Computing and Informatics, Vol. 45, 2026, 182–202, doi: 10.31577/cai_2026_1_182

CF-YOLO: TOWARDS HIGHLY EFFECTIVE SMALL FACE DETECTION IN CROWDED SCENES

Hongbo Huang, Longfei Xu, Xiaoxu Yan, Linkai Huang

Computer School, Beijing Information Science and Technology University
Beijing, 100000, China
e-mail: hhb@bistu.edu.cn

Abstract. To address the issue of low recall rates in detecting small faces within crowded scenes, this paper conducts an analysis of the primary reasons behind this challenge and introduces a real-time face detection system named CF-YOLO (Crowded-Face-YOLO). The study identifies a crucial factor contributing to this problem, which is the insufficient provision of positive samples for small faces during the training phase by conventional face detectors. To tackle this limitation, a Sa-SimOTA strategy is proposed to enhance the availability of positive samples for small targets. Additionally, in the post-processing stage, the utilization of the non-maximum suppression (NMS) algorithm for assigning optimal bounding boxes to detected faces is discussed. The traditional fixed threshold employed in the NMS algorithm for decision-making often results in the loss of small face detection boxes in crowded scenarios. To alleviate this issue, a Soft-Face-NMS algorithm is introduced, which incorporates facial feature variables into the Soft-NMS algorithm for weighted processing, facilitating the selection of face boxes with higher confidence in overlapping regions. Furthermore, to augment the feature extraction capabilities of the YOLO backbone, an EMA+ attention module is proposed, and modifications are made to the network structure of YOLOv7 to enhance the extraction of more effective features conducive to small face detection. The proposed model demonstrates impressive accuracy rates of 97.3 %, 96.4 %, and 92.8 % on the easy, medium, and hard subsets of the Wider-Face dataset, respectively. Notably, the accuracy achieved on the hard subset approaches the state-of-the-art level, which further demonstrates the effectiveness of our proposed approach for face detection in crowded scenes.

Keywords: Face detection, label assignment, NMS, EMA+

1 INTRODUCTION

Face detection is a long-term computer vision task enabling applications like biometric identity verification and personal emotion analysis. In recent years, deep-learning-based object detection methods have seen convincing rapid advancements. These breakthrough developments have been adapted to face detection, improving results under varying poses, lighting, scales, and occlusions. Nevertheless, detecting small faces in crowded real-world environments remains a tough challenge. Although much literature has analyzed these issues [1, 2], there are still some unresolved problems that can be summarized as follows.

1. For small faces, object detectors often fail to provide enough positive samples, resulting in insufficient learning and low efficiency of the network parameters for small faces.
2. During the process of the network backbone with a common architecture, the pooling operation reduces the resolution of the feature map stage-by-stage, which exacerbates the loss of information for small faces. It may even cause small facial features to completely disappear from the final feature map under some extreme cases.
3. The traditional NMS algorithm used in the post-processing stage takes a fixed threshold for hard decision, which may lead to loss of small target detection boxes in the case of occlusion, resulting in a decrease in recall.

To address the aforementioned issues, this paper conducts intensive research and proposes three corresponding solutions inspired by several recently presented methods.

Firstly, traditional detectors adopt predefined rules to distinguish positive and negative samples. For instance, RetinaNet [3] and YOLOv3 [4] take IoU as the threshold criterion. OTA [5] proposes Optimal Transport Assignment, which views the label assignment procedure as an Optimal Transport (OT) problem. Based on the OTA algorithm, SimOTA [6] proposed in YOLOx [6] simplifies the iteration by an approximate dynamic strategy. They first calculate the IoU of the target box and prediction box, then add up the largest top k IoUs to obtain the number of positive targets. However, RFLA [7] argues that small targets usually have no overlap with predicted boxes. After rounding the sum of IoUs, the resulting number of positive samples for small targets may be relatively small, which is not conducive to the detection of small targets. In this paper, we introduce a compensation strategy for small targets and propose the Sa-SimOTA (Scale-aware SimOTA) algorithm to solve this problem.

Secondly, in order to enhance the sensitivity of the model for small face prediction and improve the ability of the model's feature extraction, we design a new network based on YOLOv7 [8]. Although the YOLO series has recently been updated to YOLOv10 [9], YOLOv7 still has advantages in terms of training simplicity and convenience.YOLOv7 model consists of three parts: the backbone, the neck

and three detection heads. The backbone network consists of several BConv layers, E-ELAN layers and MPConv layers, while the neck is composed of SPPCSPC layer, multiple BConv layers, multiple MPConv layers and Rep block layers. YOLOv7 outputs three prediction heads, called P3, P4 and P5 with resolution of 8, 16 and 32, respectively. In this paper, we remove P5 prediction head with a resolution of 32 and add P2 detection layer with a size of 4 to make the network make full use of small face features. In addition, to fuse and aggregate more levels of features without increasing the cost too much, we substitute the original PAN-FPN [10, 11] structure in YOLOv7 by an improved BiFPN [12] structure, which adds two additional paths between the original input and output nodes of the backbone network. Moreover, we propose an EMA+ attention module based on the EMA [13] module and add it to the backbone network, thereby further enhancing the model's ability of feature extraction.

Thirdly, for most detection tasks, NMS (Non-Maximum Suppression) is a common operation that serves as a post-processing method for removing redundant bounding boxes of the detected objects. However, original NMS may cause the inaccurate removal of small face candidates and result in a low recall rate. Soft-NMS [14] algorithm replaces the original confidence score with a slightly lower value, rather than directly setting it to zero, which can greatly improve recall in crowded scenes. In this paper, by extending Soft-NMS, we propose the Soft-Face Non-Maximum Suppression algorithm to improve the performance of the traditional NMS method for small face detection in crowded scenes. Considering the special characteristics of the human face, we also add a new criterion to constrain the aspect ratio of the detected face to be around the value of 1.2, which we show can evidently improve the face detection performance without noticeable computational complexity increment.

Overall, in order to improve the accuracy of small face detection under crowded scenes, this paper analyzed the main problems and challenges encountered in the process of small face detection, and improved the face detection performance from multiple aspects based on YOLOv7. This proposed method has significant performance improvement for face detection in crowded scenes, and we call it CF-YOLO (Crowded Face detector by YOLO). We have conducted comprehensive experiments to verify the effectiveness of our method. The main contributions of this paper can be summarized as follows:

- We propose a Sa-SimOTA method by introducing a compensation strategy for label assignment of positive samples which is especially effective to model training with small face detection.
- We improve the backbone of the YOLO model and design the EMA+ attention module to extract more effective features beneficial for small face detection.
- We introduce the SF-NMS post-processing method, which uses a soft decision strategy and adds new criteria to constrain the aspect ratio to be more suitable for human face.

- Our proposed model attains accuracy rates of 97.3%, 96.4 %, and 92.8 % for the easy, medium, and hard subsets of the Wider-Face dataset, respectively. The accuracy on the hard subset approaches the state-of-the-art level.

2 RELATED WORK

Recently, numerous face detectors have been proposed to improve the performance of face detection. In this section, we mainly review relevant works from three perspectives: anchor design, label assignment and NMS post-preprocessing.

2.1 Anchor Design

Anchor design is quite crucial for face detection. A reasonable anchor design algorithm can provide sufficient training samples for the model. FaceBox [15] proposed the Anchor Densification Strategy in which the anchors with insufficient density will be augmented to ensure that the anchor density is balanced. S^3FD [16] proposed a scale-equitable face detection framework, which predefines anchors at different feature map and sets the scale to be within 4–128 pixels. DSFD [17] algorithm optimized the data initialization process of anchor matching by improving anchor assignment, which can enhance the model's convergence speed and detection performance during training. Inspired by the work mentioned above, this paper explores the new anchor design strategy to corporate small face detection under crowded scenes.

2.2 Label Assignment

Label assignment tries to provide as many candidate bounding boxes as possible that are consistent with the ground truth target. S^3FD proposed a scale compensation anchor matching strategy that increases the matching anchor for faces by setting lower IoU threshold. HAMBox [18] emphasized that some unmatched anchors also have strong regression ability and proposed an online high-quality anchor compensation strategy to help match abnormal faces with high-quality anchors. MogFace [19] argues that combining offline and online information can provide more accurate optimization direction and proposes the adaptive online incremental anchor mining strategy to help compensate high-quality anchors for outer ground-truths. Although these methods have achieved meaningful results, they are not specifically designed for small face detection and still fall short in detecting small-scale faces under crowded scenes. It is still necessary to find suitable strategies for assigning small face labels to provide sufficient samples for model training.

2.3 Non-Maximum Suppression

For most detection tasks, NMS (Non-Maximum Suppression) is a necessary step that serves as a post-processing method for removing redundant bounding boxes of

the detected objects. In Soft-NMS, a confidence score is set according to a monotonic function of IoU, and the predicted boxes with lower confidence scores are not directly removed from the sorted list, but are assigned lower scores to avoid some candidate boxes being forcibly deleted. DIoU-NMS [20] is another improvement to the NMS processing. In addition to the IoU (Intersection over Union) ratio criteria, It introduces the concept of DIoU (Distance-IoU) constraint, considering the distance between detection boxes and the ground truth boxes. Adaptive-NMS [21] applied a dynamic suppression strategy by designing a density-subnet network to predict the density and sparsity of the target's surroundings so that the threshold may increases or decreases accordingly with the density of the target's surroundings. Inspired by these works, this paper further considers the approximate range of facial image proportions and introduces it into the constraints of NMS to further improve the accuracy of face detection.

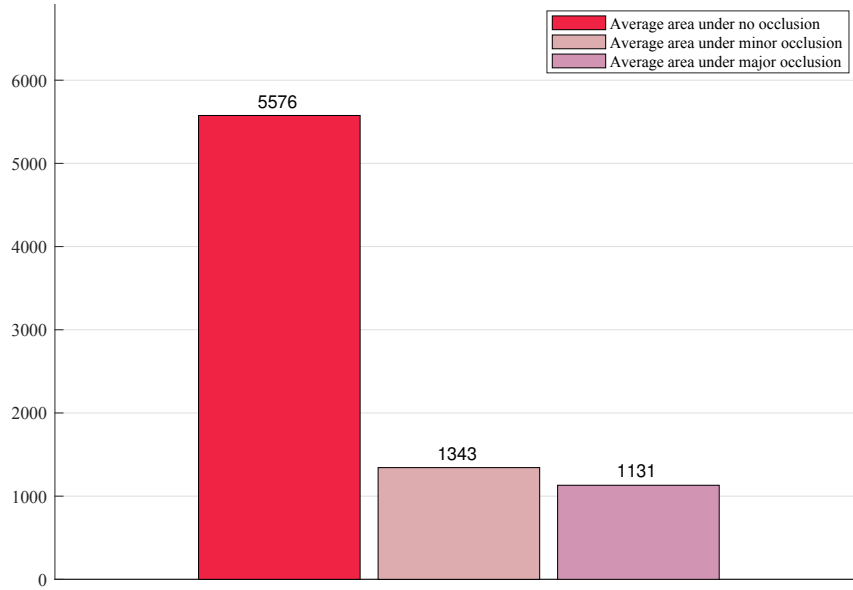
3 METHOD

To improve the performance of small face detection in crowded scences, we enhance the detector by the followig three methods: Sa-SimOTA, SF-NMS and new Network Architecture. These methods will be explained in detail respectively in the following sections.

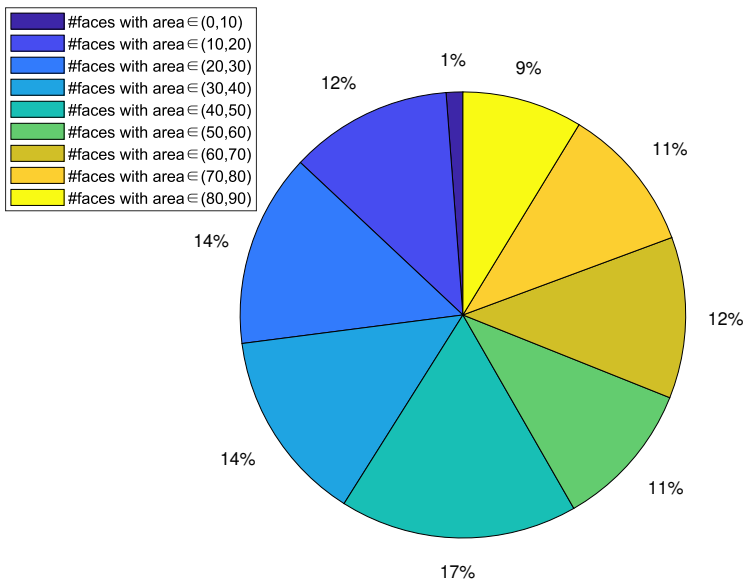
3.1 Sa-SimOTA

A statistical analysis of faces in the WiderFace [22] dataset reveals that in crowded scenes, smaller faces are more prevalent (Figure 1 a)). Therefore, when considering improving the performance of the model in crowded scenes, the ability of the model to detect small targets should also be improved. Figure 1 b) demonstrates that faces under major occlusion with area between 50 and 60 are the most numerous. Therefore, 60 is selected as the threshold for area of small targets in this paper. Besides, during the training process, we pay attention to the average number of positive samples for small targets and medium-large targets within the first five training epoches. Additionally, the training process focuses on the average number of positive samples for small and medium-large targets, demonstrating a significant disparity with small targets having only 1.03 positive samples compared to 2.26 for medium-large targets. This significant discrepancy highlights the shortcomings of current detection algorithms in the realm of small target detection.

To tackle the challenge of insufficient positive samples for small targets, we introduce the Sa-SimOTA algorithm, which builds upon the SimOTA algorithm by fine-tuning the estimation of its internal dynamic k-value. The method entails calculating the weighted sum of the maximum 10 Intersection over Union (IoU) values between every target box and its corresponding predictions. Subsequently, we devise a succinct face area reduction function to determine the adjustment coefficient, which is then multiplied by the computed initial sum (Figure 3). This



a) Average area of faces under different degrees of occlusion



b) Occupancy of different areas of faces under major occlusion

Figure 1. Face sizes analysis of WiderFace dataset. a) The average area of faces under major occlusion is relatively the smallest which is 1 131. b) Faces under major occlusion with areas between 50 and 60 are the most numerous, accounting for 17 % of the total.

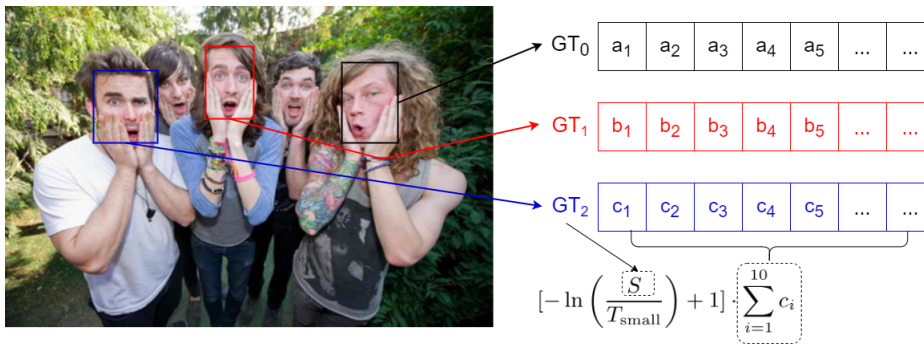


Figure 2. Schematic diagram of the Sa-Simota algorithm, where S is the area of the target box and T_{small} denotes the threshold of small faces

novel enhancement seeks to augment the number of positive samples for small targets, thereby markedly enhancing the overall face detection performance in cluttered environments. The formula for computing λ is as follows:

$$\lambda = -\ln\left(\frac{S}{T_{small}}\right) + 1, \tag{1}$$

where S is the area of the target box and T_{small} denotes the threshold of small faces. Therefore, the smaller the area of the target box, the larger the value of λ, and the more compensating positive samples will be obtained. After that, the average number of positive samples for small targets can be increased to 1.67.

3.2 SF-NMS

In this work, we introduce a SF-NMS algorithm based on the Soft-NMS algorithm. Soft NMS enhances the traditional NMS algorithm by incorporating the Intersection over Union (IoU) ratio to modulate confidence scores, thereby avoiding the direct assignment of zero values as practiced in the conventional NMS approach. The algorithmic steps of Soft NMS are delineated as follows:

$$S_i = \begin{cases} S_i * (1 - IoU(M, b_i)), & IoU(M, b_i) \geq N_t, \\ S_i, & IoU(M, b_i) < N_t, \end{cases} \tag{2}$$

where S_i refers to the confidence score of the predicted box b_i, N_t refers to the IoU threshold. IoU between M and b_i is used to form a linear function to adjust the confidence score of b_i. Hence, those predicted boxes which are very close would be assigned a penalty rather than being deleted directly. Furthermore, considering the specificity of human faces, the aspect ratio of human face is approximately 1.2. Therefore, we propose to further delete predicted boxes with low possibilities being

faces by using the aspect ratio. We suggest using the following rules to update the pruning steps:

$$S_i = \begin{cases} S_i * (1 - W\_ratio * IoU(M, b_i)), & IoU(M, b_i) \geq N_t, \\ S_i, & IoU(M, b_i) < N_t, \end{cases} \tag{3}$$

$$W\_ratio = \exp\left(\frac{\text{abs}(ratio - 1.2)}{\max(1.2, ratio)}\right), \tag{4}$$

where W_ratio defines as the disparity between the aspect ratio of the predicted box b_i and the standard aspect ratio of faces (we set as 1.2 in this paper). The W_ratio is calculated as Equation (4), where *ratio* refers to the aspect ratio of predicted box b_i. The coefficient W_ratio is used to measure the degree of conformity of the aspect ratio of the face box b_i. If the aspect ratio of one predicted box deviates from 1.2 greatly, it indicates that the predicted box is less likely to be a face, so the value ratio weight will be relatively large, thus the confidence score of this predicted box is further reduced (see Algorithm 1).

Algorithm 1 Soft-Face Non-Maximum Suppression (SF-NMS)

Input: Bounding boxes $B = \{b_1, b_2, \ldots, b_n\}$
Output: Selected bounding boxes S
Initialize an empty list S
while $B \neq \emptyset$ **do**
  Select the bounding box b_i with the highest confidence score
  Add b_i to list S
  for bounding box b_j in B **do**
    # Calculate Intersection over Union *iou* between b_i and b_j
    $iou = \frac{Area(b_i \cap b_j)}{Area(b_i \cup b_j)}$
    # Calculate r
    $r = \exp\left(\frac{absratio - 1.2}{\max(1.2, ratio)}\right)$
    # Update confidence score of b_j using decay function:
    $b_j.\text{score} = b_j.\text{score} \times (1 - r * iou)$
  end for
  Remove b_i from B
end while

3.3 Network Architecture

One-stage detectors strive to achieve fast inference speed while maintaining high detection accuracy. The YOLO series are the most representative works. Although some new upgrade YOLO detectors have shown a degree of improvement in inference speed, considering the balance between the overall accuracy and speed of the

detector, this paper still uses the typical detector YOLOv7 as the baseline. The backbone networks of the YOLO series have undergone multiple iterations and are already excellent in feature extraction capability and inference speed, but there are still room for improvement in certain aspects. In this paper, we mainly improve the attention mechanism and adjust the prediction heads to adapt to small-scale face detection.

By incorporating the SK [23] module into the EMA module, we designed an EMA+ module to enhance the model's feature extraction ability. Assuming $x \in \mathbb{R}^{H,W,C}$, the two branches in the SK module have convolutional kernel sizes of 3×3 and 5×5, respectively. For the 3×3 branch, the feature map eU is obtained; for the 5×5 branch, the feature map bU is obtained. Concatenate the output feature maps eU and bU from the two branches we can get the fusion feature U as described in Equation (5):

$$U = eU \oplus bU. \tag{5}$$

Perform global average pooling on the feature map U to obtain the statistical information $s_c \in \mathbb{R}^C$ along the channel dimension as Equation (6):

$$s_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} U_c(i,j). \tag{6}$$

Subsequently, feed s_c through a fully connected bottleneck and obtain the concise feature $z \in \mathbb{R}^{d \times 1}$.

$$z = \mathrm{ReLU}(\mathrm{BatchNorm}(Ws)). \tag{7}$$

Subsequently, compute the attention weights a_c and b_c for both branches using the softmax function as Equations (8) and (9):

$$a_c = \frac{\exp(Az_c)}{\exp(Az_c) + \exp(Bz_c)}, \tag{8}$$

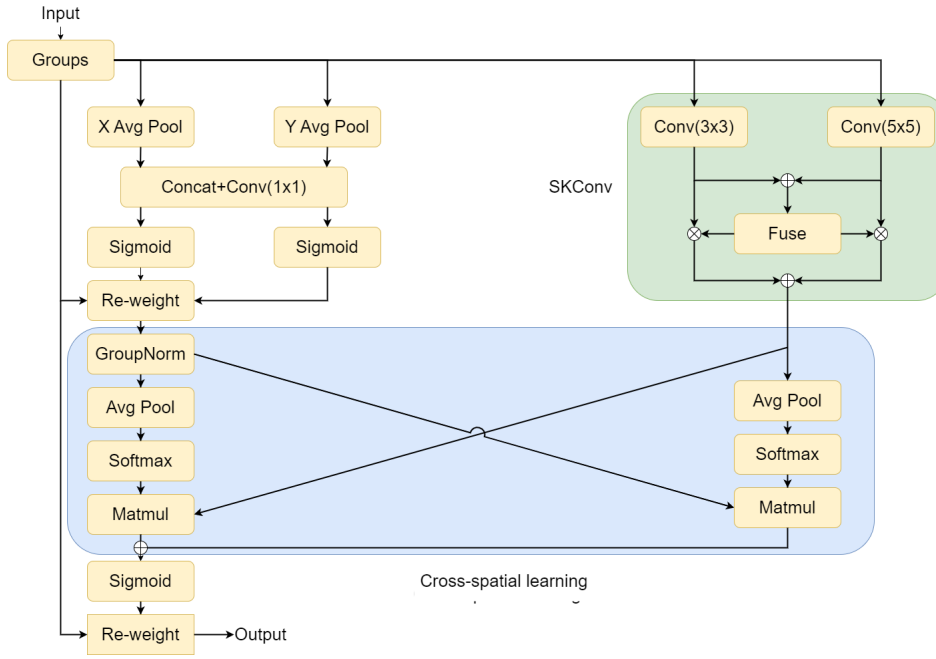
$$b_c = \frac{\exp(Bz_c)}{\exp(Az_c) + \exp(Bz_c)}, \tag{9}$$

where A and B are learnable parameters. Then, perform weighted sum of the output feature maps from the two branches based on the attention weights to obtain the final output feature map $y_c \in \mathbb{R}^{H,W}$.

$$y_c = a_c \cdot eU_c + b_c \cdot bU_c. \tag{10}$$

Following this, 2D global average pooling is employed to encode global spatial information in the SK branch and the left branch, to facilitate cross-spatial learning (see Figure 1 a)).

Furthermore, we delete P_5 prediction head with stride of 32 and add P_2 prediction head with stride of 4 (see Figure 1 b)), aiming to incorporate shallow-featured



a) Structure of EMA+ module

resolution information into P_2 prediction head for a more comprehensive information fusion. Besides, as the number of layers increases, the features become more and more abstract. In our work, we add an additional path between the P_3 layer and its corresponding output layer, as well as the P_4 layer, to fuse more hierarchical features without adding too much cost:

$$P_3^{out} = Concat\left(P_3^{CBS}, P_3^{ELAN}\right), \tag{11}$$

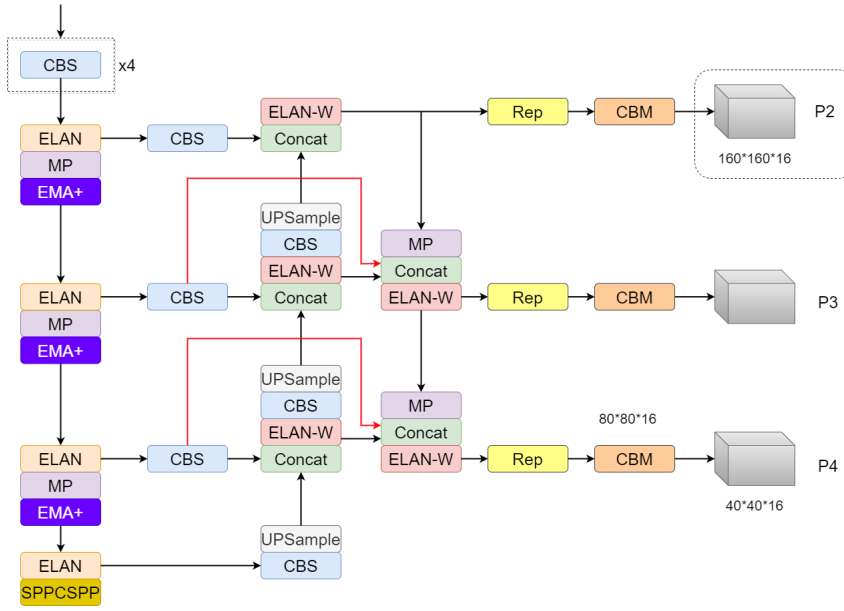
$$P_4^{out} = Concat\left(P_4^{CBS}, P_4^{ELAN}\right), \tag{12}$$

where P_3^{out} refers to the output of layer P_3, P_3^{CBS} refers to the output value of the CBS module in layer P_3, P_3^{ELAN} refers to the output value of the ELAN module in layer P_3, P_4^{out}, P_4^{CBS} and P_4^{ELAN} operate under the same principle.

4 EXPERIMENT

4.1 Dataset

We train our model on the WiderFace [22] dataset. The WiderFace dataset consists of 32 203 images and 393 703 detailed annotated face images, showing a great deal of diversity in scale, occlusion, and pose. The dataset is divided into three parts for



b) Network architecture of our proposed model

Figure 3. Subgraph a) shows network architecture of our proposed model. We plug the EMA+ module behind the ELAN module in backbone network of YOLOv7. Besides, we modify the network structure to fuse more hierarchical features. Subgraph b) shows our proposed EMA+ module, in which we introduce the SK model to replace the 3×3 branch, in an effort to obtain multi-scale feature information.

training, validation, and testing, with proportions of 40 %, 10 %, and 50 %, respectively. In addition, each part is further subdivided into three difficulty levels: easy, medium, and hard. Among them, the hard level is the most challenging and serves as the best standard to test the performance of face detection algorithms on small faces. Our model is evaluated on the validation set of the WiderFace dataset and FDDB [24] dataset. FDDB contains 2845 images, with a total of 5 171 faces from news articles on Yahoo websites. FDDB utilizes elliptical face annotations and introduces two forms of evaluation: the discontinuous score and the continuous score. We employ the discontinuous scoring metric that tallies the number of detected faces in relation to the number of false alarms. A detected bounding box is considered a true positive only when the Intersection over Union (IoU) with the ground-truth face's bounding box exceeds 0.5. We use FDDB to showcase the performance of our model on cross-domain datasets.

4.2 Training Settings

We implement our work based on the PyTorch library. Models are optimized by SGD with a batch size of 8. The initial learning rate is set to 1e−2, the final learning rate is 1e−3. The NMS IoU threshold is set to 0.5. We train the model on a 3090 GPU with 8 dataloader workers. The model is trained for a total of 250 epoches. We use the accuracy and AP (Average Precision) as the evaluation metric.

4.3 Ablation Experiment

4.3.1 SF-NMS

We compare the performance of different NMS algorithms using the same training weight. As shown in the second row of Table 1, Soft-NMS algorithm decrease the accuracy slightly on the easy and medium subsets, it may be because there are not many faces occluded in the easy and medium subsets, resulting in the algorithm can not play a normal performance on crowded faces but slightly worsen the detection of normal faces. Meanwile, it improves the accuracy by 1.3 % on the hard subset significantly, which means that it greatly improve the detection performance in crowded and occluded environments. Furthermore, the third row of Table 1 shows that with the introduction of the face aspect ratio variable, the performance of the model has been further improved, with an increase of 0.1 %, 0.1 %, and 0.5 % on the easy, medium and hard subset compared to Soft-NMS. As shown in Figure 4, in some crowded scenes, SF-NMS algorithm performs better.

Subset	Easy	Medium	Hard
Baseline	**94.7**	**93.6**	87.0
+Soft-NMS	94.5	93.4	88.3
+SF-NMS	94.6	93.5	**88.8**

Table 1. Effectiveness of Soft-NMS and SF-NMS

4.3.2 Network Architecture

From the second row of Table 3, it is evident that the modified baseline method brings about a notable enhancement in detection accuracy, achieving gains of 0.1 %, 0.4 %, and 0.5 % across the easy, medium, and hard subsets, respectively. The performance of our enhanced EMA+ module is detailed in Table 2. In comparison to the standard EMA module, the EMA+ module delivers a noticeable improvement in Average Precision (AP), increasing it by 0.08 %. Moreover, as illustrated in the third row of Table 3, the integration of the EMA+ module further augments accuracy, providing an additional increase of 0.1 % and 0.2 % in the medium and hard subsets, respectively. This progression not only reinforces the robustness of our approach but also demonstrates its potential to consistently enhance performance across varying levels of difficulty.

Figure 4. Inference results compared between traditional NMS (bottom left), Soft-NMS (top right) and our SF-NMS (bottom right). Our proposed SF-NMS performs better in crowded scenes.

Subset	AP
Baseline	78.92
+EMA	79.05
+EMA+	**79.13**

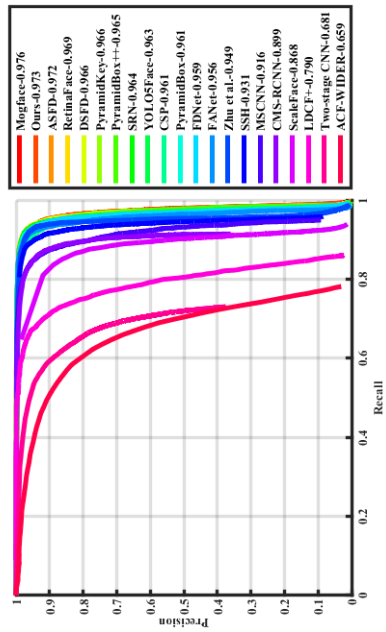
Table 2. Effectiveness of EMA+

4.3.3 Sa-SimOTA

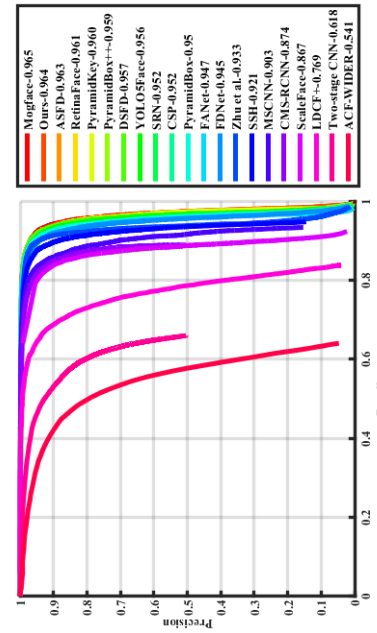
In Table 4, we provide a comparative analysis between Sa-SimOTA and SimOTA. Our Sa-SimOTA method demonstrates notable improvements in detection accuracy across different difficulty subsets, specifically achieving gains of 0.1 %, 0.1 %, and 0.3 % on the easy, medium, and hard subsets, respectively. These improvements

Subset	Easy	Medium	Hard
Baseline	94.7	93.6	87.0
Modified baseline	94.8	94.0	87.5
Modified baseline + EMA+	**94.8**	**94.1**	**87.7**

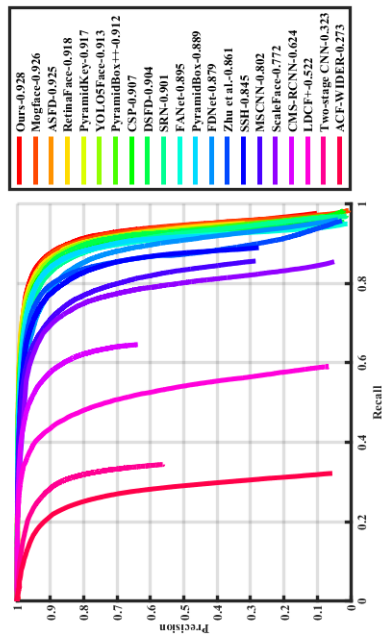
Table 3. Effectiveness of modified network architecture



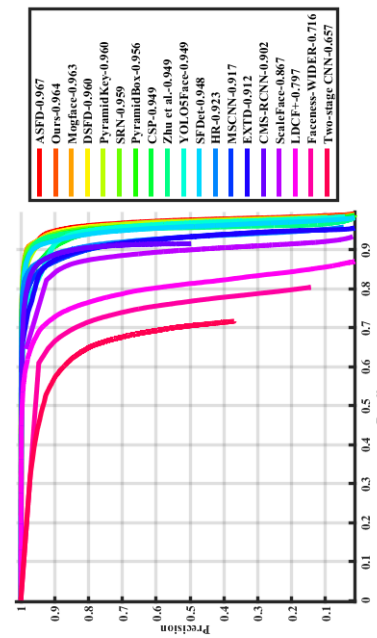
a) Val: Easy



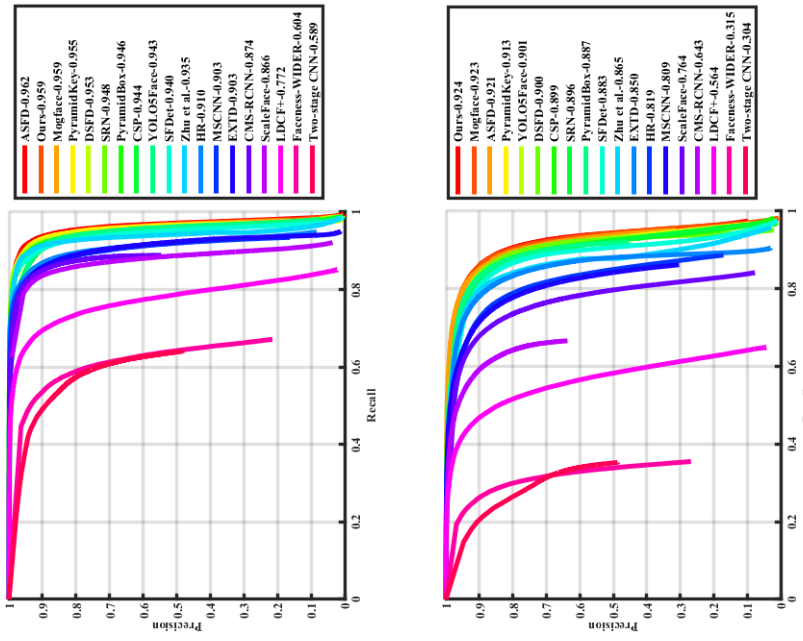
b) Val: Medium



c) Val: Hard



d) Test: Easy



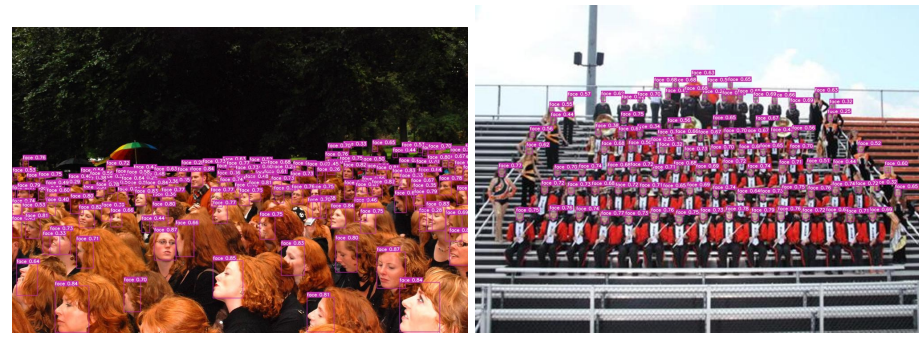
e) Test: Medium f) Test: Hard

Figure 5. Precision-Recall (PR) curves on WiderFace validation and test subsets

can be attributed to the mechanism we previously discussed, wherein increasing the number of positive samples for smaller targets enhances the model's detection performance for such targets. Consequently, our model achieves an overall accuracy of 94.8 %, 94.2 %, and 89.4 % on the easy, medium, and hard subsets, as shown in Table 4. These results underscore the efficacy of Sa-SimOTA in enhancing detection robustness, particularly in challenging scenarios.

SF-NMS	Sa-SimOTA	Network	Easy	Medium	Hard
			94.7	93.6	87.0
✓			94.6	93.5	88.8
	✓		94.8	93.7	87.
		✓	94.8	94.1	87.7
✓	✓	✓	**94.8 (+0.1)**	**94.2 (+0.6)**	**89.4 (+2.4)**

Table 4. Ablation study results on the WiderFace validation set



a) b)

c) d)

Figure 6. a), b), c) and d) are some examples of detected small faces using CF-YOLO

4.4 Comparisons with Existing Face Detectors

The precision-recall (PR) curves of our model, along with those of other state-of-the-art face detectors, are illustrated in Figure 5 and further detailed in Table 5. To provide a comprehensive evaluation, we tested our face detector on the challenging Widerface and FDDB datasets, comparing its performance against several contemporary methods [16, 19, 26, 27, 28, 30, 32, 33, 34]. Our model, based on the YOLOv7-e6e architecture, demonstrates superior performance with precision scores of 97.3 % (Easy), 96.4 % (Medium), and 92.8 % (Hard) on the validation dataset, and 96.4 % (Easy), 95.9 % (Medium), and 92.4 % (Hard) on the test dataset. Furthermore, as detailed in Table 6, our method achieves a precision of 98.59 % on the FDDB [24] dataset. Figure 6 are some actual detection examples about smallfaces on WiderFace by CF-YOLO. These results underscore the robustness and reliability of our approach in accurately detecting faces under a wide range of uncontrolled conditions, outperforming existing methodologies in terms of both precision and recall.

Detector	Easy	Medium	Hard	Params (M)	Flops (G)
DSFD [17]	96.6	95.7	90.4	101.3	515.4
SRN [25]	96.4	95.2	90.1	121.6	625.4
ASFD [26]	97.2	96.3	92.5	156.9	803.4
RetinaFace [27]	96.9	96.1	91.8	130.6	794.4
PyramidBox++ [28]	96.5	95.9	91.2	121.2	697.6
YOLO5Face [29]	96.3	95.6	91.3	141.2	188.6
MogFace [19]	**97.6**	**96.5**	92.6	164.6	902.1
PyramidKey [30]	96.6	96.0	91.7	157.2	814.1
Ours+YOLOv7	94.8	94.2	89.4	37.2	104.1
Ours+YOLOv7-e6	96.5	95.8	92.4	97.2	515.2
Ours+YOLOv7-e6e	97.3	96.4	**92.8**	151.7	843.2

Table 5. Comparison of our methods and existing face detectors on the WiderFace validation dataset

Method	AP
Ours+YOLOv7-e6e	**0.986**
BBFCN [31]	0.984
SFD [16]	0.983
FANet [32]	0.983
HR-ER [33]	0.976
DeepIR [34]	0.971
FaceBoxes [15]	0.960

Table 6. Evaluation of our model on the FDDB dataset

5 CONCLUSION

We introduce SF-YOLO, a face detector designed to address the challenge of detecting small, mutually occluded faces in crowded scenes. Experiments reveal that SF-YOLO, compared to the baseline YOLOv7, achieves 97.3 %, 96.4 %, and 92.8 % accuracy on WiderFace's easy, medium, and hard subsets, respectively, with its hard subset accuracy rivaling the state-of-the-art, showcasing a significant improvement over contemporary face detection methods. In future work, we plan to extend our approach by experimenting with YOLOv8-10, the next iteration in the YOLO series. Given the advancements anticipated in YOLOv8-10, such as improved architecture and optimized training mechanisms, applying our methods to this model could yield even greater enhancements in detection accuracy and efficiency. This exploration will help determine the scalability and adaptability of our approach, potentially setting a new benchmark for real-time object detection in complex environments.

6 DATA AVAILABILITY STATEMENT

All data underlying this research are available publicly at `https://github.com/bistuxlf/CF-YOLO/tree/master`.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 62376286 and No. 62105038) and the Research and Development Program of Beijing Municipal Education Commission (No. KM202211232001). We are grateful for the support of the organizations.

REFERENCES

[1] Chen, W.—Huang, H.—Peng, S.—Zhou, C.—Zhang, C.: YOLO-Face: A Real-Time Face Detector. The Visual Computer, Vol. 37, 2021, No. 4, pp. 805–813, doi: 10.1007/s00371-020-01831-7.

[2] Yu, Z.—Huang, H.—Chen, W.—Su, Y.—Liu, Y.—Wang, X.: YOLO-Facev2: A Scale and Occlusion Aware Face Detector. Pattern Recognition, Vol. 155, 2024, Art. No. 110714, doi: 10.1016/j.patcog.2024.110714.

[3] Lin, T. Y.—Goyal, P.—Girshick, R.—He, K.—Dollár, P.: Focal Loss for Dense Object Detection. 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2999–3007, doi: 10.1109/ICCV.2017.324.

[4] Redmon, J.—Farhadi, A.: YOLOv3: An Incremental Improvement. CoRR, 2018, doi: 10.48550/arXiv.1804.02767.

[5] Ge, Z.—Liu, S.—Li, Z.—Yoshie, O.—Sun, J.: OTA: Optimal Transport Assignment for Object Detection. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 303–312, doi: 10.1109/CVPR46437.2021.00037.

[6] Ge, Z.—Liu, S.—Wang, F.—Li, Z.—Sun, J.: YOLOX: Exceeding YOLO Series in 2021. CoRR, 2021, doi: 10.48550/arXiv.2107.08430.

[7] Xu, C.—Wang, J.—Yang, W.—Yu, H.—Yu, L.—Xia, G. S.: RFLA: Gaussian Receptive Field Based Label Assignment for Tiny Object Detection. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G. M., Hassner, T. (Eds.): Computer Vision – ECCV 2022. Springer, Cham, Lecture Notes in Computer Science, Vol. 13669, 2022, pp. 526–543, doi: 10.1007/978-3-031-20077-9_31.

[8] Wang, C. Y.—Bochkovskiy, A.—Liao, H. Y. M.: YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7464–7475, doi: 10.1109/CVPR52729.2023.00721.

[9] Wang, A.—Chen, H.—Liu, L.—Chen, K.—Lin, Z.—Han, J.—Ding, G.: YOLOv10: Real-Time End-to-End Object Detection. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (Eds.): Advances in Neural Information Processing Systems 37 (NeurIPS 2024). Curran Associates, Inc., 2024, pp. 107984–108011, doi: 10.52202/079017-3429.

[10] Lin, T. Y.—Dollár, P.—Girshick, R.—He, K.—Hariharan, B.—Belongie, S.: Feature Pyramid Networks for Object Detection. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 936–944, doi: 10.1109/CVPR.2017.106.

[11] Liu, S.—Qi, L.—Qin, H.—Shi, J.—Jia, J.: Path Aggregation Network for Instance Segmentation. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 8759–8768, doi: 10.1109/CVPR.2018.00913.

[12] Tan, M.—Pang, R.—Le, Q. V.: EfficientDet: Scalable and Efficient Object Detection. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10778–10787, doi: 10.1109/CVPR42600.2020.01079.

[13] Ouyang, D.—He, S.—Zhang, G.—Luo, M.—Guo, H.—Zhan, J.—Huang, Z.: Efficient Multi-Scale Attention Module with Cross-Spatial Learning. ICASSP 2023 – 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023, pp. 1–5, doi: 10.1109/ICASSP49357.2023.10096516.

[14] Bodla, N.—Singh, B.—Chellappa, R.—Davis, L. S.: Soft-NMS – Improving Object Detection with One Line of Code. 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5562–5570, doi: 10.1109/ICCV.2017.593.

[15] Zhang, S.—Zhu, X.—Lei, Z.—Shi, H.—Wang, X.—Li, S. Z.: FaceBoxes: A CPU Real-Time Face Detector with High Accuracy. 2017 IEEE International Joint Conference on Biometrics (IJCB), 2017, pp. 1–9, doi: 10.1109/BTAS.2017.8272675.

[16] Zhang, S.—Zhu, X.—Lei, Z.—Shi, H.—Wang, X.—Li, S. Z.: S3FD: Single Shot Scale-Invariant Face Detector. 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 192–201, doi: 10.1109/ICCV.2017.30.

[17] Li, J.—Wang, Y.—Wang, C.—Tai, Y.—Qian, J.—Yang, J.—Wang, C.—Li, J.—Huang, F.: DSFD: Dual Shot Face Detector. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5055–5064, doi: 10.1109/CVPR.2019.00520.

[18] Liu, Y.—Tang, X.—Han, J.—Liu, J.—Rui, D.—Wu, X.: HAMBox: Delving into Mining High-Quality Anchors on Face Detection. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13043–13051, doi: 10.1109/CVPR42600.2020.01306.

[19] Liu, Y.—Wang, F.—Deng, J.—Zhou, Z.—Sun, B.—Li, H.: MogFace: Towards a Deeper Appreciation on Face Detection. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4093–4102, doi: 10.1109/CVPR52688.2022.00406.

[20] Zheng, Z.—Wang, P.—Liu, W.—Li, J.—Ye, R.—Ren, D.: Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression. Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, No. 7, pp. 12993–13000, doi: 10.1609/aaai.v34i07.6999.

[21] Liu, S.—Huang, D.—Wang, Y.: Adaptive NMS: Refining Pedestrian Detection in a Crowd. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6452–6461, doi: 10.1109/CVPR.2019.00662.

[22] Yang, S.—Luo, P.—Loy, C. C.—Tang, X.: WIDER FACE: A Face Detection Benchmark. 2016 IEEE Conference on Computer Vision and Pattern Recognition

(CVPR), 2016, pp. 5525–5533, doi: 10.1109/CVPR.2016.596.

[23] Li, X.—Wang, W.—Hu, X.—Yang, J.: Selective Kernel Networks. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 510–519, doi: 10.1109/CVPR.2019.00060.

[24] Jain, V.—Learned-Miller, E.: FDDB: A Benchmark for Face Detection in Unconstrained Settings. Technical Report No. UM-CS-2010-009, University of Massachusetts, Amherst, 2010, https://web.cs.umass.edu/publication/docs/2010/UM-CS-2010-009.pdf.

[25] Zhang, S.—Zhu, R.—Wang, X.—Shi, H.—Fu, T.—Wang, S.—Mei, T.—Li, S. Z.: Improved Selective Refinement Network for Face Detection. CoRR, 2019, doi: 10.48550/arXiv.1901.06651.

[26] Zhang, B.—Li, J.—Wang, Y.—Tai, Y.—Wang, C.—Li, J.—Huang, F.—Xia, Y.—Pei, W.—Ji, R.: ASFD: Automatic and Scalable Face Detector. CoRR, 2020, doi: 10.48550/arXiv.2003.11228.

[27] Deng, J.—Guo, J.—Zhou, Y.—Yu, J.—Kotsia, I.—Zafeiriou, S.: RetinaFace: Single-Stage Dense Face Localisation in the Wild. CoRR, 2019, doi: 10.48550/arXiv.1905.00641.

[28] Li, Z.—Tang, X.—Han, J.—Liu, J.—He, R.: PyramidBox++: High Performance Detector for Finding Tiny Face. CoRR, 2019, doi: 10.48550/arXiv.1904.00386.

[29] Qi, D.—Tan, W.—Yao, Q.—Liu, J.: YOLO5Face: Why Reinventing a Face Detector. In: Karlinsky, L., Michaeli, T., Nishino, K. (Eds.): Computer Vision – ECCV 2022 Workshops. Springer, Cham, Lecture Notes in Computer Science, Vol. 13805, 2023, pp. 228–244, doi: 10.1007/978-3-031-25072-9_15.

[30] Earp, S. W. F.—Noinongyao, P.—Cairns, J. A.—Ganguly, A.: Face Detection with Feature Pyramids and Landmarks. CoRR, 2019, doi: 10.48550/arXiv.1912.00596.

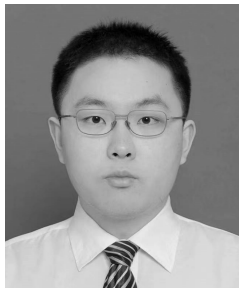
[31] Liu, L.—Li, G.—Xie, Y.—Yu, Y.—Wang, Q.—Lin, L.: Facial Landmark Machines: A Backbone-Branches Architecture with Progressive Representation Learning. IEEE Transactions on Multimedia, Vol. 21, 2019, No. 9, pp. 2248–2262, doi: 10.1109/TMM.2019.2902096.

[32] Zhang, J.—Wu, X.—Hoi, S. C. H.—Zhu, J.: Feature Agglomeration Networks for Single Stage Face Detection. Vol. 380, 2020, doi: 10.1016/j.neucom.2019.10.087.

[33] Hu, P.—Ramanan, D.: Finding Tiny Faces. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1522–1530, doi: 10.1109/CVPR.2017.166.

[34] Sun, X.—Wu, P.—Hoi, S. C. H.: Face Detection Using Deep Learning: An Improved Faster RCNN Approach. Neurocomputing, Vol. 299, 2018, pp. 42–50, doi: 10.1016/j.neucom.2018.03.030.

Hongbo Huang is Associate Professor at the Computer School of Beijing Information Science and Technology University, China. He is also Director of the Institute of Computing Intelligence of Beijing Information Science and Technology University. He received his Ph.D. degree in control science and engineering in 2015 from the University of Science and Technology Beijing. His research interests include computer vision, machine learning and video semantic analysis.



Longfei Xu received his B.Sc. degree in biotechnology from the Northwest A & F University, Shanxi, China, in 2022. He is now working toward his M.Sc. degree in computer science and technology with the School of Computer, Beijing Information Science and Technology University, Beijing, China. His research interests include object detection, video reconstruction.

Xiaoxu Yan received his B.Sc. degree in computer science and technology from the Beijing Information Science and Technology University, Beijing, China, in 2018. He is now working toward his M.Sc. degree in computer science and technology with the School of Computer, Beijing Information Science and Technology University, Beijing, China. His research interests include object detection, video reconstruction.

Linkai Huang received his B.Sc. degree in surveying and mapping from the Southwest Petroleum University, Sichuan, China, in 2021. He is now working toward his M.Sc. degree in computer science and technology with the School of Computer, Beijing Information Science and Technology University, Beijing, China. His research interests include object detection, video reconstruction.