IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 9, 2020, accepted September 6, 2020, date of publication September 14, 2020, date of current version October 5, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3023782

Face Detection Based on Receptive Field Enhanced Multi-Task Cascaded Convolutional Neural Networks

XIAOCHAO LI[1,2], **(Senior Member, IEEE), ZHENJIE YANG**[1], **AND HONGWEI WU**[3], **(Member, IEEE)**

[1]Department of Microelectronics and Integrated Circuit, Xiamen University, Xiamen 361005, China
[2]Department of Electrical and Electronics Engineering, Xiamen University Malaysia, Sepang 43900, Malaysia
[3]Xiamen Network Information Security Joint Laboratory, Xiamen 361000, China

Corresponding author: Xiaochao Li (leexcjeffrey@xmu.edu.cn)

This work was supported in part by the Xiamen Industry—University Collaboration Research Project under Grant 3502Z20203004, in part by the Xiamen University Malaysia Research Fund under Grant XMUMRF/2019-C4/IECE/0008, in part by the Fujian Key Laboratory of Integrated Circuit Design and Measurement, Xiamen University, and in part by the Innovation Fund of Xiamen University.

ABSTRACT With the continuous development of deep learning, face detection methods have made the greatest progress. For real-time detection, cascade CNN based on the lightweight model is still the dominant structure that predicts face in a coarse-to-fine manner with strong generalization ability. Compared to other methods, it is not required for a fixed size of the input. However, MTCNN still has poor performance in detecting tiny targets. To improve model generalization ability, we propose a Receptive Field Enhanced Multi-Task Cascaded CNN. This network takes advantage of the Inception-V2 block and receptive field block to enhance the feature discriminability and robustness for small targets. The experimental results show that the performance of our network is improved by 1.08% on the AFW, 2.84% on the PASCAL FACE, 1.31% on the FDDB, and 2.3%, 2.1%, and 6.6% on the three sub-datasets of the WIDER FACE benchmark in comparison with MTCNN respectively. Furthermore, our structure uses 16% fewer parameters.

INDEX TERMS Face detection, cascade convolutional neural networks, receptive field.

I. INTRODUCTION

Face detection is the basis in the field of computer vision and pattern recognition, as well as a fundamental step of face-related research, such as face recognition [1], verification [2], and tracking [3]. After decades of development and research, face detection has been widely used in various aspects of life, such as security monitoring. It has increasingly become a research hotspot in the field of video images.

There are some widely used non-neural network-based face detectors, such as skin-color detection, SVM classifier [4]. Classic image feature extraction algorithms achieve good accuracy with real-time efficiency for face detection. Ma *et al.* [5] proposed an AdaBoost-based training method to obtain cascade classifiers with multiple feature types: Haar-like, HOG for an improved discrimination ability. However, this requires high computation due to containing too many weak classifiers. An algorithm based on the Bayesian framework [6] used the Omega shape formed by a person's head and shoulder for head localization to tackle severe face occlusion. It achieves good performance in detection faces with severe occlusion, but the scene is restricted in Automatic Teller Machines. Beyond the AdaBoost-based methods, Mathias *et al.* [7] proposed face detection with deformable part models (DPM) and obtain impressive results. However, this method usually suffers from high computational cost. Another method [8] is proposed based on DPM for detecting faces with occlusion. It can reduce the false-negative face detection and error rate for detection, however, it has poor universality for only frontal face images used in the experiments.

In recent years, the face detection method based on a convolutional neural network (CNN) has made a breakthrough and become the mainstream of the face detection method. Several studies [9], [10] utilize deep CNN for face detection and have a better performance on face detection. Faster R-CNN [11] and other CNN-based two-stage or one-stage algorithms, with the help of deep convolutional networks such as VGGNet [12] and ResNet [13], achieve superior performance. Nevertheless, due to a surplus of convolutional

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

layers, the speed of detection slows down greatly. Hence, some models with a multi-stage face detection algorithm that has a relatively high True Positive Rate and a real-time speed are proposed. Wu *et al.*, [10] proposed a funnel-structured cascade (FuSt) detection constituted by multiple view-specific fast LAB cascade, multiple coarse MLP cascade, and a unified fine MLP cascade. Farfade *et al.*, [14] presented a fast CNN's cascade face detector, using a CNN with a novel pyramid architecture, multi-layer merging, knowledge distilling online and offline hard sample mining.

Multi-task Cascaded Convolutional Neural Networks (MTCNN) [9] is the dominant multi-stage and multi-task structure in recent years. Different from generic object detection, face detection features much larger scale variations (from several pixels to thousand pixels). Hence, the image pyramid method adopted by MTCNN could not perform well on faces with a high degree of variability in scale, especially for the tiny face. The main reason lies in that the first stage, the P-net of MTCNN which produces candidate windows quickly through a shallow CNN, puts the limit of the performance of the entire network. The shallow structure of P-net cannot cover all the size of the receptive field to extract the high discriminative feature with standard CNN, which come from deeper neural networks. As the kernel of standard convolution is sampled at the same center and commonly set receptive field at the same size with a regular sampling grid on a feature map, which probably induces some loss in the feature discriminability as well as robustness.

According to the discussion above, we propose a new model called Receptive Field Enhanced Multi-Task Cascaded CNN (RFE-MTCNN), which integrates the ideal from the Inception-V2 Block [15] and Receptive Field Block [16] to build a fast yet powerful face detector with the reasonable alternatives in a different stage of the cascade network. The ideal is to enhance the network's receptive field for feature representation by bringing in certain hand-crafted mechanisms rather than stubbornly deepening the model. At the same time, we import Additive Angular Margins (AM) [17] into Softmax to optimize loss function. Extensive experiments on the Wider Face and FDDB datasets show that the proposed method achieves state-of-the-art performance compared with MTCNN variants.

The major contributions of this paper are summarized as follows: (1) We propose a new face detection model RFE-MTCNN which takes advantage of the Inception-V2 block and receptive field block to enhance the feature discriminability and robustness for small targets. (2) We use the Global Average Pooling (GAP) to replace the second to last fully connected layers in order to enforce correspondences between feature maps and categories, avoid overfitting, and reduce the network parameters. (3) The AM-Softmax loss function is introduced to enhance the discriminability of the R-Net.

The remainder of the paper is organized in the following manner. Section 2 presents the related technologies involved in this paper. Section 3 provides a detailed description of our new proposed method of RFE-MTCNN. In section 4, we show the experimental settings and compare RFE-MTCNN to other state-of-the-art algorithms on FDDB, Wider Face. Finally, the paper is drawn to the conclusion in Section 5.

II. RELATED WORK

Inception-V2 Block [15] is composed of multiple different branches. And the receptive field of the feature map is enhanced by convolution kernels of different sizes. It adds a BN layer based on Inception-V1 [18], which accelerates the convergence speed of the network. At the same time, two 3*3 convolutions are connected in series to replace a 5*5 convolution, and the parameter amount is reduced under the condition that the receptive fields are the same. Compared with Inception-V2 block, its derived structure Receptive Field Block (RFB) [16] adopts multiple branch structures. Each branch is constructed using a combination of conventional convolution and dilated convolutions of different proportions. Convolution kernels of different sizes can simulate different sizes of the overall receiving field (pRF) [16] [13]. Its dilated convolution layer uses a separate eccentricity to simulate the ratio between pRF size and eccentricity. Inception-V2 has a similar structure to RFB, and it realizes a multi-size receptive field through a multi-branch structure. But the difference between the two is that in the Inception-V2 structure, the convolution has the same sampling center. So part of the edge information will be lost. But RFB exploits dilated convolution to simulate the impact of the eccentricities of pRFs in the human visual cortex. The kernel size and dilation have a similar positive functional relation as that of the size and eccentricity of pRFs in the visual cortex. The pixel in the feature map contributes the same to the output response. The spatial RF structure of Inception-V2 Block and RFB is shown in Fig. 1 and Fig. 2.

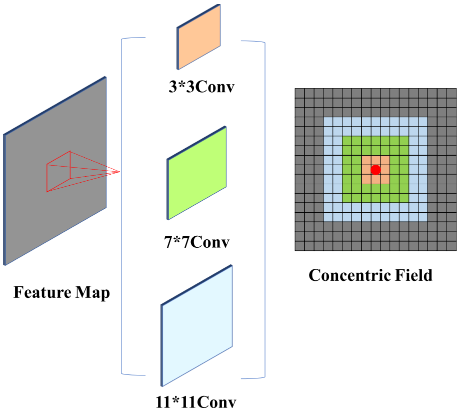

FIGURE 1. Spatial RFs of inception-V2 block.

Generally speaking, object detection needs to make predictions on the last layer of feature maps. And the receptive field of the last layer of feature maps determines the upper limit of the size that the network can detect. As usual, downsampling

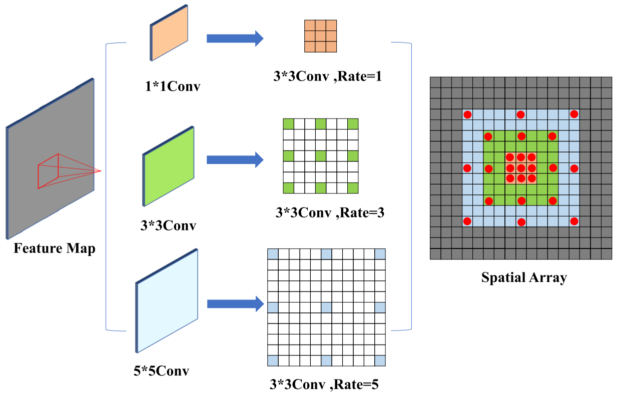

FIGURE 2. Spatial RFs of RFB.

can enhance the receptive field of the feature map, but it makes small targets difficult to detect. Common downsampling methods, such as standard convolution and pooling, can enhance the receptive field of the feature map, but its spatial resolution will be reduced, and the pooling operation will lose the information of the feature map. Dilated convolution can enhance the receptive field of the feature map without losing information. Although it has problems such as gridding effects, we use hybrid dilated convolution [19] to effectively avoid this problem.

III. THE PROPOSED METHOD

Traditional MTCNN [9] uses standard convolution. As the network depth increases, its receptive field will also increase, which is conducive to the detection of large-sized faces, but not conducive to the detection of small-sized faces. To solve this problem, we use the Inception-V2 Block and its derived Receptive Field Block (RFB) to use its multiple branches to enhance the receptive field of the feature map.

MTCNN is a framework that integrates the face detection and face alignment tasks using unified cascaded CNNs by multi-task learning. In MTCNN itself, it is made up of three networks. The first network, called Proposal Network (P-Net), mainly obtains candidate windows and their bounding box regression vector and uses non-maximum suppression (NMS) to merge the boxes that highly overlap. The second network which is known as Refine Network (R-Net), is employed to filter a large number of false candidates from P-Net and calibrates the bounding box with regression. Last but not least, the final network with the name Output Network (O-Net), outputs the final candidate windows and five facial landmarks' positions with a deeper network.

Combining the characteristics of MTCNN and two Blocks that enhance the receptive field, we introduced RFB in P-Net. It enhances the deep features in the neural network and retains the edge part of the feature map to obtain more accurate candidate boxes. In R-Net and O-Net, the input is the face detection candidate box of the superior network. If the candidate frame is the detection target, its central area contributes a lot to the output response, so Inception-V2 Block based on central sampling is introduced to improve the screening ability of R-Net and O-Net to the candidate frame. At the same time, we use the Global Average Pooling (GAP) [20] to replace the connection layers in the last layers. In order to enhance the discriminative ability of face classification in the R-Net, we adapt AM-Softmax [17] to push the face /non-face decision boundaries away from each other. Based on ablation experiments, it is shown that the fully connected layer with AM-Softmax in R-Net achieves better results.

A. FACE DETECTION NETWORK

Fig. 3, Fig. 4, and Fig. 5 shows the three proposed subnetworks of RFE-MTCNN. As can be seen from the figure, we introduced the RFB structure in P-Net. At the same time, the Inception block was introduced in R-Net and O-Net. Besides, maximum pooling is adopted between each block for feature dimensionality reduction. Fig. 6 and 7 show the RFB and Inception-V2 block, respectively.

Three tasks are used to train CNN detectors: face/non-face classification, bounding box regression, and facial landmark localization. For the first task of the face and non-face classification, we use Additive Margin Softmax (AM-Softmax) [17], which introduces additive margin to softmax loss function as follows.

$$\varphi\left(\theta\right) = cos\left(\theta\right) - m \tag{1}$$

$$\begin{aligned} L_i^{det} &= -(y_i^{det} \cdot log(\frac{e^{s\cdot(cos(\theta_i)-m)}}{e^{s\cdot(cos(\theta_i)-m)} + e^{s\cdot cos(\theta_i)}}) \\ &\quad + (1 - y_i^{det}) \cdot log(\frac{e^{s\cdot cos(\theta_i)}}{e^{s\cdot(cos(\theta_i)-m)} + e^{s\cdot cos(\theta_i)}})) \end{aligned} \tag{2}$$

where θ_i is the target angle between normalized weights and normalized features and i denotes the $i - th$ sample. The hyperparameter s and m are set to 30 and 0.35 respectively, which achieve good results in face recognition tasks [17]. The notation y_i^{det} denotes the ground-truth label. The margin is enforced by subtracting m from $cos\left(\theta\right)$ rather than m multiplied to θ, so that the derivative will not change during backpropagation. On the other hand, the additive margin enlarges the differences between face and background thus making the learning of the classification task more difficult.

On the other hand, bounding box regression is formulated in Equation 3,

$$L_i^{box} = \left\|\hat{y}_i^{box} - y_i^{box}\right\|_2^2 \tag{3}$$

where regression target $\hat{y}_i^{box}$ is obtained by the network and y_i^{box} is the ground-truth coordinate. Four coordinates are x, y of the upper left corner, height and width.

Last but not least, facial landmark localization is formulated as follows:

$$L_i^{landmark} = \left\|\hat{y}_i^{landmark} - y_i^{landmark}\right\|_2^2 \tag{4}$$

Equation. 4 is the Euclidean loss, and facial landmark detection is formulated as a regression problem. There are

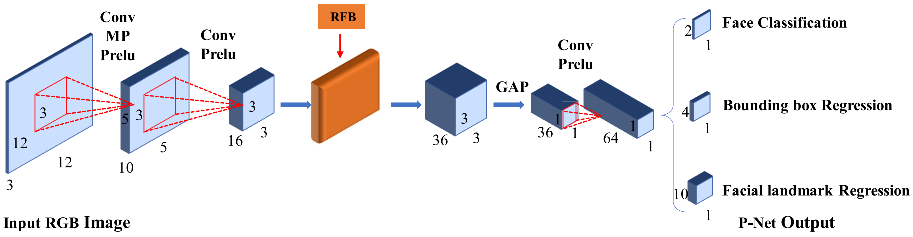

FIGURE 3. The architecture of P-Net.

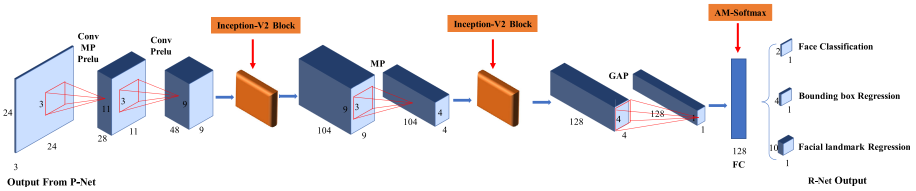

FIGURE 4. The architecture of R-Net.

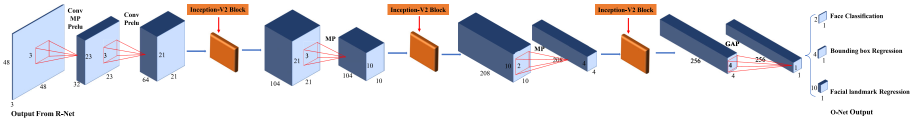

FIGURE 5. The architecture of O-Net.

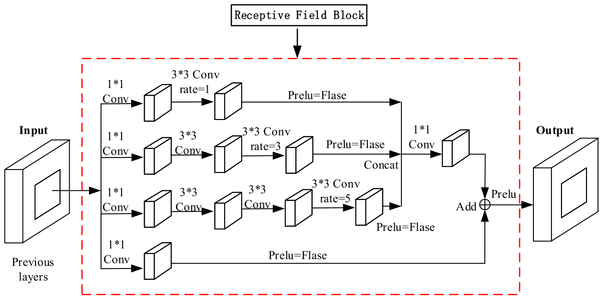

FIGURE 6. The architecture of RFB.

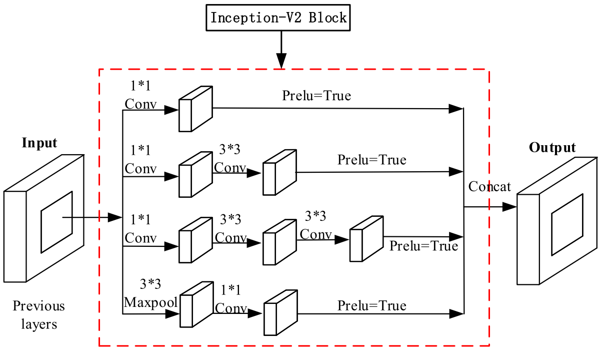

FIGURE 7. The architecture of inception-v2 block.

five facial landmarks, including left eye, right eye, nose, left mouth corner, and right mouth corner.

$$min \sum_{i=1}^{N} \sum_{j\in(det,box,landmark)} \alpha_j \beta_i^j L_i^j \tag{5}$$

Then the overall learning target can be formulated as Equation. 5, where N is the number of training samples. α_j denotes on the task importance and $\beta_i^j \epsilon (0, 1)$ is the label of the $j - th$ sample.

B. FACE DETECTION PROCESS

The training and testing phase of RFE-MTCNN are performed in the three networks, i.e., P-Net, R-Net, and O-Net.

When training the P-Net, first randomly crop images in the dataset and resize the cropped images to 12*12. Then determine the cropped image is a positive or negative sample based on the Intersection over Union(IOU) ratio of the box to ground truth. Secondly, when training the R network, detect images in the dataset with a trained P-Net model, each image will generate a large number of candidate windows. For each candidate window, according to its' IOU with ground truth, this candidate window is determined to be a positive and negative sample. After, resize these windows to 14*14 and train R-Net. Finally, similar to R-Net, the trained R-Net model is used to generate candidate windows, the candidate windows are determined to be positive and negative samples according to its' IOU with ground truth. Finally, resize these windows to 48*48 and train O-Net. The steps for training the proposed RFE-MTCNN are shown in Fig. 8.

When inference is performed, first of all, generate an image pyramid of different scales. The candidate bounding boxes and scores are initially obtained by P-Net. And then candidate bounding boxes with large overlap are eliminated through NMS. Next, merge overlapped candidates of different scales. Secondly, detect the image with the P-Net model and convert

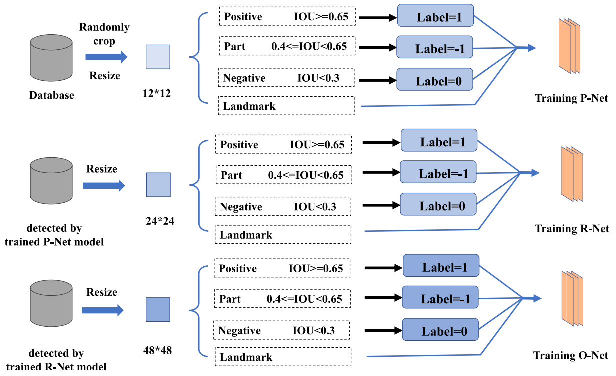

FIGURE 8. The steps for training the proposed RFE-MTCNN.

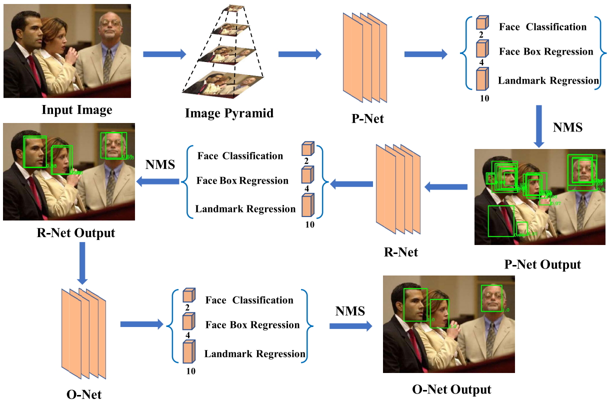

FIGURE 9. The steps for face detection with RFE-MTCNN.

the detected candidate windows of the face into the square boxes. Afterward, convert these square boxes in the original image to new boxes starting at 0 coordinates and resize the new boxes to 24×24. Subsequently, use the R-Net model to detect these new boxes and get R-Net's candidate windows of the face and scores. After, merge overlapped candidate windows with NMS. Finally, similar to R-Net, use the O-Net model to detect these new boxes and output bounding boxes and scores. The steps for face detection with RFE-MTCNN are shown in Fig. 9, respectively.

IV. EXPERIMENTS AND RESULTS

A. DATASETS USED FOR TRAINING AND TESTING

WIDER FACE [21] dataset is a face detection benchmark dataset, which is a challenging dataset and is widely used to study the problem of unconstrained face detection. It contains 393,703 faces with a high degree of variability in scale, poses, and occlusion.

CelebFaces Attributes (CelebA) Dataset [22] is a large-scale face attributes dataset with more than 200K face images and 10,177 identities, and 5 landmark locations per image.

Face Detection Data Set and Benchmark (FDDB) [23] is a widely used public dataset. It contains the annotations for 5171 faces in a set of 2845 images.

Annotated Faces in the Wild (AFW) [24] Dataset contains 205 images with 473 labeled faces.

PASCAL face dataset [25] has 1335 labeled faces in 851 images with large face appearance and pose variations. It is collected from the PASCAL person layout test subset.

B. EXPERIMENT SETTING

1) TRAINING

We choose WIDER FACE and CelebA as the training datasets for training the proposed RFE-MTCNN. Similar to MTCNN, the entire training dataset contains 215,479 images, of which the WIDER FACE dataset has 12,880 images and the CelebA dataset has 202,599 images. Four kinds of data annotation are used in the training process, negatives, positives, part faces, and landmark faces. We set the same parameter values as MTCNN. Positives mean that the Intersection-over-Union (IOU) ratio is more than 0.65 to a ground truth face. Part faces are between 0.4 and 0.65. Negatives are lower than 0.3. Landmark faces label the locations of the left eye, right eye, nose, left mouth corner, right mouth corner. We use (classification = 1, bounding box = 0.5, landmark = 0.5) in P-Net and R-Net, meanwhile, (classification = 1, bounding box = 0.5, landmark = 1) in O-Net. These numbers denote the importance of classification loss function, bounding box regression loss function, and landmark regression loss function.

Three networks are trained in order and the steps of training networks are as Fig. 8. It should be noted that the trained P-Net model is used to detect pictures to obtain the samples when R-Net is trained, so the steps of training R-Net include the steps of P-Net detection. Similarly, the trained R-Net model is used to detect pictures to obtain the samples when O-Net is trained, so the steps of training O-Net include the steps of P-Net and R-Net detection.

2) TEST

We conduct considerable test experiments on the public-domain face detection benchmark: FDDB dataset, Wider Face dataset, AFW dataset, and PASCAL dataset. The steps of face detection with RFE-MTCNN are shown in Fig. 9.

C. EXPERIMENT RESULTS

To better understand the proposed RFE-MTCNN, we conducted extensive ablation experiments to examine how the improvement of different network substructures and the introduction of AM-Softmax quantitatively affect the performance of face detection.

Table 1 shows that we conducted an ablation experiment on the proposed model. It can be seen from the table that the improved sub-network has a certain improvement to the entire network, but the performance of the overall network will also be limited to other cascaded networks. Besides, adding the Inception-V2 Block to O-Net has the greatest improvement on the network. AM-Softmax has improved network

TABLE 1. Ablation experiments of the proposed methods on FDDB dataset.

Method			Ture Positive Rate
P-Net	R-Net	O-Net	False Positive=2000
baseline	baseline	baseline	95.04%
+RFB	baseline	baseline	95.26%
+RFB	+Inc-V2	baseline	95.51%
+RFB	+Inc-V2+AM-Softmax	baseline	95.69%
baseline	+Inc-V2+AM-Softmax	+Inc-V2	96.13%
+RFB	+Inc-V2+AM-Softmax	+Inc-V2	96.35%

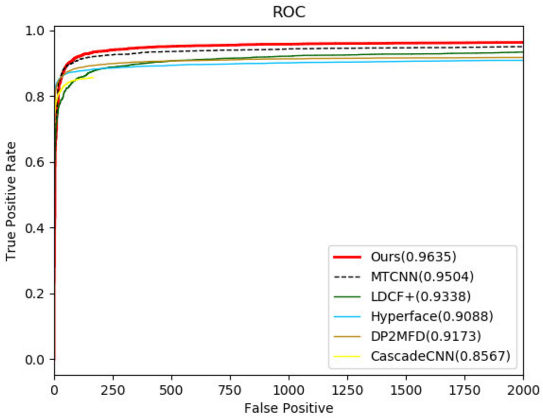

FIGURE 10. ROC curves on FDDB database.

TABLE 2. Detection performance comparison on FDDB dataset.

Method	True Positive Rate False Positive=2000
Multiscale Cascade [26]	85.67%
Hyperface [27]	90.88%
DP2MFD [28]	91.73%
LDCF+ [29]	93.38%
MTCNN [9]	95.04%
Qi et.al. [30]	95.10%
LRNet [31]	95.10%
Proposed Method	96.35%

performance to a certain extent. The detection performance on the FDDB data set has been improved by 0.18%.

Fig 10 and Table 2 show the performance evaluation of our proposed RFE-MTCNN against the state-of-the-art methods MTCNN [9], Multiscale Cascade [26], LDCF+ [29], Hyperface [27], DP2MFD [28] on FDDB. We obtained its data from the FDDB official website and evaluated the performance of the model through the ROC curve. The horizontal axis of the ROC curve represents False Positive (FP), and the vertical axis represents True Positive Rate (TPR). Besides the ROC curve, another indicator AUC is used to illustrate the pros and cons of the model, which is defined as the area enclosed by the ROC curve and coordinates. From the ROC curves in Fig 10, our proposed RFE-MTCNN outperforms the conventional MTCNN and other state-of-the-art algorithms for face detection. Compared with the MTCNN [9], the TPR of our proposed method increases by 1.3% at 2000 false positives (96.35%).

TABLE 3. Detection performance comparison on wider face.

Method	Average Precision		
	Easy	Medium	Hard
Multiscale Cascade [26]	0.691	0.634	0.345
Faceness [32]	0.713	0.664	0.424
LDCF+ [29]	0.790	0.769	0.522
MTCNN [9]	0.851	0.820	0.607
Qi et.al. [30]	0.869	0.847	0.664
Proposed Method	0.874	0.841	0.673

TABLE 4. Detection performance comparison on AFW dataset.

Method	Average Precision
DPM [7]	97.21%
Multiscale Cascade [26]	97.29%
STN [33]	98.35%
Proposed Method	98.37%

TABLE 5. Detection performance comparison on PASCAL dataset.

Method	Average Precision
DPM [7]	90.29%
Multiscale Cascade [26]	92.40%
STN [33]	94.10%
Proposed Method	95.24%

On the WIDER FACE dataset, we compared the proposed new network architecture with other excellent networks. We use the Precision-Recall (P-R) graph to measure the performance of the model. The horizontal axis of the P-R graph represents the recall of the model, and the vertical axis represents the Precision of the model. The Average Precision represents the area enclosed by the P-R graph and the coordinate axis. The better the performance. WIDER FACE has three different subsets, namely EASY, MEDIUM, and HARD. We obtained the detection data of ACF [34], Multiscale Cascade CNN [26], Faceness [32], LDCF+ [29], and MTCNN [9] on the official website of WIDER FACE, and plotted P-R diagrams. As shown in Table 3, the model we proposed has been greatly improved on three different subsets, especially on the HARD subset. Therefore, it can be seen that the model has strong robustness and detection performance.The results are shown in Fig. 11.

On the AFW and PASCAL face dataset, we compared the proposed new network architecture with our baseline and other excellent networks. We use the PR graph to measure the performance of the model. It has a great performance improvement on these two datasets. The results are shown in Table 4 and Table 5.

Fig. 12 and Fig. 13 demonstrate some qualitative results on common face detection benchmarks, including AFW, FDDB, Wider Face, and PASCAL face. The experimental results show that the proposed method has good robustness in the real environment.

D. INFERENCE EFFICIENCY

As shown in Table 6, compared with 496k parameters in MTCNN, the proposed network structure parameters are

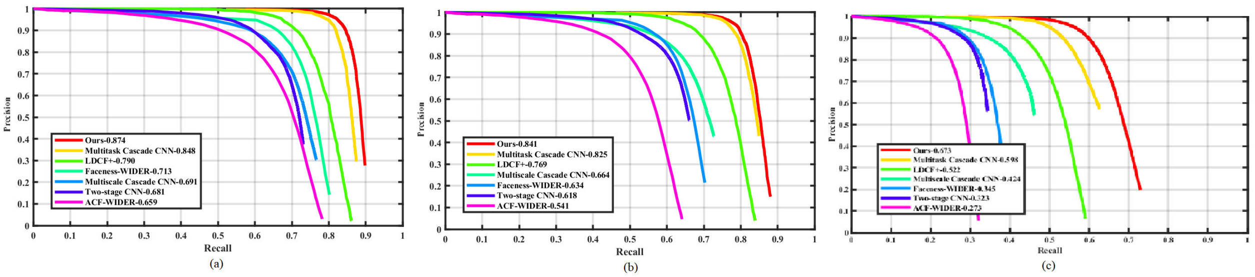

FIGURE 11. WIDER FACE Val: (a) Easy (b) Medium (c) Hard.

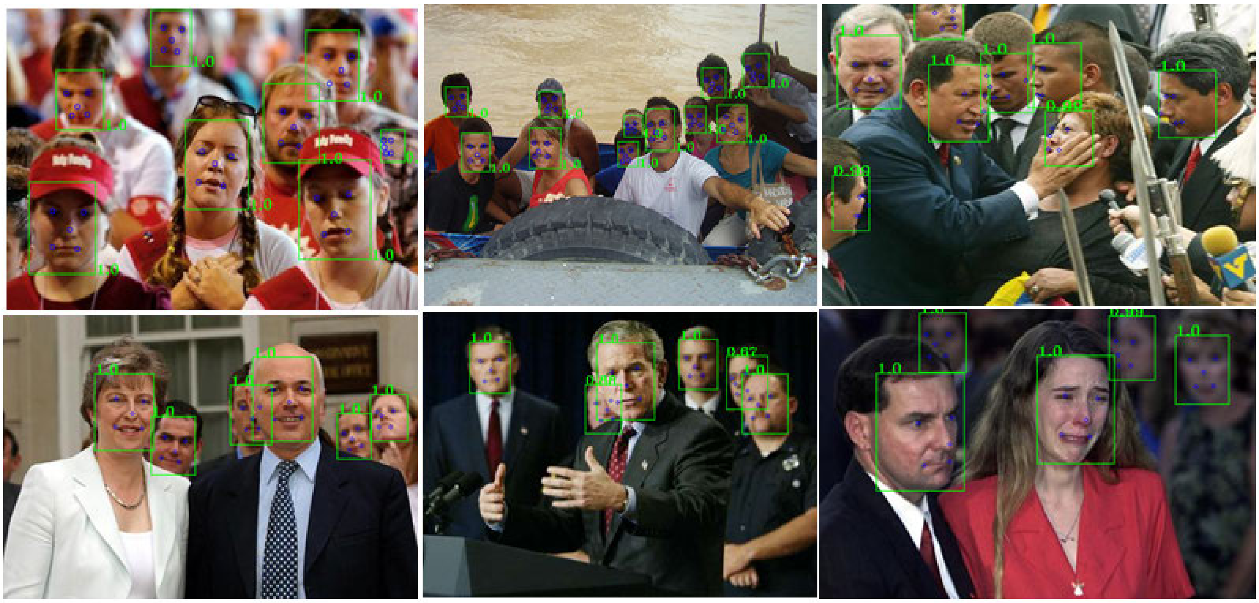

FIGURE 12. Qualitative results on AFW, PASCAL, and FDDB dataset.

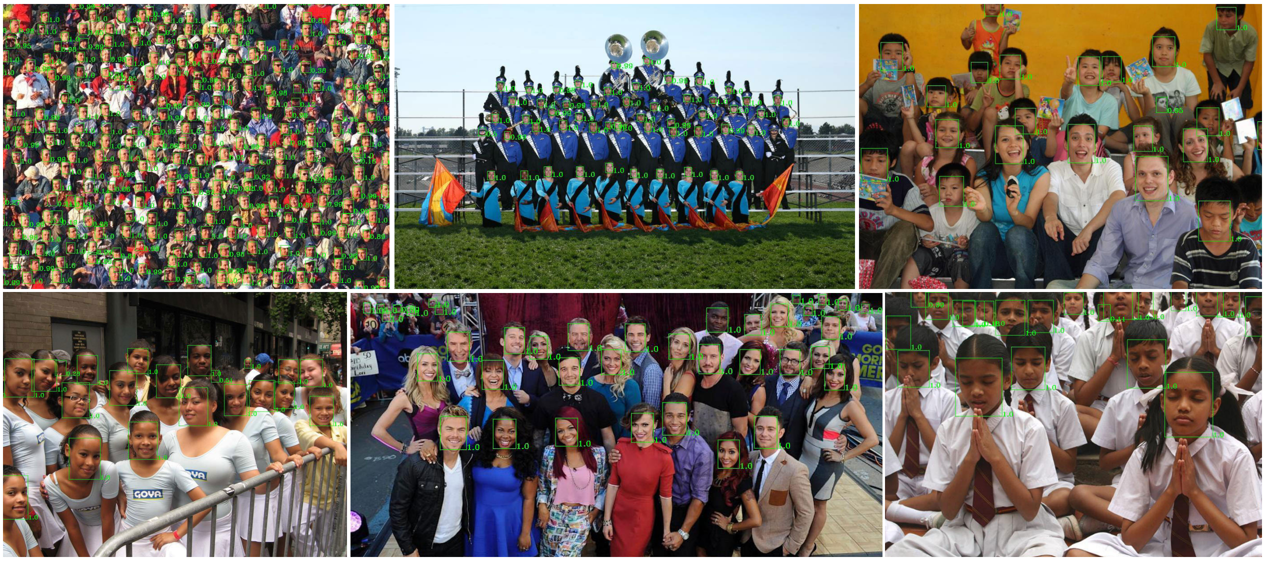

FIGURE 13. Qualitative results on wider face.

reduced by 78k, and the detection speed reaches 26 FPS on NVIDIA GTX 1070Ti. We use the Inception-V2 Block and RFB to increase the reception range of the network, and use a global average pool to replace the fully connected layer, which reduces the number of parameters and improves the detection performance of the network.

TABLE 6. Comparison Of network structure parameters.

Method	P-Net/k	R-Net/k	O-Net/k	P+R+O/k
MTCNN [9]	6.83	100.66	388.5	495.99
Qi et.al. [30]	6.79	108.28	398.9	513.97
Proposed Method	7.30	107.5	303.4	418.20

E. EXPERIMENT ENVIRONMENT

The experimental software environment is the operating system Ubuntu 18.04, CUDA 10.0, and cudnn 7.4. The deep learning framework is Tensorflow. The experimental hardware environment is Intel Core i7 8700K processor GPU for NVIDIA GTX 1070Ti.

V. CONCLUSION

In this paper, we propose a new face detection model RFE-MTCNN. According to the unique cascading characteristics of MTCNN, two different receptive field enhancement modules are used to optimize the network structure, and the AM-Softmax loss function is introduced to enhance the discriminability of the network. Experimental results show that, compared with other methods, this method has certain advantages, can improve the accuracy of face detection, and has fewer parameters.

REFERENCES

[1] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, ''ArcFace: Additive angular margin loss for deep face recognition,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4690–4699.

[2] D. Chen, C. Xu, J. Yang, J. Qian, Y. Zheng, and L. Shen, ''Joint Bayesian guided metric learning for end-to-end face verification,'' *Neurocomputing*, vol. 275, pp. 560–567, Jan. 2018.

[3] M. H. Khan, J. McDonagh, and G. Tzimiropoulos, ''Synergy between face alignment and tracking via discriminative global consensus optimization,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3811–3819.

[4] M. Drożdż and T. Kryjak, ''FPGA implementation of multi-scale face detection using HOG features and SVM classifier,'' *Image Process. Commun.*, vol. 21, no. 3, pp. 27–44, Sep. 2016.

[5] C. Ma, N. Trung, H. Uchiyama, H. Nagahara, A. Shimada, and R.-I. Taniguchi, ''Adapting local features for face detection in thermal image,'' *Sensors*, vol. 17, no. 12, p. 2741, Nov. 2017.

[6] T. Zhang, J. Li, W. Jia, J. Sun, and H. Yang, ''Fast and robust occluded face detection in ATM surveillance,'' *Pattern Recognit. Lett.*, vol. 107, pp. 33–40, May 2018.

[7] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool, ''Face detection without bells and whistles,'' in *Proc. Eur. Conf. Comput. Vis.* Springer, 2014, pp. 720–735.

[8] D. Marcetic and S. Ribaric, ''Deformable part-based robust face detection under occlusion by using face decomposition into face components,'' in *Proc. 39th Int. Conv. Inf. Commun. Technol., Electron. Microelectron. (MIPRO)*, May 2016, pp. 1365–1370.

[9] 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.

[10] S. Wu, M. Kan, Z. He, S. Shan, and X. Chen, ''Funnel-structured cascade for multi-view face detection with alignment-awareness,'' *Neurocomputing*, vol. 221, pp. 138–145, Jan. 2017.

[11] S. Wan, Z. Chen, T. Zhang, B. Zhang, and K.-k. Wong, ''Bootstrapping face detection with hard negative examples,'' 2016, *arXiv:1608.02236*. [Online]. Available: http://arxiv.org/abs/1608.02236

[12] K. Simonyan and A. Zisserman, ''Very deep convolutional networks for large-scale image recognition,'' 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

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

[14] 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 Retr. (ICMR)*, 2015, pp. 643–650.

[15] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, ''Rethinking the inception architecture for computer vision,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[16] S. Liu and D. Huang, ''Receptive field block net for accurate and fast object detection,'' in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 385–400.

[17] F. Wang, J. Cheng, W. Liu, and H. Liu, ''Additive margin softmax for face verification,'' *IEEE Signal Process. Lett.*, vol. 25, no. 7, pp. 926–930, Jul. 2018.

[18] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, ''Going deeper with convolutions,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[19] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cottrell, ''Understanding convolution for semantic segmentation,'' in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 1451–1460.

[20] M. Lin, Q. Chen, and S. Yan, ''Network in network,'' 2013, *arXiv:1312.4400*. [Online]. Available: http://arxiv.org/abs/1312.4400

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

[22] Z. Liu, P. Luo, X. Wang, and X. Tang, ''Deep learning face attributes in the wild,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3730–3738.

[23] V. Jain and E. Learned-Miller, ''Fddb: A benchmark for face detection in unconstrained settings,'' Univ. Massachusetts Amherst, Amherst, MA, USA, Tech. Rep., 2010.

[24] M. Kostinger, P. Wohlhart, P. M. Roth, and H. Bischof, ''Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization,'' in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCV Workshops)*, Nov. 2011, pp. 2144–2151.

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

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

[27] R. Ranjan, V. M. Patel, and R. Chellappa, ''Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 121–135, Jan. 2019.

[28] R. Ranjan, V. M. Patel, and R. Chellappa, ''A deep pyramid deformable part model for face detection,'' in *Proc. IEEE 7th Int. Conf. Biometrics Theory, Appl. Syst. (BTAS)*, Sep. 2015, pp. 1–8.

[29] E. Ohn-Bar and M. M. Trivedi, ''To boost or not to boost? On the limits of boosted trees for object detection,'' in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 3350–3355.

[30] R. Qi, R.-S. Jia, Q.-C. Mao, H.-M. Sun, and L.-Q. Zuo, ''Face detection method based on cascaded convolutional networks,'' *IEEE Access*, vol. 7, pp. 110740–110748, 2019.

[31] S. Hou, Y. Li, Y. Pan, X. Yang, and G. Yin, ''A face detection algorithm based on two information flow block and retinal receptive field block,'' *IEEE Access*, vol. 8, pp. 30682–30691, 2020.

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

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

[34] B. Yang, J. Yan, 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.

XIAOCHAO LI (Senior Member, IEEE) received the B.Sc. degree in electronic engineering from the Beijing Institute of Technology, China, in 1992, and the M.S. degree in electrical engineering and the Ph.D. degree in solid-state physics from Xiamen University, China, in 1995 and 2005, respectively. From 2005 to 2008, he was a Postdoctoral Fellow with Xidian University. Since 2005, he has been a Staff Member with the Electronic Engineering Department. From 2010 to 2011, he was a Visiting Scholar with North Carolina State University. From 2014 to 2016, he was a Visiting Fellow with the State Key Laboratory of Analog and Mixed-signal VLSI (AMSV), University of Macau. He is currently a Professor in electronic engineering and the Director of the Fujian Key Laboratory of Integrated Circuit Design and Measurement, Xiamen University. He is also the Head of the Electrical and Electronics Engineering Department with Xiamen University Malaysia. He has authored over 60 research articles, eight patents of invention, seven software or integrated circuit layout copyright, and one book *Mixed Signal CMOS Integrated Analog-to-Digital Convertor Design* (Tsinghua Press, 2015). His research interests include artificial intelligence, mixed signal integrated circuit design, parallel and distributed processing, and embedded systems. He is an IET Charted Engineer and enrolled with the Xiamen 200 Talents Program.

ZHENJIE YANG received the B.S. degree in electrical Engineering and automation from the Beijing Institute of Technology, Beijing, China, in 2018. He is currently pursuing the master's degree with the Department of Electronic Engineering, Xiamen University, Xiamen, Fujian, China. His current research interests include image processing and deep learning.

HONGWEI WU (Member, IEEE) received the B.S. degree in information engineering from the Nanjing University of Posts and Telecommunications, China, in 1999, and the M.S. degree in system engineering and the Ph.D. degree in circuits and systems from Xiamen University, China, in 2004 and 2013, respectively. Since 2014, he has been a Research Leader with the Xiamen Network Information Security Joint Laboratory. His research interests include cryptography and parallel and distributed processing.