATASET

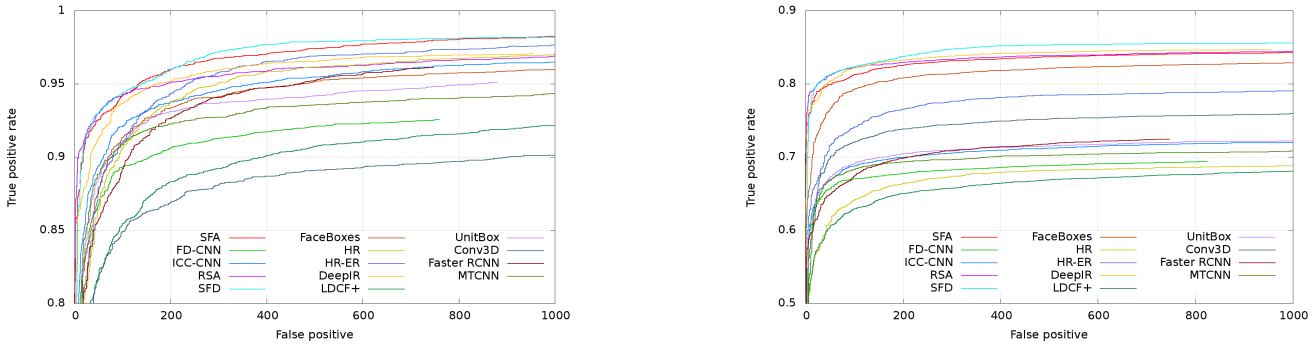
In these datasets, we resize the shortest side of the input images to 400 pixels while keeping the larger side less than 800 pixels, leading to an inference speed of more than 20 FPS. And we directly use our final SFA detector model in Experiment XIII and compare SFA against the recently published state-of-the-art methods including FD-CNN [34], ICC-CNN [35], RSA [36], S^3FD [21], FaceBoxes [22], HR [26], HR-ER [26], DeepIR [37], LDCF+ [33],



(a) Val: Easy (b) Val: Medium (c) Val: Hard

(d) Test: Easy (e) Test: Medium (f) Test: Hard

FIGURE 4. Precision-recall curves on WIDER FACE validation and test sets.



(a) Discontinuous ROC curves (b) Continuous ROC curves

FIGURE 5. Evaluation on the FDDB dataset.

UnitBox [38], Conv3D [39], Faster RCNN [40] and MTCNN [18] on FDDB dataset. For a more fair comparison, the predicted bounding boxes are converted to bounding ellipses. Fig. 5 show the discrete ROC curves and continuous ROC curves of these methods on the FDDB dataset respectively. The proposed SFA approach consistently achieves the impressive performance in terms of both the discrete ROC curves and continuous ROC curves. These results demonstrate the effectiveness and good generalization capability of SFA to detect unconstrained faces.

E. INFERENCE TIME

In this section, we report the inference time of our proposed SFA face detector on the WIDER FACE validation set. Benefit from the single stage of SFA, our inference time is independent of the number of faces in an image. It can detect faces from images with arbitrary size. Specifically, the inference time of our proposed method is determined by two aspects as follows: (1) The size of entry feature map ($W\times H$); (2) The number of scale N_s in MS-testing. A feature map is a tensor of size $C\times W\times H$, where C is the number of channels, W and H are the width and the height of feature map respectively. When the max scale S_{max} in MS-Testing and Max Size are set, the size of entry feature map in SFA is determined at the same time. More precisely, W is set to S_{max} and H is set to Max Size.

The speed is measured by using NVIDIA GeForce GTX 1080Ti GPU and cuDNN v5.1 with Intel Core i7-6850k CPU@3.60GHz. Tab. 4 shows the inference time with respect to the size of entry feature map ($W\times H$) and the number of scales N_s in MS-Testing. For the 1200×1600 entry feature map, our real-time SFA face detector can run at 5 FPS as well as maintain high performance, as the row 5 listed in Tab. 4. Our final SFA detector as the last row of Tab. 4 described can take 1.4s to detect faces from an image with 1500×1600 entry feature map. From Experiment XII in

(a)

(b)

FIGURE 6. Qualitative results of SFA on the WIDER FACE validation set. Red bounding boxes are the faces that annotated on the WIDER FACE validation dataset. Green bounding boxes represent the detection results. Best viewed in color. Please zoom in to see some small detections.

TABLE 4. SFA inference time with respect to the size of entry feature map (W×H) and the number of scales N_s in MS-Testing.

ID	W×H	MS-Testing	S_{max}	Max Size	N_s	Time (ms)
1	400×800	400	400	800	1	49
2	480×640	480	480	640	1	49
3	600×1000	600	600	1000	1	74
4	720×1280	720	720	1280	1	100
5	1200×1600	1200	1200	1600	1	200
6	1500×1600	4-scale	1500	1600	4	750
7	1500×1600	wide-scale	1600	1600	9	1400

Tab. 3 and the last second row in Tab. 4, we can see that when using MS-Training and MS-Testing with the same 4-scale, our method can take 0.75s to detect faces from an image with 1500×1600 entry feature map. It achieves slightly lower detection performance against our final SFA detector but reduces half of inference time. Therefore, our SFA face detector is fit for real applications by simultaneously using MS-Training and MS-Testing with 4-scale.

F. QUALITATIVE RESULTS

Fig. 6 shows some examples of the face detection results using the proposed SFA on the WIDER FACE validation dataset. Fig. 6(a) lists some difficult cases. Our method is able to detect faces with different scales, especially for small faces (see the first row in Fig. 6(a)). Besides, SFA can

FIGURE 7. Qualitative results of SFA on the FDDB dataset. Red bounding ellipses are the faces that FDDB labeled; Green bounding boxes are the detection results. Best viewed in color. Please zoom in to see some small detections.

also achieve satisfied detection results on hard faces caused by atypical pose, heavy occlusion, exaggerated expression, make up, extreme illumination and blur (see the last two rows in Fig. 6(a)). Fig. 6(b) lists some selected false positives. In fact, most of the false positives in SFA are actually human faces caused by missing labels (see the first two rows in Fig. 6(b)). For other false positives, we find errors made by our model are rather reasonable. They all have the pattern of human face and fool our model to treat it as a face (see the last row in Fig. 6(b)).

Fig. 7 shows some examples of the face detection results generated by SFA on the FDDB dataset. Fig. 7(a) lists some difficult cases including faces with different scale, atypical pose, heavy occlusion, exaggerated expression, and blur. Benefit from excellent performance of SFA in detecting small faces and hard faces, we can find a lot of faces from human perspective but lack of labels on the FDDB dataset, as seen in Fig. 7(b). Our method is able to find extra faces with small scale which are not labeled (see the first row in Fig. 7(b)). Besides, some faces with atypical pose can also be detected

(see the second row in Fig. 7(b)). The detection results of faces with heavy occlusion, blur, and wrong label are shown in the last row of Fig. 7(b).

V. CONCLUSION

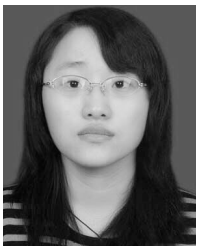
In this paper, we propose a novel face detector, named Small Faces Attention (SFA) face detector, to deal with the open problem of anchor-based detection methods whose performance drops sharply as the faces becoming small. Multiple strategies are deployed in SFA for the sake of better detecting small faces, such as multi-branch detection architecture, small faces sensitive anchors design, feature map fusion strategy, multi-scale training, and multi-scale testing strategy. These strategies make SFA rapid, efficient, and robust to detect faces in unconstrained settings, especially for small faces. Extensive experiments demonstrate that our method outperforms most of the recently published face detectors and achieves promising performance on challenging face detection benchmarks like WIDER FACE and FDDB datasets with competitive inference speed.

REFERENCES

[1] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Proc. CVPR*, Jun. 2013, pp. 532–539.
[2] A. Jourabloo, M. Ye, X. Liu, and L. Ren, "Pose-invariant face alignment with a single CNN," in *Proc. ICCV*, Oct. 2017, pp. 3219–3228.
[3] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. CVPR*, Jun. 2015, pp. 815–823.
[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.*, Oct. 2016, pp. 499–515.
[5] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Proc. CVPR*, Jun. 2014, pp. 1891–1898.
[6] Y. Taigman, M. Yang, M. A. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *Proc. CVPR*, Jun. 2014, pp. 1701–1708.
[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556
[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, Jun. 2016, pp. 770–778.
[9] R. Girshick, "Fast R-CNN," in *Proc. ICCV*, Dec. 2015, pp. 1440–1448.
[10] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 21–37.
[11] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. CVPR*, Jun. 2016, pp. 779–788.
[12] 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.
[13] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.
[14] H. Jin, Q. Liu, H. Lu, and X. Tong, "Face detection using improved LBP under Bayesian framework," in *Proc. 3rd Int. Conf. Image Graph.*, Dec. 2015, pp. 306–309.
[15] S. Liao, A. K. Jain, and S. Z. Li, "A fast and accurate unconstrained face detector," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 211–223, Feb. 2016.
[16] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, "A convolutional neural network cascade for face detection," in *Proc. CVPR*, Jun. 2015, pp. 5325–5334.
[17] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "From facial parts responses to face detection: A deep learning approach," in *Proc. ICCV*, Dec. 2017, pp. 3676–3684.
[18] 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.
[19] D. Wang, J. Yang, J. Deng, and Q. Liu, "*FaceHunter*: A multi-task convolutional neural network based face detector," *Signal Process., Image Commun.*, vol. 47, pp. 476–481, Sep. 2016.
[20] M. Najibi, P. Samangouei, R. Chellappa, and L. S. Davis, "SSH: Single stage headless face detector," in *Proc. ICCV*, Oct. 2017, pp. 4875–4884.
[21] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, "S3FD: Single shot scale-invariant face detector," in *Proc. ICCV*, Oct. 2017, pp. 192–201.
[22] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, "FaceBoxes: A CPU real-time face detector with high accuracy," in *Proc. IEEE Int. Joint Conf. Biometrics*, Oct. 2017, pp. 1–9.
[23] H. Wang, Z. Li, X. Ji, and Y. Wang, "Face R-CNN," 2017, *arXiv:1706.01061*. [Online]. Available: https://arxiv.org/abs/1706.01061
[24] C. Zhu, Y. Zheng, K. Luu, and M. Savvides, "CMS-RCNN: Contextual multi-scale region-based CNN for unconstrained face detection," in *Deep Learning for Biometrics*, B. Bhanu and A. Kumar, Eds. Cham, Switzerland: Springer, 2017, pp. 57–79, doi: 10.1007/978-3-319-61657-5_3.
[25] Y. Wang, X. Ji, Z. Zhou, H. Wang, and Z. Li, "Detecting faces using region-based fully convolutional networks," 2017, *arXiv:1709.05256*. [Online]. Available: https://arxiv.org/abs/1709.05256
[26] P. Hu and D. Ramanan, "Finding tiny faces," in *Proc. CVPR*, Jul. 2017, pp. 1522–1530.
[27] C. Zhu, R. Tao, K. Luu, and M. Savvides, "Seeing small faces from robust anchor's perspective," in *Proc. CVPR*, Jun. 2018, pp. 5127–5136.
[28] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proc. CVPR*, Jun. 2016, pp. 761–769.
[29] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "WIDER FACE: A face detection benchmark," in *Proc. CVPR*, Jun. 2016, pp. 5525–5533.
[30] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. Int. Conf. Multimedia*, 2014, pp. 675–678.
[31] V. Jain and E. Learned-Miller, "FDDB: A benchmark for face detection in unconstrained settings," Univ. Massachusetts Amherst, Amherst, MA, USA, Tech. Rep. UM-CS-2010-009, 2010.
[32] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 354–370.
[33] E. Ohn-Bar and M. M. Trivedi, "To boost or not to boost? On the limits of boosted trees for object detection," in *Proc. ICPR*, pp. 3350–3355, Dec. 2016.
[34] D. Triantafyllidou, P. Nousi, and A. Tefas, "Fast deep convolutional face detection in the wild exploiting hard sample mining," *Big Data Res.*, vol. 11, pp. 65–76, Mar. 2018.
[35] K. Zhang, Z. Zhang, H. Wang, Z. Li, Y. Qiao, and W. Liu, "Detecting faces using inside cascaded contextual CNN," in *Proc. ICCV*, Oct. 2017, pp. 3190–3198.
[36] Y. Liu, H. Li, J. Yan, F. Wei, X. Wang, and X. Tang, "Recurrent scale approximation for object detection in CNN," in *Proc. ICCV*, Oct. 2017, pp. 571–579.
[37] X. Sun, P. Wu, and S. C. H. Hoi, "Face detection using deep learning: An improved faster RCNN approach," *Neurocomputing*, vol. 299, pp. 42–50, Jul. 2018.
[38] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. Huang, "UnitBox: An advanced object detection network," in *Proc. Int. Conf. Multimedia*, 2016, pp. 516–520.
[39] Y. Li, B. Sun, T. Wu, and Y. Wang, "Face detection with end-to-end integration of a ConvNet and a 3D model," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 420–436.
[40] H. Jiang and E. Learned-Miller, "Face detection with the faster R-CNN," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, Jun. 2017, pp. 650–657.

SHI LUO received the B.S. degree in software engineering from Jilin University, in 2016, where he is currently pursuing the Ph.D. degree with the College of Computer Science and Technology. His research interests include computer vision, pattern recognition, especially for object detection, and face detection.

XIONGFEI LI received the B.S. degree in computer software from Nanjing University, in 1985, the M.Sc. degree in computer software from the Chinese Academy of Sciences, in 1988, and the Ph.D. degree in communication and information system from Jilin University, Changchun, China, in 2002. Since 1988, he has been a member of the faculty of the Computer Science and Technology, Jilin University, where he is currently a Professor of computer software and theory. He has authored more than 100 research artiles. His research interests include datamining, intelligent networks, and image processing and analysis.



RUI ZHU received the B.S. degree in software engineering from Jilin University, in 2016, and the master's degree. She is currently pursuing the Ph.D. degree with Jilin University. Her research specializes in image processing and deep learning.

XIAOLI ZHANG received the M.Sc. degree in computer science and technology from Jilin University, in 2012, where he is currently pursuing the Ph.D. degree with the College of Computer Science and Technology. He has published more than 20 articles in journals and conferences. His research interests include information fusion, algorithm evaluation, and data mining.