Face Detection With Different Scales Based on Faster R-CNN

Wenqi Wu, Yingjie Yin, Xingang Wang, and De Xu, *Senior Member, IEEE*

***Abstract*—In recent years, the application of deep learning based on deep convolutional neural networks has gained great success in face detection. However, one of the remaining open challenges is the detection of small-scaled faces. The depth of the convolutional network can cause the projected feature map for small faces to be quickly shrunk, and most detection approaches with scale invariant can hardly handle less than 15×15 pixel faces. To solve this problem, we propose a different scales face detector (DSFD) based on Faster R-CNN. The new network can improve the precision of face detection while performing as real-time a Faster R-CNN. First, an efficient multitask region proposal network (RPN), combined with boosting face detection, is developed to obtain the human face ROI. Setting the ROI as a constraint, an anchor is inhomogeneously produced on the top feature map by the multitask RPN. A human face proposal is extracted through the anchor combined with facial landmarks. Then, a parallel-type Fast R-CNN network is proposed based on the proposal scale. According to the different percentages they cover on the images, the proposals are assigned to three corresponding Fast R-CNN networks. The three networks are separated through the proposal scales and differ from each other in the weight of feature map concatenation. A variety of strategies is introduced in our face detection network, including multitask learning, feature pyramid, and feature concatenation. Compared to state-of-the-art face detection methods such as UnitBox, HyperFace, FastCNN, the proposed DSFD method achieves promising performance on popular benchmarks including FDDB, AFW, PASCAL faces, and WIDER FACE.**

***Index Terms*—Deep convolutional neural network (DCNN), deep learning, face detection, Faster R-CNN.**

I. Introduction

FACE detection has been being a research hotspot in the field of computer vision. The face detection technology regarding nearly frontal faces has been well-developed. Viola and Jones [1], [2] proposed a robust real-time face detection method which has been widely used. In their method, rectangular Haar-like features are applied in an AdaBoost cascaded classifier to achieve real-time face detection, which greatly increases the precision of face detection, while nonfrontal faces can hardly be detected. To solve such problems, new features which are more complicated and more robust against the nonfrontal face detection have been chosen, such as SURF [3], HOG [4], and ACF [5]. Furthermore, Hayashi and Hasegawa [54] and Zheng *et al.* [56] proposed a low-resolution face detector based on the AdaBoost cascaded classifier to improve the detection rate of small face. Other improvements include human face detection through support vector machines [6] and Random Forest [7] ever since researchers started to lay their emphases on the uncontrolled face detection under a complicated background. Factors that affect the precision of face detection include pose, illumination, expression, and occlusion. Detection algorithms based on the deformable part models (DPM) dominate face detection and provide favorable detection performance of human faces in different poses and angles. The unified model [8], structural model [9], and vanilla DPM [10] are representative DPM-based face detection methods which improved the detection accuracy effectively. Zhang *et al.* [58] proposed a boosted method for detection, they divide the detector learning process into two stages, and formulate it as a weak-to-strong learning framework. Zhang *et al.* [59] proposed an adaptive patch-of interest composition approach for boosting both accuracy and speed for detection while maintaining low complexity. In recent years, methods based on deep convolutional neural networks (DCNN) [11]–[17] have been a significant success in the field of human face detection. Compared to the hand-crafted feature in the past, DCNN is more robust in automatically extracting features of human faces through large amounts of training data. Furthermore, more complicated features are utilized by He *et al.* [53] and Kong *et al.* [57] to improve the expression ability of features. Besides, human face detectors based on general object detection framework RCNN [19] achieve significant precision, such as the human face detection framework proposed by Chen *et al.* [18], which is based on the region proposal network (RPN) [20] and RCNN [19]. Le *et al.* [63] proposed a multiscale Faster-RCNN method, which fuses multilayer feature maps to achieve multiscale object detection. The network parameters are difficult to adapt to a wide range of object scales since only one set of network parameters is used to extract features for different scale objects. On the other hand, the network depth of DCNN

Manuscript received December 11, 2017; revised March 5, 2018 and April 23, 2018; accepted July 20, 2018. Date of publication August 14, 2018; date of current version July 19, 2019. This work was supported in part by the National Natural Science Foundation of China under Grant 61421004, Grant 61573349, and Grant 61703398, and in part by the National High Technology Research and Development Program of China (863 Program) under Grant 2015AA042308. This paper was recommended by Associate Editor H. Lu. *(Corresponding author: Yingjie Yin.)*

The authors are with the Research Center of Precision Sensing and Control, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: wuwenqi2013@ia.ac.cn; yingjie.yin@ia.ac.cn; xingang.wang@ia.ac.cn; de.xu@ia.ac.cn).

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

Digital Object Identifier 10.1109/TCYB.2018.2859482

makes it impossible to obtain satisfactory detection result for small-scaled images. Features extracted from such images contain high-layer semantic information, but they can scarcely express the accurate human face because the projected feature map for small faces quickly shrinks in the last convolution layer of DCNN.

Concerning the problems above, a new face detection network based on Faster R-CNN [20], namely, different scales face detector (DSFD), is proposed in this paper, which improves the detection precision of small images of human face and, thus, improves the precision of uncontrolled face detection under a complicated background. With the RPN in Phase 1, ROIs of the human face are detected with ease in the original image through a boosting cascade face detector. With the ROIs as constraints, anchors are generated through the sliding window on the top feature map, and combined with facial landmarks to extract face proposals, resulting in an increasingly efficient extraction of proposals. In the testing stage, Nontop K suppression (NKS) is applied instead of the nonmaximum suppression (NMS) of Faster R-CNN, which can further improve the recall rate [18]. At the second phase, in order to improve the precision of detection in terms of small-scaled images of human faces, three scale-sensitive Fast R-CNN networks are designed to detect faces in different scale proposals. Proposals produced by RPN are divided into three modes according to the percentages they cover in the images, namely, small, medium, and large. In the testing stage, these three groups of proposals act as the input of each Fast R-CNN network to achieve the face detection featuring the scale of face objects, so as to improve the precision of face detection. Based on the VGG-16 model [22], the proposed face detection system can reach 130 ms to process a frame of image on GPU, and remarkably perform on current mainstream benchmarks, including FDDB, AFW, PASCAL faces, and WIDER FACE. The main contributions of this paper are as follows.

1) A highly efficient multitask RPN of boosting face detector is proposed to improve the extraction efficiency of proposal and the recall rate.
2) A parallel-type Fast R-CNN, which consists of three networks differing from each other in the weight of feature map concatenation, is developed. According to the different percentages they cover on the images, the proposals are assigned to the three corresponding Fast R-CNN networks to improve the precision of face detection.
3) The proposed face detection system acquires remarkable performances on several mainstream benchmarks of face detection, including FDDB, AFW, PASCAL faces, and WIDER FACE.

This paper is structured as follows. Section II reviews the related work. The proposed Faster R-CNN-based face detection framework is described in detail in Section III. Section IV presents the application details. The experimental results and discussion of the proposed DSFD are shown in Section V. Finally, Section VI is devoted to the conclusions and future work.

II. Related Work

Large amounts of works have been proposed in terms of face detection, some of which have already been put to use in our daily life. Just as the developing process of face detection described in [30], early algorithms of human face detection were based on hand-craft features and shallow classifiers. In recent years, approaches based on deep learning have gradually dominated the research of face detection and brought much higher precision.

A. Hand-Craft Feature-Based Face Detection

Viola and Jones proposed the application of Haar feature, AdaBoost-based learning, and cascade-based inference for real-time face detection [1], [2]. Since then, new methods appeared one by one, including building up new local features [5], [31] and new algorithms [32], and to adopt a new cascade structure [34]. As a new requirement for different poses of human faces started to play a role in face detection, some efficient cascade structures [36] based on multiple models were also proposed. DPM [37] was another important breakthrough, in which deformable parts on top of the HOG feature were used to represent objects. Many other improving strategies [8]–[10], [38] emerged on the basis of DPM by applying supervised parts, more pose partition, and better training to improve performance. Shen *et al.* [39] proposed a retrieval-based method combined with discriminative learning. However, it took too much time in the training and testing stage. Wang *et al.* [29] proposed a framework called adaptive sparse representation-based classification, in which sparsity and correlation are jointly considered. After that, Vaillant *et al.* [40] built up a new model by combining face detection with face alignment, which performed remarkably in terms of both precision and speed.

B. Neural Network and Deep Learning-Based Face Detection

In 1994, Vaillant *et al.* [40] used the neural network in face detection for the first time. They proposed a method to detect a face in an image window and scan the whole image with the network at all possible locations by training a convolutional neural network (CNN). As a coarse-to-fine detection strategy, the whole network was divided into two stages. Later, Rowley *et al.* [41] exploited a rationally connected neural network for the detection of an upright frontal face. Osadchy *et al.* [42] presented a joint learning method of face detection, which trained a convolutional network for simultaneous face detection and pose estimation, and improved the precision of face detection. In recent years, the improvements in the performance of human face detection mainly come from the approaches based on deep learning, including [11]–[15], [18], [43], and [44]. Li *et al.* [11] proposed a cascade architecture built on CNNs with very powerful discriminative capability while maintaining high performance. The cascade architecture consists of six convolutional networks, three of which are applied in the positioning of a box while the other three are in calibration.

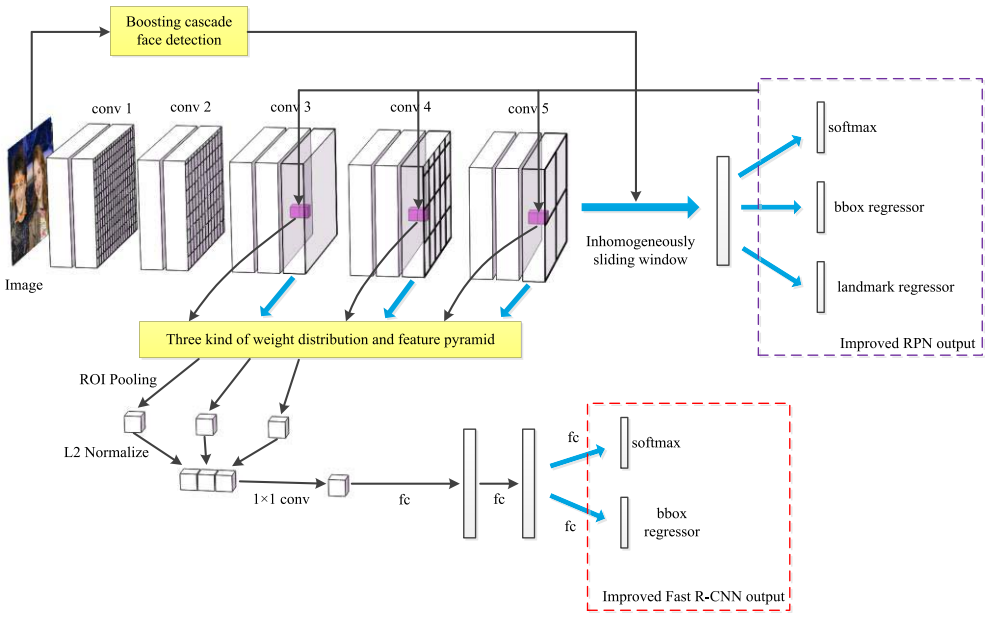


Fig. 1. Architecture of the proposed face detection network.

The six CNNs are independently trained, which results in a very complicated and overloaded process of training. In order to solve such a problem, Qin *et al.* [12] proposed a joint-training network architecture which makes it possible to proceed end-to-end training for several cascade convolutional networks. Ranjan *et al.* [43] exploited a deep multitask learning framework for face detection, landmark localization, pose estimation, and gender recognition. Such a network can predict several tasks related to face features at the same time. In [45], a multitask learning loss function with a master/slave (primary/secondary) relationship was proposed, in which the slaves related to the human face can help improve the prediction of the master. Moreover, some of new methods [33], [35] about the facial expression task achieved good performance on benchmarks. Chen *et al.* [18] exploited facial landmarks as supervision signals to improve face detection performance. In this paper, a general object detection network based on RCNN is applied as the mainframework of the network. Compared with the face detection systems above, our network structure is achieved on the basis of Faster R-CNN [20], which is a general object detection framework. With the properties of RPN and Fast R-CNN [21], such as sharing features, such a framework can guarantee the precision while increasing processing speed simultaneously.

III. Improved Faster R-CNN

A. Network Architecture

In this part, the network structure of our proposed DSFD method is introduced. As stated in Fig. 1, the proposed DSFD method consists of two parts.

The first part is the multitask RPN based on the boosting cascade face detection constraint [18]. With the boosting cascade face detection as a prefilter, the candidate regions of human faces are chosen and then their unions are calculated to obtain one or more ROIs of the human face. The ROI in the original image is then projected to the feature map of conv5_3, in terms of the distribution of which the anchor is inhomogeneously generated. The output of RPN involves the following three tasks: 1) a softmax layer for object classification; 2) a regression function for bounding-box regression; and 3) the least square as loss function for facial regression landmark (reg-landmark). In [23], the experimental result of the JDA detector confirms that facial landmarks are effective in the improvement of face detection.

The second part is a parallel-type Fast R-CNN [21] based on the proposal scale. In the case of human faces in small scales and low resolutions, when ROI pooling is performed to the top feature map, there is a great chance that the projected ROI pooling region is very low in pixels, which results in difficult classification and prediction. To solve the problem above, Bell *et al.* [24] extracted information at multiple scales and levels of abstraction with the aid of ROI pooling. Each of these descriptors becomes a fixed-length descriptor after a series of operations, including L2 normalization, concatenation, scaling, and 1×1 convolution. Feature modification is introduced ahead of ROI pooling. The RPN is divided into three groups, namely, small, medium, and large, according to the percentage its proposal covers on the feature map, with each group corresponding to an independent feature modification. The modifications of the three groups are different from each other in terms of the weight distribution of conv3_3, conv4_3, and conv5_3 of VGG-16.

In order to describe our proposed DSFD method more intuitively, the flowchart for the detection process is shown in Fig. 2, which describes the detailed testing process of our method. The first part is the process of improved RPN, and the second part is the process of improved Fast R-CNN (IFR). The proposals are generated by the improved RPN, and these proposals are distributed to the corresponding Fast R-CNN by their scale for face detection. At last, the detection results of three Fast R-CNNs are merged into the final result.

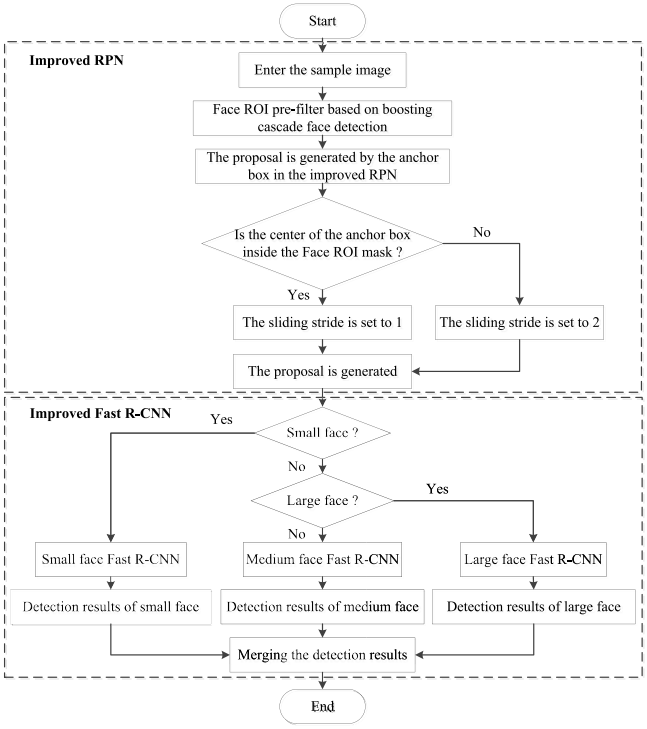


Fig. 2. Algorithm flowchart of the proposed face detection network.

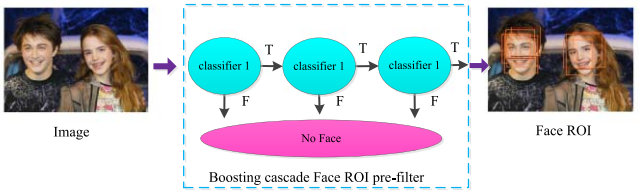


Fig. 3. Illustration of boosting cascade face detection.

B. Face ROI Prefilter

As shown in Fig. 3, a cascade prefilter in [18] is applied, in which some candidate face regions are generated by boosting cascade face detection. Similar to the detector proposed by Voila–Jones [1], the strong classifier consists of 500 weak classifiers. Since a Fern is more powerful and more efficient than a single Haar feature, every weak classifier adopts boosted Fern [25] with each fern containing eight binary nodes. The splitting function is shown as in (1). It compares the threshold value θ_i and the difference values of two pixels at different positions. When the difference value is less than the threshold value θ_i, $s_i = 1$. Otherwise, $s_i = 0$

$$s_i = \begin{cases} 1 & p(x_{1i}, y_{1i}) - p(x_{2i}, y_{2i}) < \theta_i \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where $p(x_{1i}, y_{1i})$ is the value of pixel (x_{1i}, y_{1i}). $(x_{1i}, y_{1i}, x_{2i}, y_{2i}, \theta_i)$ are parameters obtained from training. Cascade classification is learned through the Real-boost algorithm [55]. In each space partition, the classification score is computed as

$$\frac{1}{2}\log\left(\frac{\sum_{\{i\in \text{piece}\cap y_i=1\}}\omega_i}{\sum_{\{i\in \text{piece}\cap y_i=0\}}\omega_i}\right) \tag{2}$$

where ω_i is the weights of positive and negative samples in the space partition.

C. Improved RPN

Chen *et al.* [18] applied boosting cascade face detection for image pretreatment and divided the face ROI into different groups according to their scales. They also proposed an approach concerning ROI convolution to decrease the area of each layer of convolution. Since convolution spends 90% of the time for the whole network process, such an approach can greatly increase the speed of the detector. However, by applying such an approach, ROI masks, the only standard of the decreasing convolution area of each convolutional layer is directly produced from face ROI. Thus, human face object region that cannot be detected in cascade face detection will not be included in ROI masks either. As a result, convolution will not be operated into the face object region in the original image, which results in failures of face detection. Concerning the problem above, we propose an improved RPN aimed at the face object to balance precision and speed, as shown in Fig. 4.

Fig. 4(a) describes how face ROI masks are produced. The face ROI masks are the regions which are the union of all face ROIs output from the face ROI prefilter (FRP). Then, the ROI masks are downsampled to match the scale of the conv5_3 feature map of the RPN. Our proposed improved RPN, which is used to produce face proposal, is shown in Fig. 4(b). A series of anchor boxes is output by a shared small network which slides on the conv5_3 feature map of the RPN. The shared small network consists of an intermediate layer and a full connected layer. The intermediate layer shown in Fig. 4(b) is a convolution layer. The input size and the size of the convolution kernel are both 3×3 in the intermediate layer. It is different from the original RPN [20] that the shared small network is inhomogeneously sliding on the conv5_3 feature map of our improved RPN. The sliding stride in the regions corresponding to the face ROI masks is set to 1, whereas the sliding stride in the other regions is set to 2. Meanwhile, the shape of the human face is considered as a vertical rectangular in the case of most people. Thus, the naive aspect ratios of the anchor are set as 1:1 and 1:2, while the naive scales are set as 128^2, 256^2, and 512^2 pixels. The amount of anchors can be reduced by 45% if such an inhomogeneous sliding window is applied. With the region constraint of face ROI masks, the recall rate of the network is improved as a result.

In improved RPN, when $n = 3$, the 512-D vector flows into three fully connected layers, namely, a box-classification layer (cls), a regression-box layer (reg-box), and a facial regression-landmark layer (reg-landmark). We minimize an objective function following the multitask loss in Fast R-CNN [21], and the loss function for an image is defined as

$$L(\{p_i\},\{t_i\},\{q_i\}) = \frac{1}{N_{\text{cls}}}\sum_i L_{\text{cls}}(p_i, p_i^*) + \lambda\frac{1}{N_{\text{reg}}}\sum_i p_i^* L_{\text{reg}}(t_i, t_i^*) + \beta\frac{1}{N_{\text{reg-landmark}}}\sum_i L_{\text{reg-landmark}}(q_i, q_i^*). \tag{3}$$

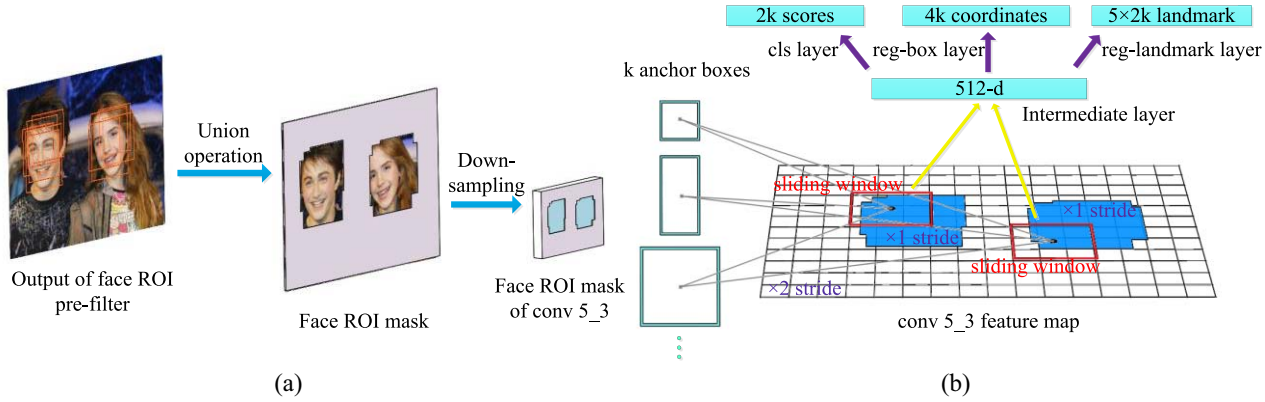


Fig. 4. Improved RPN for face proposal. (a) Process of face ROI to face ROI masks of conv5_3, (b) the multitask RPN with the mode of inhomogeneously sliding windows.

In (3), the first two terms are defined the same as in Faster R-CNN [20]. i is the index of an anchor in a mini-batch. L_{cls} is the log loss over two classes. The regression loss L_{reg} is the robust loss function (smooth L1), being used for the regression parameterization of the four coordinates. In the third term, however, $L_{\text{reg}-\text{landmark}}$ employs the least square as the loss function, in which q_i stands for the coordinates of five facial landmarks of the predicted bounding box and q_i^* is that of the ground-truth box associated with a positive anchor. Besides, the outputs of the cls, reg, and reg-landmark layers consist of $\{p_i\}$, $\{t_i\}$, $\{q_i\}$, respectively. The three terms are normalized with N_{cls}, N_{reg}, and $N_{\text{reg}-\text{landmark}}$, and two balancing weights λ and β. In the discussion of experiments in Section IV, the recall rate of face proposal can be improved by adding reg-landmark into the loss function.

In addition, in the testing stage [19], after the region proposals, NMS is always used to decrease the number of proposals for efficiency. But the problem is that in NMS, even the candidate with the highest confidence score can still be refused by successive Fast R-CNN. In this paper, NKS is applied instead of NMS. The main purpose of NKS is to keep K candidate regions with the highest confidence score in regard to each potential human face, which can effectively improve the recall rate.

D. Improved Fast R-CNN for Different Scale of Proposal

As shown in Fig. 1, the network shared the convolution layer of VGG-16 model with RPN. Small-scaled object detection is challenging for standard Fast R-CNN because the receptive field in the last convolution layer conv5_3 in the standard Faster R-CNN [20] is quite large. For example, a 64×64 face proposal in an image only results in 4×4 pixels in conv5_3 output, which is insufficient to encode facial informative features. On the other hand, the deeper the convolution layer is, the more convolutional information outside the ROI region is gathered for each pixel on the feature map with corresponding depth. In order to solve such a problem, we improve the Fast R-CNN as shown in Fig. 1. The multiscale representation has a positive influence on the improvement of detection precision [24]. We adopt the layers of conv3_3, conv4_3, and conv5_3 in order to stack extract features, where the high resolution information of the lower-level layer will not be lost in terms of small-scaled face objects. On the other hand, we introduce three network-connecting modes ahead of the ROI pooling layer as shown in Fig. 5, including three different weight distributions. In addition, we better integrate the high-level features of low-resolution and high-semantic information into the low-level features of high-resolution and low semantic information. As a result, all features at all scales have rich semantic information. Proposals produced from RPN are divided into three modes according to their coverage in the image, namely, small group, medium group, and large group, in accordance with

$$G = \begin{cases} 1, & 0 < w_p h_p / w_{oi} h_{oi} \le 1/100 \\ 2, & 1/1100 < w_p h_p / w_{oi} h_{oi} \le 1/10 \\ 3, & 1/10 < w_p h_p / w_{oi} h_{oi} \end{cases} \tag{4}$$

where w_p and h_p represent the width and length of the human face proposal, respectively, and w_{oi} and h_{oi} are the width and length of the original image. Being the number of the group, G represents the small, medium, and large groups, respectively, when it is equal to 1, 2, or 3. In the testing stage, the three groups of proposals are input into three corresponding Fast R-CNNs for detection, respectively, where face detection is achieved concerning the scale of face objects, so as to improve its precision.

In Fig. 5, the input of the three network connections is the feature map of conv3_3, conv4_3, and conv5_3 [27]. The network of Fig. 5(a) is used to detect the small faces, and the networks of Fig. 5(b) and (c) are used to detect the medium faces and large faces. The outputs of the network of Fig. 5 are concatenated after ROI pooling and L2 normalization. Fig. 5(a) shows the network structure for detecting small-scaled faces (the small group). Nearest-neighbor upsampling operating on conv4_3 and conv3_3 is processed by the 1×1 convolutional layer to reduce channel dimensions. The two processed feature maps are then added to obtain the layer $P3\_3$ with higher-level semantic information. Finally, ROIs corresponding to the proposals in the feature map $P3\_3$, conv4_3, and conv5_3 are sent to the ROI pooling layer. The three weight values, α_{samll}, β_{samll}, and γ_{samll}, are added to restrain the ROI amplitudes of three different resolutions,

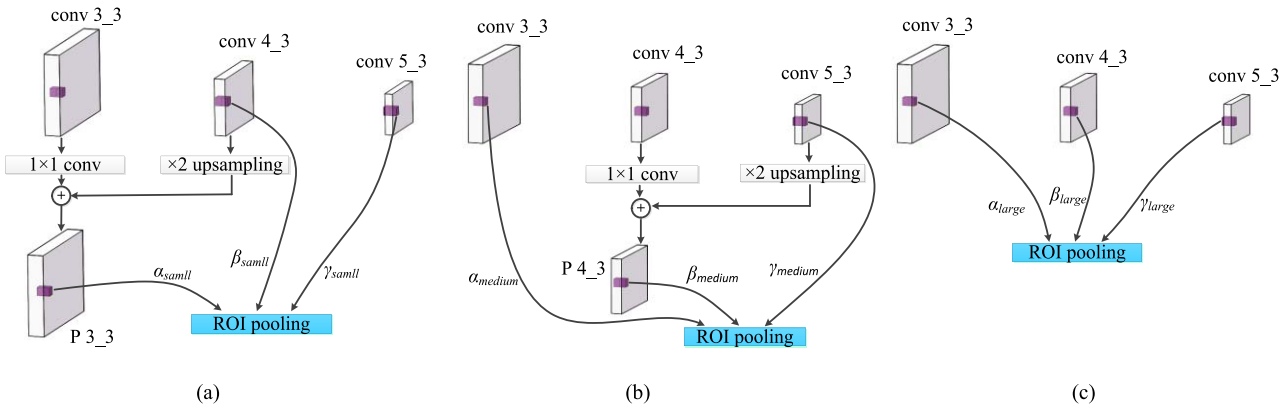


Fig. 5. Three network structures in accordance with different proposal scales. Network structure aiming at (a) small proposals, (b) medium proposals, and (c) large proposals.

respectively. In order to increase the influence produced by *P*3_3 after feature concatenation, α_{samll}, β_{samll}, and γ_{samll} are set as 1/2, 1/3, and 1/3, respectively. Fig. 5(b) shows the network structure for detecting medium-scaled faces (the medium group). Similar to Fig. 5(a), the feature map of conv5_3 is upsampled to match the scale of conv4_3, and the merged feature map obtained by add operation is marked as *P*4_3. The three weights α_{medium}, β_{medium}, and γ_{medium}, valued as 1/3, 1/2, and 1/3 are applied to restrain the ROI amplitudes of conv3_3, *P*4_3, and conv5_3. Fig. 5(c) shows the network structure for detecting large-scaled faces (the large group). Compared to the former two types of connections, Fig. 5(c) does not include upsampling and the 1×1 convolutional layer. The three weights α_{large}, β_{large}, and γ_{large} are valued as 1/3, 1/3, and 1/2 to promote the effect of the high-level feature.

As shown in Fig. 1, the three ROIs corresponding to face proposals are input to three corresponding ROI pooling layers, and normalization outputs are concatenated to be a new feature map with a fixed-length descriptor of size 512×7×7, which is followed by a convolution layer and two fully connected layers. In the final step of this part, there are both a softmax layer for object classification and a regression function for bounding-box refinement.

E. L2 Normalization

As shown in Fig. 1, in order to extend the deep features of face proposal in different scales, the three feature tensors are combined after ROI pooling. Usually, the number and scale of channels differs on each layer of VGG-16, with a smaller scale on a deeper layer. Thus, direct combination of tensors from ROI pooling may result in unpleasant performance, because there is great difference among tensors in terms of their scales. A large-scale feature may take the dominating place, which then weakens the robustness of the algorithm. In order to solve such a problem, every ROI pooling tensor is normalized. L2 normalization is applied to each tensor within each pixel in the pooled feature map tensor [26] ahead of concatenation. The scaling is applied on each tensor independently after normalization. For a d-dimensional input $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)$, L_2-norm[2] is applied as in (5) to normalize the tensor. The L_2 norm of input x is defined as in (6)

$$\hat{x} = \frac{x}{\|x\|_2} \tag{5}$$

$$\|x\|_2 = \left(\sum_{i=1}^{d} |x_i|^2\right)^{\frac{1}{2}} \tag{6}$$

where x is the original pixel vector, $\hat{x}$ is the normalized pixel vector, and d is the number of channels in each ROI pooling tensor. Here, a scaling parameter γ_i is introduced to scale the normalized value through

$$y_i = \gamma_i \hat{x}_i. \tag{7}$$

In the training stage, the scaling factor γ and input data x are calculated with back-propagation and the chain rule as defined in

$$\frac{\partial l}{\partial \hat{x}} = \frac{\partial l}{\partial y} \cdot \gamma \quad \frac{\partial l}{\partial x} = \frac{\partial l}{\partial \hat{x}} \left(\frac{I}{\|x\|_2} - \frac{xx^T}{\|x\|_2^3}\right) \quad \frac{\partial l}{\partial \gamma_i} = \sum_{y_i} \frac{\partial l}{\partial y_i} \hat{x}_i. \tag{8}$$

F. Hard Negative Mining

Hard negative mining has been confirmed as a highly efficient strategy to improve object detection based on the deep convolutional network [28]. The areas that cannot be detected by the network are hard negative. Thus, there has been a vital imbalance between positive and negative training examples. Instead of using all of the negative examples, they are sorted using the highest confidence for each box and the top ones are picked so that the ratio between the positives and negatives is at most 1:3.

G. Data Augmentation

In order to maximize the robustness of our model toward the different sizes and shapes of detection objects, all of the training images are randomly chosen in accordance with the following rules.

1) Use the entire original input image.
2) Randomly sample a patch. The scale of each sampled patch is set in [0.5, 1] of the original image and the aspect ratio in [0.5, 2] of the original image.

3) Randomly mirror the original image.
4) Randomly crop the original image.

IV. Implementation Details

The proposed network is trained through the WIDER FACE dataset [47] which includes 32 203 images and 393 703 labeled human faces that greatly differ from each other in terms of size, pose, and occlusion. Twenty-five thousand images are randomly chosen from the WIDER FACE dataset [47] for the training, while other images are used as a validation set. Five elements are labeled in these images, including the middle point of the left and right eye, nasal tip, and both corners of the mouth. As for the selection of negative samples, labeled regions of human face images from the WIDER FACE dataset [47] are randomly covered with color blocks from which 20 000 images are randomly selected and applied as the first part of the negative samples. Since these images contain a great many elements of human bodies, the Coco database [48], which has pixel-level annotations of various objects, including human beings, is applied as in the previous study [18]. Similarly, all labeled regions of human bodies are covered with color blocks in the images from which 80 000 images are selected and included as the second part of the negative samples. Thus, 100k images of negative samples are applied in the training. For anchors, we use three naive scales with box areas of 128^2, 256^2, and 512^2 pixels, and two naive aspect ratios of 1:1 and 1:2, where six different anchors are produced and applied in the improved RPN part.

At the training stage, Caffe framework [49] is used to train the whole network. The improved RPN can be used as a fully convolutional network. As a result, the improved RPN can be trained end-to-end by back-propagation and stochastic gradient descent (SGD). VGG-16 is applied as the main framework of DSFD which had been pretrained on ImageNet. All images are resized to the scale of 1024/max (w, h) where w and h, respectively represent the width and height of the images. A joint learning strategy similar to the one in [20] is adopted to enable the improved RPN and the IFR to share convolutional layer. The three networks based on Fast R-CNN in our proposed network are separately combined with the improved RPN and trained as three independent networks, which results in a different training procedure from the one in [20]. First, the improved RPN is trained following the steps above, in which the network is initialized by ImageNet pretrained model and fine-turned end-to-end for region proposal task. Meanwhile, proposals are divided into three groups according to (4), namely the small, medium, and large group. Second, all proposals obtained from step one are used to train the testing network, respectively. A lacking IFR (without weight distribution) is trained before the three testing networks being trained in each corresponding group as divided in step one. Thus, three pairs of RPN and Fast R-CNN, namely the small, medium, and large group, are trained. Next, the training of RPN is initialized by using lacking IFR during which the shared convolutional layer is locked while only the exclusive layer of RPN is fine-tuned. Last, the exclusive layer of IFR of the small, medium, and large groups are, respectively, fine-tuned while the shared convolutional layer remains locked. In this way, improved RPN and IFR share one convolutional layer and form a unified network. We fine-tune the model with the SGD solver 60k iterations with a fixed learning rate of 0.001, and run another 20k iterations reducing the learning rate to 0.0001. We also use a momentum of 0.9 and a weight decay of 0.0005.

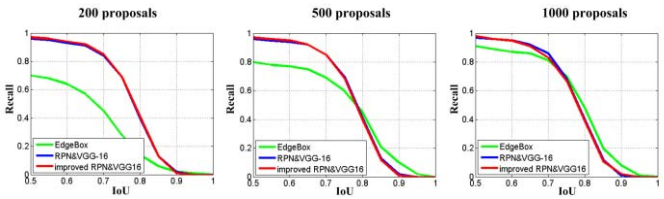


Fig. 6. Recall versus IoU overlap ratio on the FDDB dataset.

V. Experimental Results and Discussion

First, the improved RPN is compared to other methods of proposal extracting in FDDB dataset, including EdgeBox [51] and Faceness-Net [44]. Second, ablation experiments are carried out for the new elements introduced to our method, DSFD, including FRP, improved RPN (with facial landmark), IFR (with three different network structure), NKS, hard negative mining and data augmentation (H&D). Third, our proposed method is compared to current methods of face detection in three benchmark datasets, including FDDB, AFW, PASCAL faces, and WIDER FACE.

A. Evaluation of Face Proposals

First, the performance evaluation method as in [20] is adopted where the Recall-to-IoU is applied to diagnose the method of producing proposal, so as to verify the advantage of our proposed method in maintaining the stability of proposals. Fig. 6 shows the result of involving 200, 500, and 1000 proposals. EdgeBox [51], RPN&VGG-16, and our improved RPN&VGG-16 are compared in FDDB dataset. EdgeBox evaluates the confidence score of each proposal based on the distribution of edge responses. The curves prove the significant stability of our proposed improved RPN when the number of proposals decreases from 1000 to 200, and demonstrates that learning-based methods perform better than heuristic-based methods. It also shows that face detection with a fairly high quality can still be achieved for face detection even with a small amount of proposals in the testing stage.

Then, the performance evaluation method as in [44] is adopted where EdgeBox [51], Faceness-Net [44], and our improved RPN&VGG-16 are compared in FDDB. In Faceness-Net, five CNNs are trained for the different facial attribution parts, including hair, eyes, nose, mouth, and beard. The confidence score of each proposal is based on the response maps of the five different networks. Fig. 7 shows the effectiveness of each method represented by the relation between the number of proposals and detection rate when IoU is set in different values. It is shown that our proposed method achieves greater precision when the same amount of proposal is produced, while less amount of proposals are needed to achieve

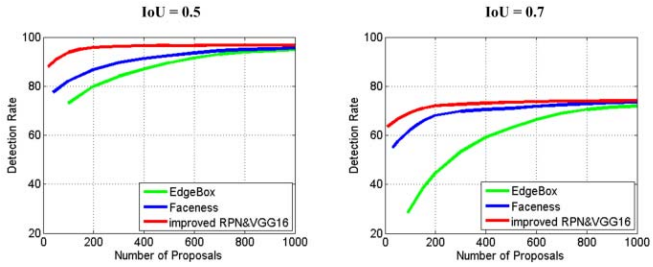


Fig. 7. Detection rate *versus* number of proposals on the FDDB dataset.

TABLE I
EVALUATION OF NINE PARTS WITH ABLATION EXPERIMENT

Group	1	2	3	4	5	6	7	8	9
FRP	No	Yes	No	No	No	No	No	Yes	Yes
FRL	No	No	No	No	Yes	No	No	No	Yes
IR	No	No	No	No	No	Yes	No	Yes	Yes
IFR-CNN	No	No	No	No	No	No	Yes	Yes	Yes
NKS	No	No	Yes	No	No	No	No	Yes	Yes
H&D	No	No	No	Yes	No	No	No	Yes	Yes
Recall Rate (%)	87.3	88.2	87.9	87.6	88.6	89.6	90.1	89.9	90.6

Note: "No" means the corresponding strategy module in the first column is inactive, "Yes" means active.

a precision of the same level. It is demonstrated that in our proposed improved RPN, the method of inhomogeneous sliding window on multitask with facial landmark and feature map is greatly beneficial for the significant effectiveness in producing proposals.

B. Ablation Experiments

Compared to original Faster R-CNN, Ablation experiments are conducted with six new elements introduced in our proposed DSFD including FRP, facial reg-landmark layer (FRL), improved RPN with facial landmark (IR), IFR with three different network structure (IFR), NKS, and H&D. In order to test the performance and effectiveness of each strategy, similar experiment settings are applied as in [18], including evaluating the recall rate in FDDB when the false alarm number is 10. As shown in Table I, nine groups of experiments are processed under different conditions. In group 1, none of the strategies is introduced where an original Faster R-CNN (baseline) is applied in the face detection. In groups 2 to 4, FRP, NKS (K=3), and H&D are introduced, respectively. Compared to the original baseline, the three strategies are slightly effective to the increase of recall rate, which are 0.9%, 0.6%, and 0.3%.

In group 5, the recall rate is improved 1.3% by increased FRL. The reg-landmark layer serves as an auxiliary task to increase the recall rate of main task face detection in the multitask loss function. Groups 6 and 7 represent the increase of two major networks RPN and Fast R-CNN, respectively. In group 6, the method of facial landmark strategy and the inhomogeneous sliding window strategy on feature map in the original RPN accelerate the producing of proposals, which greatly increase the recall rate to 2.3%. In group 7, the recall rate is increased by 2.8%, for the three Fast R-CNNs corresponding to the number of proposals produced from RPN make the detection more object-oriented, which significantly improve the precision of face detection. In group 9, all strategies are combined for application where the highest recall rate is achieved. Comparing the experimental results of group 8 and group 9 in Table I, we can see that the recall rate of our method is reduced by 0.7% in the absence of reg-landmark layer.

C. Comparing on Benchmark Datasets

Comparative experiment is made on three widely used face detection benchmark datasets. The proposed method achieves promising performance on all of the three datasets.

The FDDB dataset [46] consists of 2845 images containing 5171 faces collected from Faces in the Wild dataset. Each face involves more than one challenge, such as pose, low resolution, occlusion, and fuzziness. This dataset is the most popular benchmark for nonrestrictive face detection. The proposed method is compared to several major methods that are currently available, including some traditional methods and the latest research achievement. As shown in Fig. 8, there are two types of scoring the detection in an image, namely the continuous score Fig. 8(a) and the discrete score Fig. 8(b). It can be seen from Fig. 8(a) that our method achieves a recall rate of 83.16% with 200 false positives and 84.52% with 700 false positives, outperforming the baseline method of Jiang and Learned-Miller [50], respectively, by 13.16% and 12.15% in recall rate. Furthermore, our method can improve the effect by 0.92% in recall rate, compared with the currently most efficient method of Wan *et al.* [52] with 350 false positives. As shown in Fig. 8(b), our method achieves a recall rate of 95.32% with 200 false positives and 96.69% with 700 false positives, outperforming the baseline method of Jiang and Learned-Miller [50] by 2.67% and 0.64%, respectively, in recall rate. In addition, our method can improve by 0.74% when compared with the currently most efficient method of Wan *et al.* [52] with 350 false positives. It is shown that our proposed DSFD method achieves promising performance compared to the current methods. Some qualitative results of face detection in FDDB dataset are shown in Fig. 12(a). In addition, our method is compared with Le's method [63] on the FDDB dataset with 200 and 700 false positives. Our method improved by 9.59% and 9.94% in the continuous ROC result, by 7.02% and 5.33% in the discrete ROC result, respectively.

The AFW dataset [8] is built using Flickr images. It contains 205 images with 473 labeled faces. For each face, annotations include one rectangular bounding box, six landmarks, and the pose angles. The PASCAL faces dataset [9] is a subset from PASCAL VOC, and contains 1335 faces from 851 image. Several methods are compared in these two datasets. The PR curves in Figs. 9 and 10 show the state-of-the-art detection accuracies are achieved by our proposed DSFD method which owns 98.4% and 94.43% average precisions (AP), respectively, on AFW and PASCAL faces. Some qualitative results of face detection in AFW and PASCAL faces dataset are shown in Fig. 12 (b) and (c).

The WIDER FACE dataset [47] has 32 203 images and 393 703 labeled faces with a high degree of variability in pose,

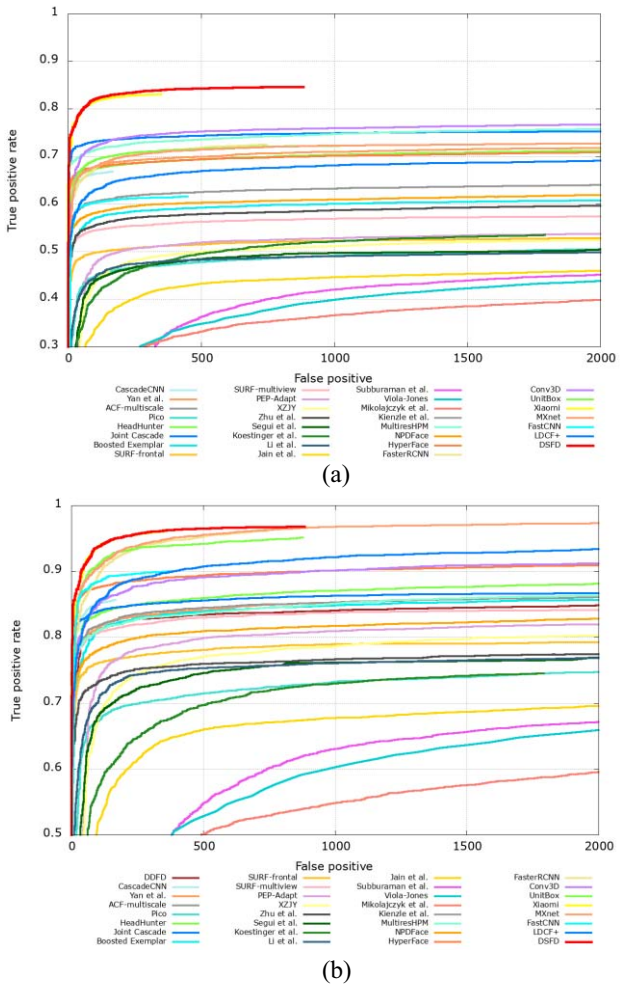


Fig. 8. Comparisons of face detection with state-of-the-art methods on the FDDB dataset. (a) Continuous ROC result. (b) Discrete ROC result.

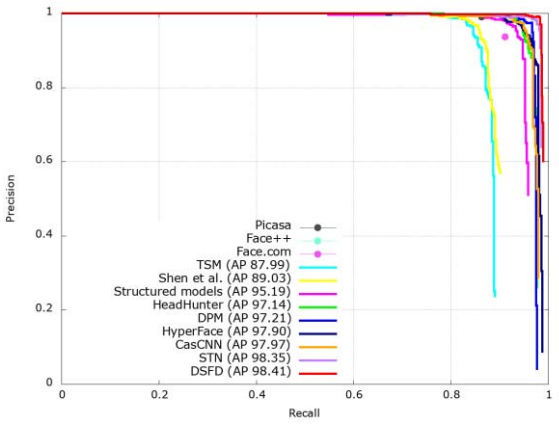


Fig. 9. Precision–recall curves on the AFW dataset.

occlusion, and scale. The dataset is split into three set including training (40%), validation (10%), and testing (50%) set. In addition, the images are divided into three levels (easy, medium, and hard subset) according to the difficulties of the face detection. WIDER FACE dataset is a more challenging dataset and contains a plenty of small faces, especially

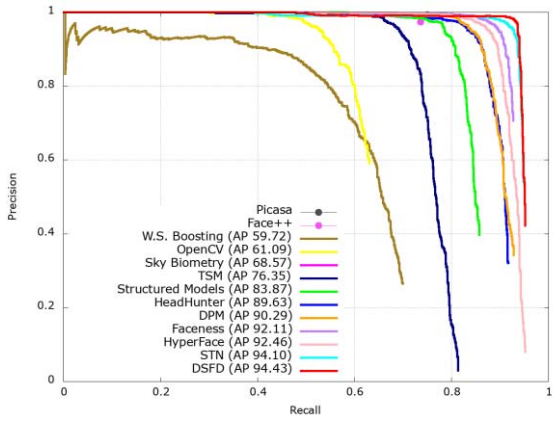


Fig. 10. Precision–recall curves on the PASCAL face dataset.

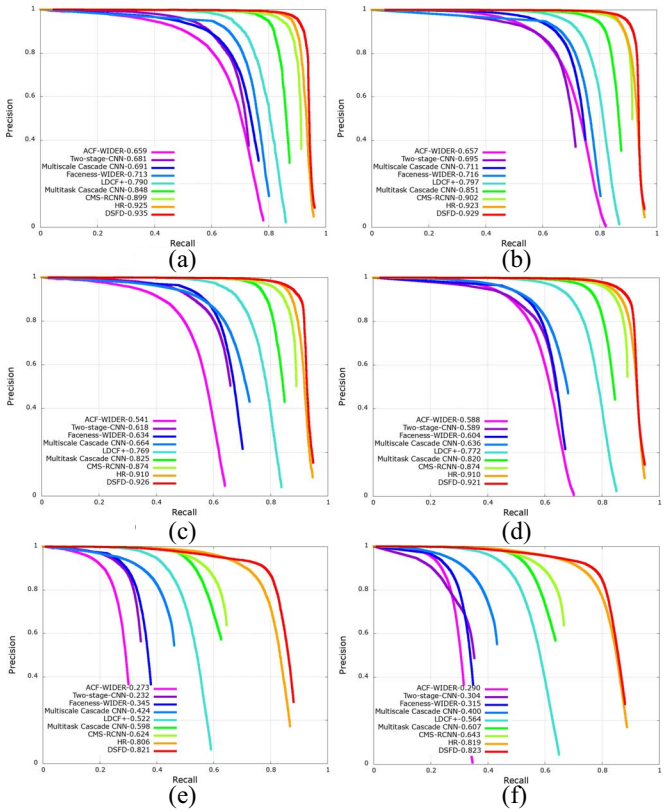


Fig. 11. Precision–recall curves on the WIDER FACE validation and test sets.

in the hard subset. The proposed method is compared to several major methods including HR [60], CMS-RCNN [61], Multitask cascaded CNN [62], *et al.* on the validation and testing set. The precision–recall curves and AP values are shown in Fig. 11. Our method achieves the best AP in all subset, 0.935 (easy), 0.926 (medium), and 0.821 (hard) for validation set, and 0.929 (easy), 0.921 (medium), and 0.823 (hard) for testing set. Some qualitative results of face detection in WIDER FACE dataset are shown in Fig. 12(d). These results demonstrate the proposed method is effectiveness for detecting the small faces.

Fig. 12. Some results of face detection in the four datasets. (a) FDDB dataset. (b) AFW dataset. (c) PASCAL face dataset. (d) WIDER FACE dataset.

D. Comparing on Different Scale Faces

We group faces into three subsets according to different face scales on the FDDB and WIDER FACE database: 1) small scale face subset (10×10 to 50×50 pixels); 2) medium scale face subset (51×51 to 150×150 pixels); and 3) large scale face subset (151×151 pixels or more).

TABLE II
THREE SCALE SUBSETS EXPERIMENTAL RESULTS ON FDDB DATASET

Subset	Method	Recall rate (%) Continuous score	Discrete score
Small Scale Faces	Our Method	**76.02**	**85.08**
	Faster R-CNN[20]	73.42	83.33
Medium Scale Faces	Our Method	**84.01**	**95.04**
	Faster R-CNN[20]	82.71	93.89
Large Scale Faces	Our Method	**93.51**	**99.22**
	Faster R-CNN[20]	93.35	98.96

TABLE III
THREE SCALE SUBSETS EXPERIMENTAL RESULTS ON WIDER FACE DATASET

Subset	Average Precision (%) Our Method	HR[61]	Faster R-CNN[20]
Small Scale Faces	**73.82**	73.27	52.68
Medium Scale Datasets	**95.36**	93.83	93.10
Large Scale Datasets	**98.12**	98.10	97.79

On the FDDB dataset, for each subset, the proposed method is compared to original Faster R-CNN [20] (baseline) with the continuous score and discrete score on 400 false positive. The experimental results are shown in Table II. As can be seen, the bold red numbers indicate the highest recall rate for each subset. Our method achieves better performance than original Faster R-CNN. On the other hand, among the three subsets, the mean difference value of recall rate is 2.18% (small scale datasets), 1.23% (medium scale datasets), 0.21% (large scale datasets), respectively. The experimental results show that our method is more effective in small face detection task.

On the WIDER FACE dataset, for each subset, the proposed method is compared to original Faster R-CNN [20] (baseline) and the second best method HR [60]. The experimental results are shown in Table III. As can be seen, our method achieves an AP of 73.82%, 95.36%, and 98.12 outperforming the baseline method Faster R-CNN [20] by 21.14%, 2.26%, and 0.33% on the small, medium, and large scale face subset, respectively. In addition, from the experimental results we can find out that the performance improvement on the small scale face subset is significantly greater than the other two datasets. It is worth pointing out that our method is effective for improving the accuracy of the small face. On the other hand, our method has a 0.55%, 1.53%, and 0.02% improvement over the second best method HR [60] on the small, medium, and large scale face subset, respectively. The method of HR [60] combines the image pyramid and feature pyramid to improve the detection result of small face. In contrast, our method adopts three networks differing from each other in the weight of feature map concatenation to extract better features for different scale faces and achieves better performances.

TABLE IV
RECALL RATE AND RUNTIME WITH DISCRETE SCORES

Method	Recall rate (%)	Runtime (ms/frame)
Our method	96.69	130
Faster R-CNN	96.05	140
MXnet	96.10	230
UnitBox	94.61	110

E. Runtime Analysis

The testing experiments are carried out on a computer with a Intel Xeon E5@2.0GHz owning 8 cores and a GTX TITAN-X GPU. In the Table IV, we compare our method DSFD, baseline method original Faster R-CNN, MXnet, and UnitBox on 500 images randomly sampled from the FDDB dataset. Under the discrete score with 700 false positives, our method achieves a recall rate of 96.69% with 130 ms to process a frame of image, the original Faster R-CNN and MXnet use 140 ms and 230 ms. Our method is a little faster than these two methods, besides their recall rates are 0.64% and 0.59% lower than our method. UnitBox achieves the best runtime with 110 ms to process a frame of image, but it uses a shallow network to extract features lack of strong semantic information. Therefore, its recall rate is 94.61% lower than our method.

VI. CONCLUSION

In this paper, a new method of face detection based on Faster R-CNN is proposed. The multitask RPN based on cascade face detector is proposed to obtain high-quality face proposals. A parallel-type Fast R-CNN which consists of three networks is also introduced, and each network in the parallel-type Fast R-CNN is chosen to work based on the proposals' scales. In the meantime, the feature concatenation is designed to strengthen the features for face detection. The experimental results show that our proposed DSFD method achieves promising performance on popular benchmarks including FDDB, AFW, PASCAL faces, and WIDER FACE.

Our future work will explore how to improve the efficiency of parallel-type fast R-CNN and optimize the generation process of face proposal. On the other hand, the network compression methods such as network pruning methods and network structured simplification methods will be researched to obtain a compact and fast model which owns less weight parameters and faster running speed.

REFERENCES

[1] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Dec. 2001, pp. 511–518.

[2] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.

[3] J. G. Li and Y. M. Zhang, "Learning SURF cascade for fast and accurate object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Dec. 2016, pp. 3468–3475.

[4] Q. Zhu, M.-C. Yeh, K.-T. Cheng, and S. Avidan, "Fast human detection using a cascade of histograms of oriented gradients," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2006, pp. 1491–1498.

[5] B. Yang, J. Yan, S. Z. Lei, and S. Z. Li, "Aggregate channel features for multi-view face detection," in *Proc. IEEE Int. Joint Conf. Biometrics*, Sep. 2014, pp. 1–8.

[6] C. A. Waring and X. W. Liu, "Face detection using spectral histograms and SVMs," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 35, no. 3, pp. 464–476, Jun. 2005.

[7] R. Khan, A. Hanbury, and J. Stoettinger, "Skin detection: A random forest approach," in *Proc. Int. Conf. Image Process.*, Sep. 2010, pp. 4613–4616.

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

[9] J. J. Yan, X. X. Zhu, Z. Lei, and S. Li, "Face detection by structural models," *J. Image Vis. Comput.*, vol. 32, no. 10, pp. 790–799, 2014.

[10] M. Mathias, R. Benenson, M. Pedersoli, and L. V. Gool, "Face detection without bells and whistles," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2014, pp. 720–735.

[11] H. X. Li, Z. Lin, X. H. Shen, J. Brandt, and G. Hua, "A convolutional neural network cascade for face detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 5325–5334.

[12] H. W. Qin, J. J. Yan, X. Li, and X. L. Hu, "Joint training of cascaded CNN for face detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 3456–3465.

[13] S. S. Farfade, M. J. Saberian, and L.-J. Li, "Multi-view face detection using deep convolutional neural networks," in *Proc. 5th ACM Int. Conf. Multimedia Retrieval*, Jun. 2015, pp. 643–650.

[14] M. Opitz, G. Waltner, G. Poier, H. Possegger, and H. Bischof, "Grid loss: Detecting occluded faces," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2016, pp. 386–402.

[15] Y. Z. Li, B. Y. Sun, T. F. Wu, and Y. Wang, "Face detection with end-to-end integration of a ConvNet and a 3D model," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2016, pp. 420–436.

[16] S. Yang, P. Luo, C. C. Loy, and X. O. Tang, "From facial parts responses to face detection: A deep learning approach," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3676–3684.

[17] C. Zhang and Z. Y. Zhang, "Improving multiview face detection with multi-task deep convolutional neural networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2014, pp. 1036–1041.

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

[19] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[20] S. Q. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[21] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1440–1448.

[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv Preprint arXiv: 1409.1556*, 2014.

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

[24] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick, "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Dec. 2016, pp. 2874–2883.

[25] M. Ozuysal, P. Fua, and V. Lepetit, "Fast keypoint recognition in ten lines of code," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Dec. 2007, pp. 1–8.

[26] W. Liu, A. Rabinovich, and A. C. Berg, "ParseNet: Looking wider to see better," *arXiv preprint arXiv:1506.04579*, 2015.

[27] T. Y. Lin *et al.*, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 936–944.

[28] W. Liu *et al.*, "SSD: Single shot MultiBox detector," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2016, pp. 21–37.

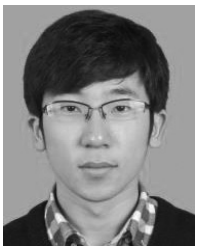
[29] J. Wang *et al.*, "Robust face recognition via adaptive sparse representation," *IEEE Trans. Cybern.*, vol. 44, no. 12, pp. 2368–2378, Dec. 2014.

[30] S. Zafeiriou, C. Zhang, and Z. Y. Zhang, "A survey on face detection in the wild: Past, present and future," *Int. J. Comput. Vis. Image Understanding*, vol. 138, pp. 1–24, Sep. 2015.

[31] L. Zhang, R. F. Chu, S. M. Xiang, S. C. Liao, and S. Z. Li, "Face detection based on multi-block LBP representation," in *Proc. Int. Conf. Biometrics*, 2007, pp. 11–18.

[32] C. Hang, H. Z. Ai, Y. Li, and S. H. Lao, "High-performance rotation invariant multiview face detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 4, pp. 671–686, Apr. 2007.

[33] A. Majumder, L. Behera, and V. K. Subramanian, "Automatic facial expression recognition system using deep network-based data fusion," *IEEE Trans. Cybern.*, vol. 48, no. 1, pp. 103–114, Jan. 2018.
[34] L. Bourdev and J. Brandt, "Robust object detection via soft cascade," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 236–243.
[35] P. Rodriguez *et al.*, "Deep pain: Exploiting long short-term memory networks for facial expression classification," *IEEE Trans. Cybern.*, to be published.
[36] S. T. Li *et al.*, "Statistical learning of multi-view face detection," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2002, pp. 67–81.
[37] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.
[38] J. J. Yan, Z. Lei, L. Y. Wen, and S. Z. Li, "The fastest deformable part model for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2497–2504.
[39] X. H. Shen, Z. Lin, J. Brandt, and Y. Wu, "Detecting and aligning faces by image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3460–3467.
[40] R. Vaillant, C. Monrocq, and Y. Le Cun, "Original approach for the localisation of objects in images," *IEE Proc. Vis. Image Signal Process.*, vol. 141, no. 4, pp. 245–250, Aug. 1994.
[41] H. A. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 1, pp. 23–38, Jan. 1998.
[42] M. Osadchy, Y. Le Cun, and M. L. Miller, "Synergistic face detection and pose estimation with energy-based models," *Mach. Learn. Res.*, vol. 8, pp. 1197–1215, Jan. 2007.
[43] R. Ranjan, V. M. Patel, and R. Chellappa, "HyperFace: A deep multi-task learning framework for face detection, landmark localization, pose estimation, gender recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published.
[44] S. Yang, P. Luo, C. C. Loy, and X. O. Tang, "Faceness-Net: Face detection through deep facial part responses," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 8, pp. 1845–1859, Aug. 2018.
[45] Z. P. Zhang, P. Luo, C. C. Loy, and X. O. Tang, "Facial landmark detection by deep multi-task learning," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2014, pp. 94–108.
[46] V. Jain and E. G. Learned-Miller, "FDDB: A benchmark for face detection in unconstrained settings," Univ. Massachusetts at Amherst, Amherst, MA, USA, Rep., 2010.
[47] S. Yang, P. Luo, C. C. Loy, and X. O. Tang, "WIDER FACE: A face detection benchmark," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Dec. 2016, pp. 5525–5533.
[48] T. Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2014, pp. 740–755.
[49] Y. Q. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22nd AMC Int. Conf. Multimedia*, 2014, pp. 675–678.
[50] H. Z. 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.
[51] P. Dollár and C. L. Zitnick, "Fast edge detection using structured forests," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 8, pp. 1558–1570, Aug. 2015.
[52] S. H. Wan, Z. J. Chen, T. Zhang, B. Zhang, and K.-K. Wong, "Bootstrapping face detection with hard negative examples," *arXiv Preprint arXiv: 1608.02236v1*, 2016.
[53] K. M. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2017, pp. 2980–2988.
[54] S. Hayashi and O. Hasegawa, "Robust face detection for low-resolution images," *J. Adv. Comput. Intell. Intell. Inf.*, vol. 10, no. 1, pp. 93–101, 2016.
[55] B. Wu, H. Z. Ai, C. Huang, and S. H. Lao, "Fast rotation invariant multi-view face detection based on real Adaboost," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, May 2004, pp. 79–84.
[56] J. Zheng, G. A. Ramírez, and O. Fuentes, "Face detection in low-resolution color images," in *Proc. 7th Int. Conf. Image Anal. Recognit.*, Jun. 2010, pp. 454–463.
[57] T. Kong, A. Yao, Y. Chen, and F. Sun, "HyperNet: Towards accurate region proposal generation and joint object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, Dec. 2016, pp. 845–853.
[58] X. Zhang, H. Xiong, W. Lin, and Q. Tian, "Weak to strong detector learning for simultaneous classification and localization," *IEEE Trans. Circuits Syst. Video Technol.*, to be published.
[59] S. Zhang, W. Lin, P. Lu, W. Li, and S. Deng, "Kill two birds with one stone: Boosting both object detection accuracy and speed with adaptive patch-of-interest composition," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, Jul. 2017, pp. 447–452.
[60] P. Hu and D. Ramanan, "Finding tiny faces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1522–1530.
[61] C. Zhu, Y. Zheng, K. Luu, *et al.*, "CMS-RCNN: Contextual multi-scale region-based CNN for unconstrained face detection," in *Deep Learning for Biometrics*, 2016.
[62] 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, Apr. 2016.
[63] T. H. N. Le, Y. Zhang, C. Zhu, K. Luu, and M. Savvides, "Multiple scale faster-RCNN approach to driver's cell-phone usage and hands on steering wheel detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2016, pp. 46–53.



Wenqi Wu received the B.Sc. degree in electronic and information engineering from the North China University of Technology, Beijing, China, in 2013 and the Ph.D. degree in pattern recognition and intelligent system from the Institute of Automation, Chinese Academy of Sciences, Beijing, in 2018.

He is currently a Researcher with Tencent Research, Beijing. His current research interests include deep learning, machine learning, and computer vision.

Yingjie Yin received the B.Sc. degree in control science and engineering from the Ocean University of China, Qingdao, China, in 2011 and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2016.

He is currently an Assistant Professor with the Research Center of Precision Sensing and Control, Institute of Automation, Chinese Academy of Sciences. His current research interests include computer vision and machine learning.

Xingang Wang received the B.Sc. degree in semiconductor physics and devices from Tianjin University, Tianjin, China, in 1995 and the Ph.D. degree in mechanical manufacturing and automation from the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, China, in 2002.

He is currently a Professor with the Research Center of Precision Sensing and Control, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His current research interests include image processing and machine learning.

De Xu (M'05–SM'09) received the B.Sc. and M.Sc. degrees from the Shandong University of Technology, Jinan, China, in 1985 and 1990, respectively, and the Ph.D. degree from Zhejiang University, Hangzhou, China, in 2001, all in control science and engineering.

Since 2001, he has been with the Institute of Automation, Chinese Academy of Sciences, Beijing, China, where he is currently a Professor with the Research Center of Precision Sensing and Control. His current research interest includes robotics and automation, such as visual measurement, visual control, intelligent control, visual positioning, microscopic vision, and micro-assembly.