Journal of Information Hiding and Multimedia Signal Processing ©2021 ISSN 2073-4212
Ubiquitous International Volume **12**, Number **4**, December **2021**

Occluded and tiny face detection network for dense crowd

Zhuo-Fan Xu

Beijing Jiaotong University
Haidian District, Beijing

Hui-Hui Bai

Beijing Jiaotong University
Haidian District, Beijing

Ji-Min Xiao

Xi'an Jiaotong-liverpool University
Suzhou, Jiangsu

Fei-Ran Jie

Luoyang Institute of Electro-optical Equipment
Luoyang, Henan

Yao Zhao

Beijing Jiaotong University
Haidian District, Beijing

Received June 2021; revised August 2021

Abstract. *Face detection, which palys a crucial part in face related applications, has been widely developed in the past decades. However, detecting occluded and tiny faces still remains great challenges due to two inherent problems: the excessive number of redundant smaples, and few effective features for detection. Most previous approachs cannot essentially filter out the redundant samples and fail to take full advantage of existing effective features. Therefore, the detection accuracy of occluded face and tiny face cannot be improved simultaneously. To overcome these problems, we put forward a occluded and tiny face detection network (OTFD), aiming to enhance the ability of detecting occluded and tiny faces simultaneously especially in dense crowd. OTFD consists of filter subnet, connection subnet and precise subnet in a parallel structure. Among them, the filter subnet is used to delete the excessive redundant samples and reduce the complexity of training process. The connection subnet is designed to achieve the fusion of features from shallow and deep layers, and the semantic information from deep layers are supplemented to shallow layers. In precise subnet, the anchors and features have been filtered and enriched in filter and connection subnets, which is helpful to make precise classification and regression in this subnet. Furthermore, in order to solve the problem of few effective features for detection, in filter subnet, we propose the feature compensation module (FCM) to reduce the inevitable feature loss caused by convolution operation, which makes the most of existing features in occluded and tiny faces. Last but not least, with the purpose of improving the anti-interference ability of detector in dense crowd, a repulsive loss function is presented to enhance the robustness of detector in complex backgrounds. Extensive experiments are conducted on wider face dataset, and our method achieves superior performance compared with other corresponding methods.*

Keywords: face detection, occluded tiny face, repulsive loss, feature compensation

1. **Introduction.** Face detection, which has been widely developed in the past decades, is a fundamental task in face applications such as face alignment, face verification and face recognition. The performance of face detection directly affect these subsequent tasks. Therefore, it is an essential and practical problem to improve the accuracy of face detection. As the pioneering work for face detection, Viola Jones *et al.*[1] focus on designing robust features and training effective classifiers, and then adopt AdaBoost algorithm with handcrafted features to detect faces. It is the representative traditional face detection method. Lu [2] proposes a novel algorithm for automatically detecting human faces in digital still color images under non-constrained scene conditions, such as the presence of a complex background and uncontrolled illuminations. Some methods also detect faces from the perspective of skin. Hu[3] presents a three-stage scheme for real-time reliable face detection. The proposed three-stage scheme is a feature-based method that is mainly based on skin color and facial features. Zhang[4] improves GLHS space and detects faces on the basis of it.

With the rapid development of deep learning, traditional hand crafted face features have been replaced by deeply learned features from convolutional neural network (CNN). Li *et al.*[5] propose CascadeCNN, a deep convolutional network implementation of the classical Viola Jones method, which plays a milestone role in the future CNN method. Since then, face detection methods based on CNN have flourished. CMS-RCNN[6], Face-RCNN[7], FDNet[8] are RCNN-based detectors which can achieve high detection accuracy but are time-consuming. Based on Region-based Fully Convolutional Networks (R-FCN), Face R-FCN[9] proposed in a fully convolutional fashion, is more accurate and computational efficient face detector compared with the previous R-CNN based face detectors. MTCNN[10], ScaleFace[11] and MSCNN[12] argue that different scale faces need different network structures to detect that achieve multi-scale face detection. STN[13] and Faceboxes[14] enable real-time face detection with high accuracy.

Although face detection has made big breakthrough with deep learning, the detection under occluded and tiny conditions still faces great challenges.

In order to solve tiny face detection, Finding tiny face[15], SSH[16] and PyramidBox[17] use context such as shoulders or heads to improve the detection accuracy of tiny face. S^{3}FD [18] and Faceboxes[14] increase the sampling ratio of tiny anchors and change the anchor settings to make the detection framework friendly to tiny face. Attention mechanism is applied in FAN[19] to improve the ability of detecting occluded face. Although many methods are devoted to enhancing detection accuracy of occluded or tiny face to some extent, the results are still unsatisfactory and these methods can not improve the performance of detecting tiny face and detecting occluded face at the same time. It can be found out that the main difficulties lie in two aspects: excessive number of redundant smaples, and few effective features for detection.

As for the first difficulty, the samples can be divided into two types: redundant and face-related samples. As shown in the FIGURE 1, the contents in the green anchors are houses and trees, whose features are essentially different from human face features, called redundant samples. The blue anchors are related to facial features, called face-related samples. It can be found out that some blue anchors are real faces while others only have the same context as the faces, such as the back of head. These face-related samples require further precise judgment. In practice, the number of redundant samples far exceeds the number of face-related samples in occluded and tiny face detection. Excessive redundant samples are harmful for precise classification. If classification and regression are conducted directly with such a large number of redundant samples, the accuracy of results will be greatly compromised. Therefore, these redundant samples need to be screened out before precise classification and regression.

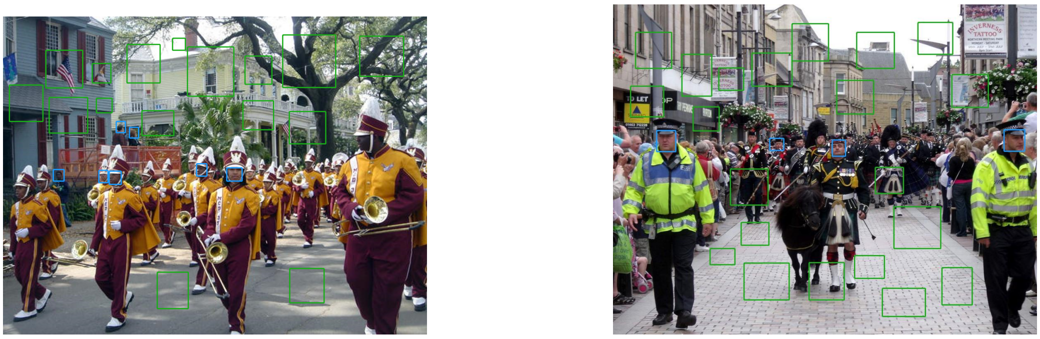

Figure 1. Some examples of redundant samples and face-related samples

2. **Related Work.** Recently, face detection is of great academic and application value and has been highly concerned. Viola Jones *et al.* [1] use Haar features and AdaBoost algorithm to train classifiers that achieve good accuracy with real-time efficiency, which is an early face detection method using traditional technology. Zhen *et al.* [21] come up with multi-view detection framework, in which the concept of channel feature is proposed and the image channel is extended to the histogram of gradient direction. CascadeCNN proposed by Li *et al.* [5] develops a cascade architecture based on CNNs with powerful discriminative capability and high performance. MTCNN[10] takes advantage of multi-task structure to jointly solve face detection and alignment. MS-CNN[12] uses different layers to detect faces of different scales that achieves multi-scale detection. CMS-RCNN[6] is on the basis of Faster R-CNN[22] with body contextual information as an assistant to improve the accuracy of face detection. Faceboxes[14], a lightweight and speed-fast face detector, comes up with rapidly digested convolutional layers to accelerate the speed of extracting the features for real-time detection. Besides, it presents multiple scale convolutional layers to achieve multi-scale detection.

As the general face detection technology is becoming more and more mature, many researchers find that occluded and tiny face detection still exits great obstacles. Some methods are devoted to improving the ability of detector in detecting occluded or tiny face. Finding tiny face[15] trains separate detectors for different scales and uses context such as shoulders to jointly judge if it is a face. SSH[16], which adopts the same idea as Hu *et al.*[15], add large filters on each prediction head to merge the context information and take use of different detection modules to detect face in various scales. Zhu *et al.*[23] propose expected max overlapping score to evaluate the quality of anchor matching. Besides, extra shifted anchors and stochastic face shifting are designed to improve the ability of detector. Faceness[24] trains a series of networks for facial attribute recognition to detect partially occluded faces. S^{3}FD[18] based on the architecture of SSD[25] utilizes the scale compensation anchor matching strategy to increase the recall rate of tiny face. And it comes up with a scale-equitable framework to deal with the problem that the existing face detection framework is friendly to detect medium and big scale face and not conducive to tiny face. Similar to this idea, Facesboxes[14] changes the settings of conventional anchor, increasing the sampling density of tiny scale anchor and forming a tiny scale friendly detection framework. As for occluded face, FAN[19], on the basis of RetinaNet[26], adds attention mechanism to pay more attention to the unoccluded parts to promote occluded face detection. P-SFD[20], this framework supports face and key points detection by using 3-order cascaded CNN architecture. DPSSD[28] proposes pyramid single shot face detector, which detects faces with large scale variations (especially tiny faces). It can well balance the relationship between detection accuracy and detection speed. CAHR[29] takes the influence of context into account to balance the image resolution and the spatial context range for the purposes of locating small faces. As for VGG16-SSH[30], the combination of SSH and VGG16 simplifies the network structure when considering the effectiveness of SSH. [31]compares the differences between yolo and mtcnn in face detection. It is used to compare the performance difference between YOLO5 and MTCNN in face detection, and it is found that the performance is much better than MTCNN, especially in tiny face detection. S^3FD[18] proposes maxout operation and compensation anchor strategy to improve tiny face detection. In maxout operation, it does not treat the background class as only one class but as multi-class, thus generates N candidate scores for background and selects the largest score from these candidate scores as the final background confidence. This operation enhances background discrimination and improves the reliability of the detection results. As for compensation anchor strategy, it divides the matching of the anchor into two stages. In the first stage, it follows conventional anchor matching method but decreases face threshold from 0.5 to 0.35 in order to increase the average number of matched anchors. In the second stage, it picks out anchors whose confidence are higher than 0.1, then sorts them to select top-N as matched anchors of this face. N is the number of matched faces in the first stage. It is because of these two operations that the detection of tiny faces is further enhanced to some extent. However, S^3FD is a single-track network from shallow to deep layers so that the features of shallow and deep features are unable to merged well, which hinders the performence of occluded and tiny face detection.

3. **Proposed Network.** In this section, we will introduce occluded and tiny face detection network (OTFD) in detail. We choose S^3FD with VGG16 as our backbone because of its simple structure and high efficiency. In this paper, S^3FD is modified into a parallel structure with the filter subnet, the connection subnet and the precise subnet as FIGURE 2 illustrated. Besides, several FCMs are inserted behind the three feature maps in the filter subnet, aiming to take advantage of existing features. Finally, a repulsive loss layer is presented to enhance the robustness of detectors in complex backgrounds.

3.1. Parallel network.

3.1.1. *Filter subnet.* In the filter subnet, we use the network of S^3FD with VGG16 as backbone for features extraction. In VGG16, the layers from conv11 to pool5 are remained

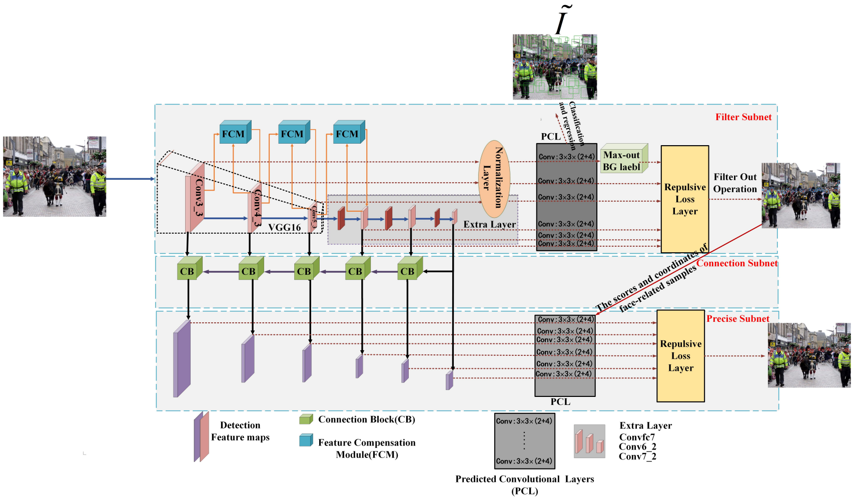

Figure 2. Architecture of our detector

and other layers are removed. Beisdes, it switches fc6 and fc7 from full connection layers into convolutional layers to decrease the number of parameters, then adds four extra convolutional layers behind them called extra layers. As illustrated in FIGURE 2, the size of feature maps in this subnet decreases progressively that contributes to mutil-scale detection. The conv3_3, conv4_3, conv5_3, convfc7, conv6_2 and conv7_2 are selected as the detection layers. The shallow layers such as conv3_3 and conv4_3 contain more location information, which are more beneficial to the tiny anchors filtration. Deep layers such as conv6_2 and conv7_2, which are rich in semantic information, are used to filter medium and large anchors. Followed by the conv3_3, conv4_3 and conv5_3 is normalization layer that is made up of L2-Norm operation to reduce the complexity of training. Then the features from these six detection layers will make classification in predicted convolutional layer (PCL) to judge if it is a redundant sample. The PCL is composed of several 3×3×(N+4) convolutional layers. 3×3 represents the size of the convolution kernel, and (N+4) represents the number of output channels in which N channels is for classification and 4 for regression. In PCL, N=4 is in the first convolution operation followed by Max-out BG label and N=2 for other operations. Here, N=4 means that four scores are generated for the current anchor, where three candidate scores are used to evaluate the anchor if it is a redundant sample that needs to be filtered out and one score is used to determine if it can be remained. Then, the largest of the three candidate scores is selected as the final score for the redundant sample judgement and the other two scores will be abandoned. Similarly, N=2 represents that only two scores are generated for the current anchor where one score is for remaining and one score for filtering.

After the PCL and maxout operation (only first convolution layer in PCL has maxout operation), the scores and coordinates of the anchors will be obtained. According to the scores, the anchors can be divided into redundant samples (green anchors) and face-related samples (blue anchors) by setting a threshold, as shown in the $\widetilde{I}$ of FIGURE 2. It can be seen that there are much more redundant samples than face-related samples. Since the features in these redundant samples are completely different from the human face, the large number of them will interfere with precise face classification. Therefore, these redundant samples need to be filtered out. And then the scores and coordinates of anchors are put into the repulsive loss layer, which includes the classification loss function and the regression loss function. After passing through this layer, there is a filter out operation to delete all these redundant samples which do not participate in the subsequent training and judgment. Besides, the features from conv3_3, conv4_3, conv5_3 will through feature compensation module (FCM) to reduce the unnecessary feature loss. The details of FCM and repulsive loss layer are introduced in Section3.2 and Section3.3. Through this subnet, redundant samples will be removed and face-related samples will be remained. It reduces the negative impacts from redundant samples and contributes to further accurate classification and regression in precise subnet.

3.1.2. *Connection subnet.* This subnet plays a major role in the fusion of the shallow and deep layers features. The features of shallow layers are sufficient in spatial information but short of semantic information. With each convolution operation from VGG16 to extra layer, the spatial information is gradually weakened and semantic information is gradually enhanced. In order to aggregate semantic information to the shallow layers features, the connection subnet uses five connection blocks (CB), whose structure is illustrated in FIGURE 3. The inputs of each CB are the shallow features from the filter subnet and the deep features from the next CB, except for the last CB which input features from filter subnet and conv7_2. When the shallow features input into CB, they will first pass through two 3×3 convolution layers and a relu layer. As for deep features, the size and channels

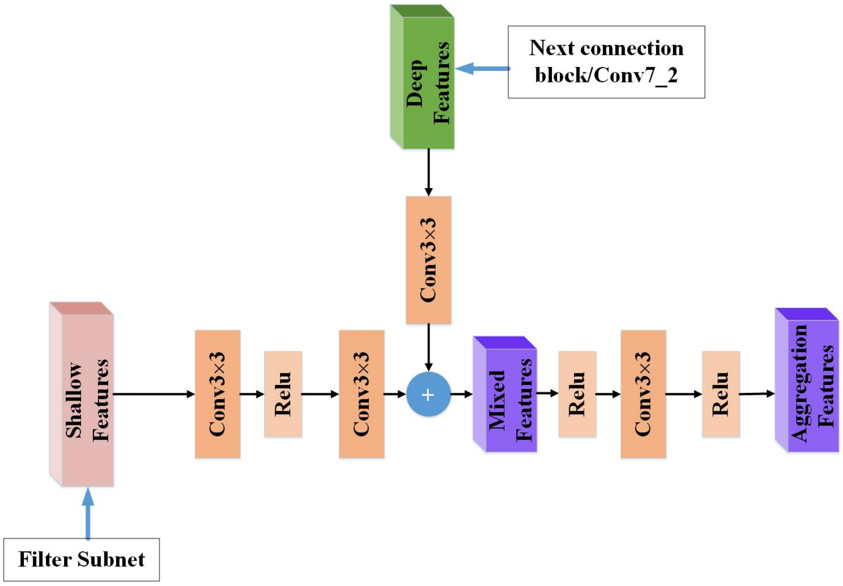

FIGURE 3. Connection block

will be resized by deconvolution operation. And then mixed features can be obtained by sum operation. Followed by a convolution layer and two relu layer, the aggregation features will be output. Through this block, the semantic information of the deep layers can be well integrated with the features of the shallow layers. In the connection subnet, from deep layers to shallow layers, the five CBs are arranged to enrich low level semantic information with high level features, which is helpful to occluded and tiny face detection.

3.1.3. *Precise subnet.* The main role of this subnet is to make final prediction to accurately find and locate the faces. It uses the face-related samples without redundant samples and six feature maps from filter and connection subnets. Occluded and tiny face detection mainly depends on shallow detection layers such as the first layer and the second layer. The features in these shallow layers are not only rich in the spatial information but also have the strong semantic information, which benefit from the connection subnet. The PCL in this subnet is composed of six 3×3×(2+4) convolution layers. The 3×3 represents the size of the convolution kernel and (2+4) stands for the numbers of output channels. 2 channels are for classification and 4 for regression. Through this layer, the scores and locations of faces can be obtained. PCL in this subnet further classifies the face-related samples and finds out exactly real faces and negative samples just similar to faces in context. Finally, the scores and coordinates will enter the repulsive loss layer to calculate the classification loss and regression loss. It is because the use of the face-related samples without redundant samples and information-rich features that realize the accurate classification and regression in this subnet.

3.2. **Feature Compensation Module.** Occluded and tiny faces contain few valid features. With convolution operation, the available features will be inevitably lost. In order to make full use of the available features, the task of this module is to achieve the lost features of current layer compared with previous layer and then compensate them to current layer. As illustrated in FIGURE 4, this module is composed of three main operations: upsample, sum, subtraction. The two inputs of this module are the features from the current layer and previous layer, respectively. Firstly, the current features can be upsampled to the same size as the previous features and then a subtraction operation is made between them. The features obtained are considered as the lost features caused by

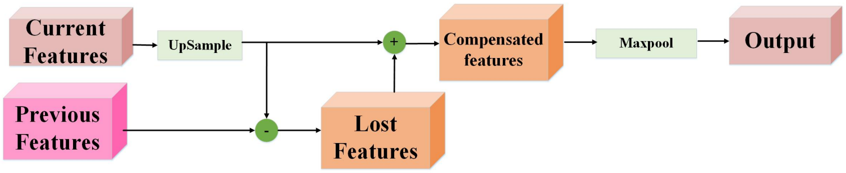

FIGURE 4. Feature compensation module

convolution. After that, the lost features are compensated to the current features by sum operation called compensated features. Finally, the size and channels of the compensated features are changed back to the size of input current features by maxpool operation. We adopt this module after conv3_3, conv4_3 and conv5_3 in the filter subnet to compensate the loss of features in the propagation process and make best use of existing features.

3.3. **Repulsive Loss Layer.** It is very challenging to detect occluded and tiny faces especially in dense crowd because there exists highly overlap between faces which disturb with each other. We solve this problem from the perspective of loss function. Here, the concept of repulsive term is put forward in this layer.

$$\begin{aligned} L = & \frac{1}{N_{cls}} \sum_i p_i^* (L_{attr}(t_i, t_i^*) + L_{repu}(t_i, t^*)) \\ & + \frac{\lambda}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*) \end{aligned} \tag{1}$$

The composition of the repulsive loss layer in filter and precise subnet is completely identical, so we only explain the precise subnet in detail. As shown in equation (1), the loss function can be divided into two parts, that is, regression and classification parts. We use the binary cross entropy loss function as the classification loss function, which is defined as L_{cls}. The parameter p_i stands for predicted probability that anchor i is a face. p_i^* is groundtruth label, and $p_i^* = 1$ when the anchor is positive, $p_i^* = 0$ otherwise. The classification loss function will eventually be normalized by N_{cls}, which is the number of positive anchors and λ is a balancing parameter. As for regression part, we further divide the regression function into attractive part L_{attr} and repulsive part L_{repu}. They are all made up of Smooth L1. t_i is the coordinates of predicted box, and t_i^* stands for the coordinates of groundtruth i. t^* is the coordinates of all groundtruths. For an anchor, the groundtruths can be divided into target-groundtruth and interference-groundtruths. Target-groundtruth is the groundtruth which has the largest overlap with this anchor, and other groundtruths are regarded as interference-groundtruths. L_{attr} is on behalf of the conventional concept of loss function. The variables in L_{attr} are the distance of predicted box and target-groundtruth. The smaller distance means the more accurate results of prediction. Therefore, the purpose of L_{attr} is to minimize the distance between them.

As for repulsive part L_{repu}, it plays an important role in the dense scenes especially when there exits highly overlap between faces. The variables in L_{repu} are the overlap value, that is, IOGs of the predicted box and each interference-groundtruths. The smaller IOG means the farther distance between them. L_{repu} makes interference-groundtruths be as far away from the predicted box as possible, which can reduce the disturbance of interference-groundtruths to the predicted box. We select IOG to measure the overlap between the predicted box and the interference-groundtruth. IOG are defined as equeation (2),

Algorithm 1 Divide the groundtruths into target groundtruth and interference-groundtruths($Array, GT, Repu, P, Map$)

```
input: All GroundTruths: $GT$; Prediction Box: $P$
output: The set of attractive terms: $Attr$; The set of repulsive terms: $Repu$; The map
    of IOGs between groundtruths and the current anchor: $Map$;
Initialize set: $Attr$, $Repu$; Initialize map: $Map$;
for $gt \in GT$ do
    $iog \leftarrow IOG(P, gt)$
    $Map.put\{gt : iog\}$;
end for
$Map \leftarrow sorted(Map.values)$
$targetgt \leftarrow Map[0]$
$Attr.append(targetgt)$
$Map.remove(targetgt)$
if $len(Map) == 0$ then
    return $Attr$
else if $len(Map) = 1$ then
    return $Repu.append(Map[0])$
else
    return $Repu.append(Map[0], Map[1])$
end if
```

$$IOG(B,G) = \frac{area(B \cap G)}{area(G)} \tag{2}$$

where B represents the predicted box and G represents the groundtruth. The $area(G)$ is the area of the groundtruth, and the $area(B \cap G)$ is the overlap area of predicted box and the groundtruth. Given $area(G)$, the IOG can be reduced by decreasing the $area(B \cap G)$.

In order to divide the groundtruths into target-groundtruth and interference-groundtruths, Algorithm 1 is designed. As explained in Algorithm 1, firstly, we calculate the IOGs between the predicted box and each groundtruths, and then arrange the groundtruths in descending order according to the IOGs. The groundtruth with the largest IOG is regarded as the target-groundtruth and put into Attr set to calculate L_{attr} with the anchor. Others are regarded as the interference-groundtruths, where the larger IOG value means the stronger interference to anchors. In order to simplify the calculation, if the number of interference-groundtruths are greater than two, we choose the two most disturbing groundtruths to put them into Repu set and calculate L_{repu}. With the adjustment by L_{repu}, IOG is getting smaller and smaller which means the predicted box and interference-groundtruths are getting farther and farther. L_{repu} is helpful to reduce the disturbance of the interference-groundtruths to anchors for classification and regression, and improve the anti-interference ability of detector especially in dense scene.

4. **Experiment.** In this section, we first compare the performance of our detector with other classical detectors such as S^{3}FD, SSH. Then, we conduct ablation experiments to analyze the effectiveness of our parallel network, feature compensation module and loss fucntion with repulsive term. Finally, the influence of the threshold in the filter subnet will be verified. We use the wider face dataset to evaluate the performance of our detector. The wider face dataset is a benchmark data set for face detection, including 32203 images and 393703 human faces. Among them, 158989 pictures are divided into train subset,

39496 in the validation subset and the rest images are divided into test subset. Each subset is divided into easy, medium and hard parts, which roughly correspond to difficulty of detection. In the easy part, most of the faces are relatively large in size, close to the camera, less occlusion and sufficient light. However, in hard part, the scale of faces is tiny, occluded, dense and fuzzy. Therefore, it is more difficult to improve the accuracy in hard part than in easy part. We chose the validation subset to do the ablation experiments and compare with other methods on the test subset.

TABLE 1. Detection results on wider face test subset

network	easy	medium	hard
ACF-WIDER[21]	64.2	52.6	25.2
LDCF+[27]	79.7	77.2	56.4
MTCNN[10]	84.8	82.5	59.8
CMS-RCNN[6]	90.2	87.4	64.3
ScaleFace[10]	86.7	86.6	76.4
MSCNN[12]	91.7	90.3	80.9
HR[15]	92.3	91.0	81.9
Face-RCNN[7]	93.2	91.6	82.7
S^{3}FD[18]	92.8	91.3	84.0
SSH[16]	92.7	91.5	84.4
FAN[19]	94.6	93.6	88.5
DPSSD[28]	92.5	90.8	85.7
CAHR[29]	92.1	90.1	83.2
P-SFD[20]	93.8	90.1	80.1
VGG16-SSH [30]	92.1	91.8	84.5
yolo-mtcnn[31]	—	—	85.7
our detector	94.7	93.8	86.1

4.1. **Comparison With Other Classic Networks.** TABLE 1 compares our detector with good performing methods including traditional and deep learning methods on the wider face test subset. Firstly, we compare our detector with the traditional classical methods ACF and LDCF+. The proposed OTFD performs much better than these two methods, mainly because of the deep learning techniques. And then the comparisons with the CNN methods are conducted. In order to verify our improvements, we carry out comparative experiments with our baseline: S^{3}FD. In the easy, medium and hard test subset, the mAP of our detector outperforms the S^{3}FD by 1.9, 2.5, 2.1 respectively. These improvements demonstrate the effectiveness of our detector. MTCNN and MSCNN are known for multi-scale detection. From TABLE 1 ,it can be found out that these two methods are worse than the OTFD by 9.9, 11.3, 26.3 and 3.0, 3.5, 5.2, respectively. OTFD also outperforms SSH by 2.0, 2.3, 1.7, since SSH may not makes full use of facial features. HR, known for its ability in detecting tiny face, achieves 92.3 91.0 81.9 mAP, which is less than our detector. More importantly, OTFD is more accurate than the two stage network CMS-RCNN and Face-RCNN by 4.5, 6.4, 21.8 and 1.5, 2.2, 3.4.

4.2. **Ablation study.** To understand our detector better, we conduct ablation experiments to verify the effectiveness of the proposed modules. The baseline for all the ablation experiments is S^{3}FD. We use five different settings to verify our method.

(1) S^{3}FD (R): This setting changes the loss function and turns it into a loss function with the repulsive term, and other settings and network structure are consistent with S^{3}FD.

(2) S^{3}FD (P): It changes S^{3}FD to parallel structure without feature compensation module and repulsive term in loss function.

(3) S^{3}FD (FCM): This experiment only applies the feature compensation module in S^{3}FD.

Table 2. Detection results on wider face validation subset

network	easy	medium	hard
S^{3}FD	93.7	92.4	85.2
S^{3}FD(R)	94.5	93.4	85.5
S^{3}FD(P)	94.2	93.4	85.3
S^{3}FD(FCM)	94.1	93.2	85.7
S^{3}FD(R+P)	94.6	93.8	85.9
S^{3}FD(R+P+FCM)	94.8	94.0	86.0

(4) S^{3}FD (P+R): It changes S^{3}FD to parallel structure and adds repulsive term to loss function.

(5) S^{3}FD (R+P+FCM): It is our complete detector OTFD that contains the idea of parallel structure, feature compensation module and repulsive term on the basis of S^{3}FD.

As listed in TABLE 2, repulsive loss term is effecient compared with the original version of S^{3}FD listed in the first line. Just adding a repulsive term in loss function can make mAP of our detector higher than that of S^{3}FD by 0.8, 1.0, 0.3 respectively. This is due to the repulsive loss function which minimizes the interference from surrounding faces.

From the comparison between the S^{3}FD (P) and S^{3}FD, it can be seen that the mAP can be increased by 0.5, 1, 0.1. The reason for the growth of performance mainly comes from that the redundant samples are filtered out in filter subnet so that precise subnet can concentrate on face-related samples.

From S^{3}FD (FCM), the feature compensation module is also vaild. It can make our detector achieve AP scores of 94.1, 93.2, 85.7. This module makes up for the lost features caused in the propagation process and takes advantage of existing features adequately.

The comparison between the S^{3}FD and S^{3}FD (P+R) in table 2 indicates that the mixture of parallel and replusive term also makes the whole detector more effective than adding them respectively, which can achieve the mAP of 94.6, 93.8, 85.9. This setting can not only deletes the superfluous redundant samples, but also makes the predicted box unaffected by other surrounding faces.

In the last experiment, the parallel structure, feature compensation module, repulsive term in loss function are all applied together which can achieve the mAP of 94.8, 94.0, 86.0. The detection results on the validation subset is shown in FIGURE 5. Compared with S^{3}FD, our detector increases 1.1, 1.6, 0.8 of mAP in the easy, medium, hard part. Through experiments, we find that it is slightly easier to achieve growth in easy and medium than in hard part. Our detector not only has the ability of anti-interference and deletes redundant samples, but also has the ability to make existing features be fully utilized by adding the feature compensation module.

4.3. **The threshold in the filter subnet.**

This section discusses the impact of threshold in the filter subnet on the wider face validation subset. As listed in TABLE 3, we set four thresholds: 0.05, 0.1, 0.15 and 0.2. Other settings keep the same. The effect is the worst

Table 3. Varing threshold in the filter subnet

thresold	easy	medium	hard
0.05	86.9	85	76
0.1	94.8	94.0	86.0
0.15	94.7	93.8	85.9
0.2	91.7	91.3	85.6

when the threshold is set to 0.05, mainly due to the fact that the redundant samples are not screened thoroughly. However, when the threshold is set to 0.2, the mAP decreases compared to 0.15 and 0.1. The main reason is that a lot of potentially face-related samples are filtered out. As for 0.1 and 0.15, the threshold of 0.1 is the optimal and we choose 0.1 as the threshold of the filter subnet.

5. **Conclusion.** We propose a network OTFD which changes S^{3}FD to a parallel network to filter out the excessive number of redundant smaples. In addition, feature compensation module is proposed to compensate the inevitable loss caused by convolution. This module can make best of features which is helpful to solve the problem of few effective features in occluded and tiny faces. Last but not least, we improve the loss function by adding an repulsive term to the conventional loss function. It can enhance the robustness of the detector especially in dense scenes. The experiments demonstrate that these ideas we put forward are valid and achieve greater performance on wider face than S^{3}FD and other classical methods.

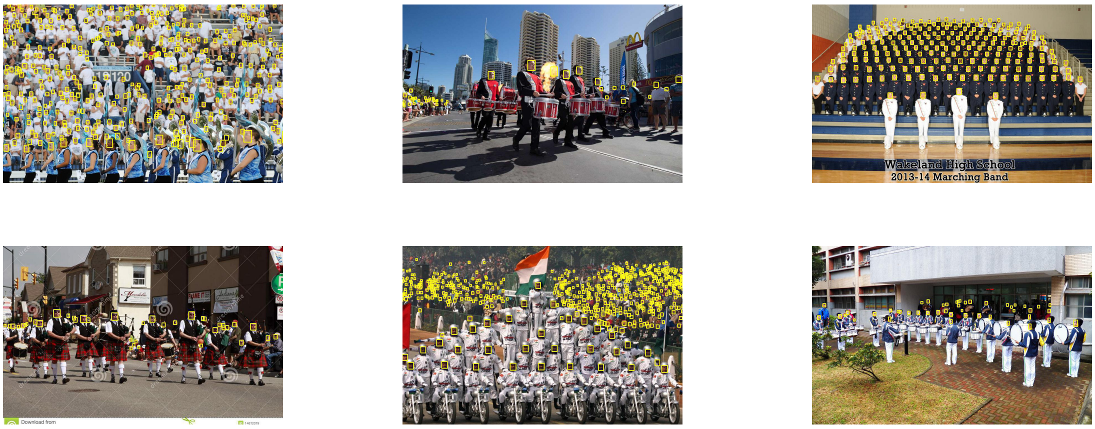

Figure 5. Detection Results

Acknowledgement. This work was supported in part by National Natural Science Foundation of China (No.61972023) and Fundamental Research Funds for the Central Universities(2019JBZ102).

REFERENCES

[1] P. Viola, M. J. Jones, Robust real-time face detection, *International Journal of Computer Vision*, vol.57, no.2, pp.137–154, 2004.

[2] Z. Lu, X. Xu and J. Pan, Face Detection Based on Vector Quantization in Color Images, International Journal of Innovative Computing, Information and Control, vol. 2, no. 3, pp. 667-672, 2006.

[3] W. Hu, C. Yang, D. Huang, Feature-based Face Detection Against Skin-color Like Backgrounds with Varying Illumination, *Journal of Information Hiding and Multimedia Signal Processing*, Vol.2, no. 2, pp. 123-132, 2011.

[4] Z. Zhang, M. Wang, A Skin Color Model Based on Modified GLHS Space for Face Detection, *Journal of Information Hiding and Multimedia Signal Processing*, Vol. 5, no. 2, pp. 144-151, 2014.

[5] H. Li, L. Zhe, X. Shen, A convolutional neural network cascade for face detection, *Computer Vision Pattern Recognition*, Boston, 2015.

[6] C. Zhu, Y. Zheng, K. Luu, Cms-rcnn: contextual multi-scale region-based cnn for unconstrained face detection, https://arxiv.org/pdf/1606.05413.pdf.

[7] H. Wang, Z. Li, X.Ji, Face R-CNN, *2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, Salt Lake City, Hawaii, 2017.

[8] S. Hoi, X. Sun, P. Wu, Face detection using deep learning: an improved faster rcnn approach, *Neurocomputing*, vol.299, no.19, pp.42–50, 2018.

[9] Z. Zheng, Y. Wang, J. Xing, Detecting faces using region-based fully convolutional networks. https://128.84.21.199/pdf/1709.05256.pdf

[10] K. Zhang, Z. Zhang, Z. Li, Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, vol.23, no.10, pp.1499-1503, 2016.

[11] S. Yang, Y. Xiong, C. L. Chen. Face detection through scalefriendly deep convolutional networks. *IEEE Signal Processing Letters*, vol.23, no.10, pp.1239-1251, 2016.

[12] Z. Cai, Q. Fan, R. S. Fe. A unified multi-scale deep convolutional neural network for fast object detection. *European Conference on Computer Vision*, Amsterdam, pp.329–342, 2016.

[13] D. Chen, G. Hua, F. Wen. Supervised transformer network for efficient face detection. *European Conference on Computer Vision)*, Amsterdam, pp.432–442, 2016.

[14] S. Zhang, Zhu X, Lei Z, Faceboxes: a cpu real-time face detector with high accuracy. *International Joint Conference on Biometrics*, Denver, 2017.

[15] P. Hu, D. Ramanan. Finding tiny faces. *2017 IEEE Conference on Computer Vision and Pattern Recognition*, Hawaii, 2017.

[16] M. Najibi , P. Samangouei, R. Chellappa , SSH: Single stage headless face detector. *2017 IEEE International Conference on Computer Vision*, Venice, 2017.

[17] X. Tang, D. K. Du, Z. He, Pyramidbox: a context assisted single shot face detector. *European Conference on Computer Vision*, Venice, 2017.

[18] S. Zhang , X. Zhu, Z. Lei, S3FD: Single shot scale-invariant face detector. *2017 IEEE International Conference on Computer Vision*, Venice, 2017.

[19] J. Wang , Y. Yuan, G. Yu, Face attention network: an effective face detector for the occluded faces. *2017 IEEE Conference on Computer Vision and Pattern Recognition*, Hawaii, 2017.

[20] H. Xu, Z. Wu, J. Ding, B. Li, FPGA Based Real-Time Multi-Face Detection System With Convolution Neural Network. *2019 8th International Symposium on Next Generation Electronics* .

[21] L. Zhen, B. Yang, J. Yan, Aggregate channel features for multi-view face detection. *2017 IEEE International Conference on Computer Vision*, pp.2999–3007,2014.

[22] S. Ren, K. He, R. Girshick, Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol.39, no.6, pp.1137–1149,2017.

[23] C. Zhu, R. Tao, K. Luu, Seeing small faces from robust anchors perspective. *2018 IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 2018.

[24] S. Yang, P. Luo, C. C. Loy, Faceness-net: Face detection through deep facial part responses. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, pp.1–1, 2017.

[25] D. Erhan, W. Liu, D.Anguelov, SSD: single shot multibox detector. *European Conference on Computer Vision*, Amsterdam, 2016.

[26] R. Girshick, T. Y. Lin, P. Goyal, Focal loss for dense object detection. *2017 IEEE International Conference on Computer Vision*,vol.2, no.4, pp.2999–3007, 2017.

[27] M. M. Trivedi, B. Ohn, To boost or not to boost? on the limits of boosted trees for object detection. *International Conference on Pattern Recognition* , Beijing, 2017.

[28] R. Ranjan, A. Bansal, J. Zheng, H. Xu, A fast and accurate system for face detection, identification, and verification. *IEEE Transactions on Biometrics Behavior & Identity Science*, vol.5, no.2, pp.1–1, 2019.

[29] T. Wu, D. Liang, J. Pan, Context-Anchors for Hybrid Resolution Face Detection. *2019 IEEE International Conference on Image Processing* , 2019.

[30] L. L. Zhu, F. C. Chen, Improvement of Face Detection Algorithm Based on Lightweight Convolutional Neural Network, *2020 IEEE International Conference on Image Processing* , 2020.

[31] N. Zhang, J. Min. L, Research on Face Detection Technology Based on MTCNN, *2020 IEEE International Conference on Image Processing* , 2020.