https://doi.org/10.46335/IJIES.2024.9.9.7 e-ISSN: 2456-3463
Vol. 9, No. 9, 2024, PP. 33-38
International Journal of Innovations in Engineering and Science, www.ijies.net

Attendance Marking System using YOLOv5

Vaibhav Deshpande, Ayush Gupta, Jay Shinde, Jyotiraditya Parihar, Mayur Khonde

,St. Vincent Pallotti College of Engineering & Technology Gavasi Manapur, Nagpur, Maharashtra, India

vdeshpande@stvincentngp.edu.in

***Received on**: 5 May,2024* ***Revised on**: 12 July,2024* ***Published on**: 15 July ,2024*

I. INTRODUCTION

Face detection is indispensable for many visual tasks and has been widely used in various practical applications, such as intelligent surveillance for smart cities, face unlocking in smartphones, and beauty filters. However, face detection still has many challenges due to the interference of shooting angle, background noise, image quality, face scale, and other factors. In practical scenarios, the missing detection problem of small-scale faces results in poor performance of former face detectors. Thus, many scholars have launched researches on blurring small-size human faces.

Over the past decades, convolutional neural networks (CNNs) have been certified to be useful models for processing a wide range of visual tasks, and we have witnessed the rapid development of general object detectors. The commonly used target detection framework is divided into two branches [1], two-stage detectors and one-stage detectors. Typical algorithms of two-stage detectors include faster R-texture information and improves the authenticity of visual perception. The algorithm proposed in the paper is called SR-YOLOv5, which guarantees the detection speed while improving the detection accuracy of small targets.

II. RELATED WORK

In this section, we introduce the related work from three following parts. First, we review recent progress on face detection in low-resolution conditions. Second, we give an overall description of the YOLO series. Third, we describe the principle of the SR network.

A. Face Detection

Face detection has received much attention due to its wide practical applications [14]. Before deep convolutional neural network (deep CNN) was widely used, hand-made features were a very important part of face detectors. Researchers proposed many robust hand-made features [15], such as HAAR [16], HOG [17], LBP [18], SIFT [19], DPM [20], and ACF [21]. However, the performance of these feature extractors has been far surpassed by deep CNN. In recent years, numerous models have emerged, and deep CNN has shown excellent performance in general target detection tasks. The target detection task is modeled as two problems of classification and regression of target candidate regions. There are many object detection networks including RCNN family [2, 5, 15,22], SSD From the multiscale, small face, low light, dense scene, and other challenges encountered in face detection, face detection is the same as general target detection. Thus, face detection networks can learn from general object detection networks. There are also some specific problems containing scale, pose, occlusion, expression, and makeup. Many researchers developed methods to deal with the above problems, such as Cascade CNN, MTCNN, HR, and SSH. They also test their algorithm on public datasets [27].

B. SR

In the actual application scene, some images will be fuzzy and of low quality because of the limitation of environment and shooting technology. Such images have poor performance in the region of interest (RoI). Therefore, the researchers proposed the image super resolution reconstruction technology to enrich the detailed information of low- resolution images and improve the expression ability of images. Currently, super resolution reconstruction technology [13] based on deep learning is widely used. Among them, the super resolution image generated by the Generative Adversarial Networks (GAN) [12] has a better visual effect, which is called SRGAN. By training a generation function, SRGAN converts the input low-resolution image into the corresponding super resolution image [28]. Based on SRRes- Net, SRGAN uses perceptual loss and adversarial loss to make the generated images closer to the target images.

The SRGAN network is composed of a generator and a discriminator, and its network model is shown as in Figure 1 [13] below. The core of the generator network is multiple residual blocks, each residual block containing two 3 × 3 convolutional layers. After the convolutional layer is a batch normalization layer, PReLU is used as the activation function [29]. The discriminant network uses a network structure similar to VGG-19, but without maximum pooling. The discriminant network contains eight convolutional layers. As the number of network layers increases, the number of features increases, and the size of features decreases. Leaky ReLU acts as an activation function. Finally, the network uses two full convolution layers and a sigmoid activation function to capture the potentiality of the learned real sample, which is used to determine whether the image comes from the high-

https://doi.org/10.46335/IJIES.2024.9.9.7 e-ISSN: 2456-3463
Vol. 9, No. 9, 2024, PP. 33-38
International Journal of Innovations in Engineering and Science, www.ijies.net

resolution image of the real sample or the super resolution image of the fake sample.

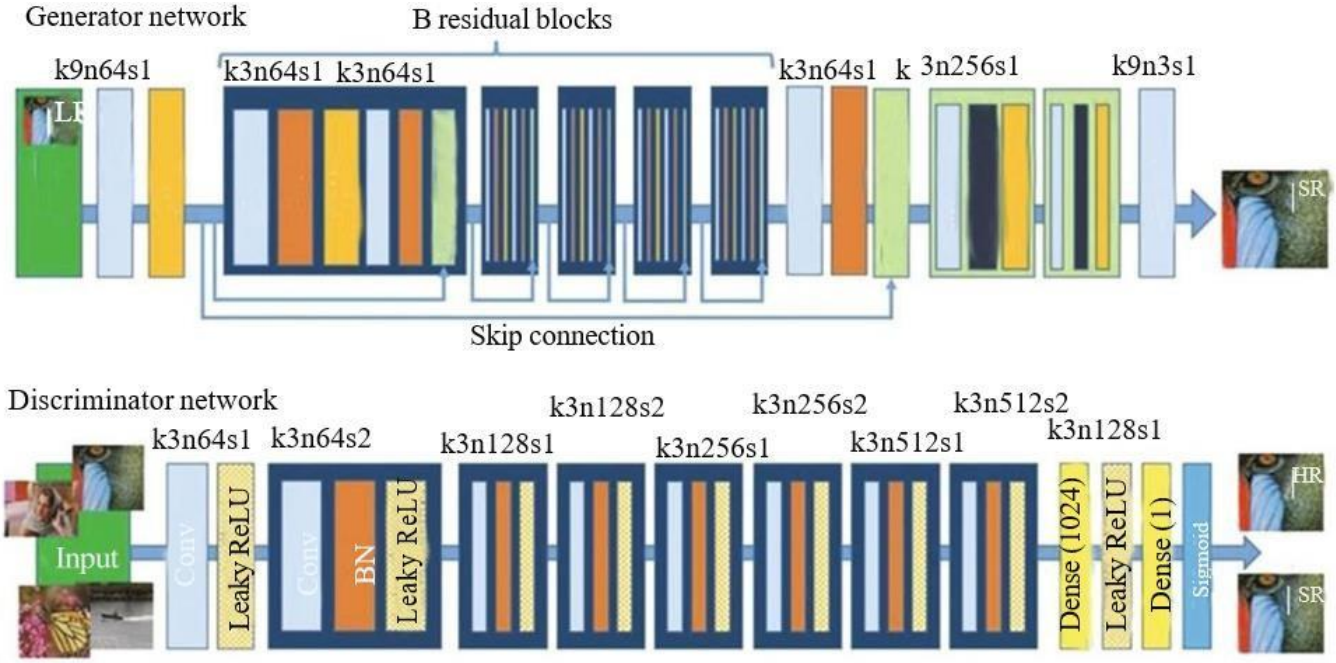
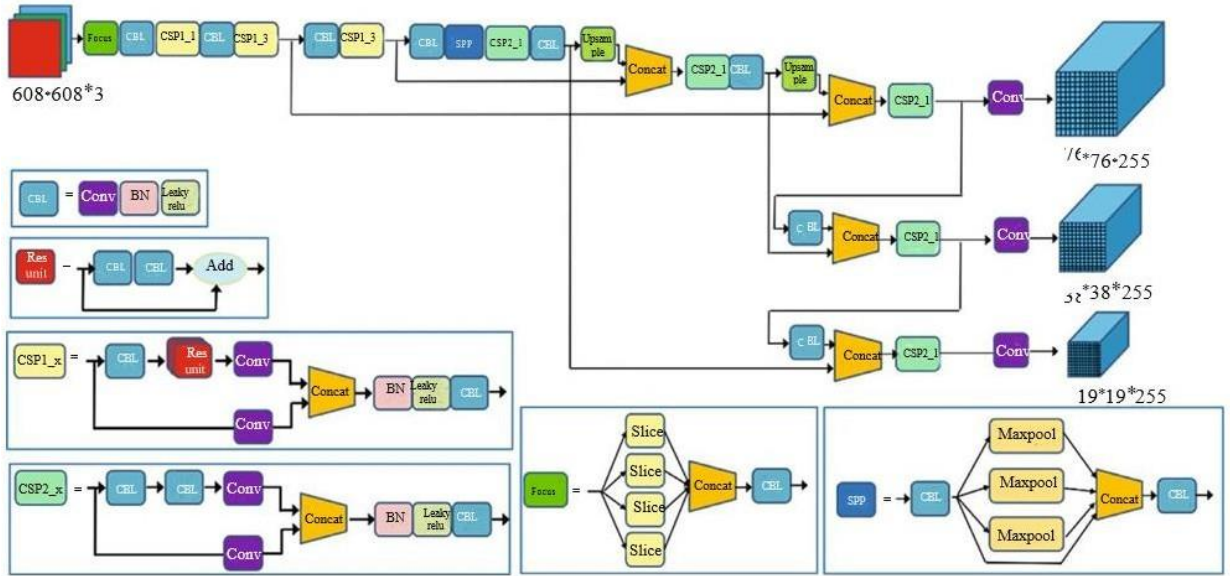


Fig 1: SRGAN Network Model



target. C is the total number of target categories. If the center of the target is in this grid, then the target will be acquired and judge whether it is a human face. The position of the regression box of the target can be obtained by the following formula:

$$C_i^j = P_{i,j} * \mathrm{IOU}_{\mathrm{pred}}^{\mathrm{truth}}. \quad (1)$$

In the above parameters, i and j represent the jth regression box of the ith grid, C_i^j represents the confidence score of the jth bounding box of the ith grid. $P_{i,j}$ represents whether there is a target, if the target is in the jth box, the value of $P_{i,j} = 1$; otherwise, $P_{i,j} = 0$. The $\mathrm{IOU}_{\mathrm{pred}}^{\mathrm{truth}}$ is a widely used parameter that represents the intersection over union between the predicted box and ground truth box [31]. The higher the IOU score, the more accurate the position of the predicted box.

1.3. Loss Function of YOLOv5s. The loss function can be expressed as follows:

$$\mathrm{loss} = l_{\mathrm{box}} + l_{\mathrm{cls}} + l_{\mathrm{obj}}, \quad (2)$$

where l_{box}, l_{cls}, and l_{obj} are bounding box regression loss function, classification loss function, and confidence loss function, respectively.

The bounding box regression loss function is defined as

$$l_{\mathrm{box}} = \lambda_{\mathrm{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{ij}^{\mathrm{obj}} bj(2 - w_i \times h_i) \quad x_i - x\wedge_i^j \; ^2 + y_i - y\wedge_i^j \; ^2 + w_i - w\wedge_i^j \; ^2 + h_i - h\wedge_i^j \; ^2. \quad (3)$$

https://doi.org/10.46335/IJIES.2024.9.9.7 e-ISSN: 2456-3463
Vol. 9, No. 9, 2024, PP. 33-38
International Journal of Innovations in Engineering and Science, www.ijies.net

III. METHODOLOGY

This paper focuses on improving the detection accuracy of small faces in surveillance images. Because of the comparison of the four versions of YOLOv5 including YOLOv5m, YOLOv5l, YOLOv5x, and YOLOv5s, the YOLOv5s model is smaller and easier to deploy quickly. Therefore, our research is based on the YOLOv5s model. We optimize the backbone, then integrate image super resolution technology on the head and improve the loss function to ensure efficient detection speed.

1. SR-YOLOv5

1.1. Adaptive Anchor

The calculation of adaptive anchor is added in YOLOv5s. Before each training, the K-means algorithm is used to cluster the ground truth of all samples in the training set and to find out the optimal group of anchors point frames in the high complexity and high recall rate. The results of anchor boxes clustered by the algorithm are shown in Table

1.2. Network Architecture

1.2.1 Backbone: The overall architecture of improved YOLOv5s is depicted in Figure 3 which consists of the backbone, detection neck, and detection head. Firstly, a newly designed backbone named CSPNet is used. We change it with a new block called CBS consists of Conv layer, BN layer, and a SILU [32]. Secondly, a stem block is used to replace the focus layer in YOLOv5s. Thirdly, a C3 block is used to replace the original CSP block with two halves One is passed through a CBS block, some bottleneck blocks, and a Conv layer, while another consists of a Conv layer. After the two paths with a CONCAT and a CBS block followed, we also change the SPP block [4] to improve the face detection performance. In this block, the size of the three kernels is modified to smaller kernels.

1.2.2 Detection Neck: The structure of the detection neck is also shown in Figure 3 which consists of a normal feature pyramid network (FPN) [23] and path aggregation network (PAN) [3]. However, we modify the details of some mod- ules, such as the CS block and the CBS block, we proposed.

1.2.3 Detection Head: Through feature pyramid structure and path aggregation [33] network, the front segment of the network realizes the full fusion of low-level features and high-level features to obtain rich feature maps, which can detect the most high-resolution face samples. However, for low-resolution images, feature fusion cannot enhance the original information of the image, and through layers of iteration, the prior information of small faces is still lacking. To enhance the detection rate of small faces in low-resolution images, SR is fused in the detection head part of the network. For the grid to be determined, the region information is input into SRGAN to carry out super resolution reconstruction and face detection again through its coordinate information. Finally, the output of the two-stage face detector is integrated and output.

2. LOSS FUNCTION

IOU is a frequently used index in target detection. In most anchor-based [34] methods, it is used not only to judge the positive and negative sample but also to assess the distance between the location of the predicted boxand the ground truth. The paper proposes that a regression positioning loss [35] should be considered: overlapping area, center point distance, and aspect ratio, which have aroused wide concern. At present, more and more researchers propose better performance algorithms, such as IOU, GIOU, DIOU, and CIOU. In this paper, we propose to replace GIOU in YOLOv5s with CIOU and nonmaximal suppression (NMS). Our bounding box regression loss function is defined as the combination of CIOU and NMS, the candidate box in the same grid can be judged and screened several times through the cyclic structure, which can effectively avoid the problem of missed detection.

IV. EXPERIMENTATION

1.1. Dataset and Experimental Environment Configuration. This experiment uses a face detection benchmark called wider face [27], which is recognized as the largest one among public available datasets. The details of publicly available datasets are shown in Table 2. These faces in the wider face dataset have great changes in scale, posture, and occlusion with an average of 12.2 faces per image, and there are many dense small faces.

$$l'_{\text{CIOU}} = 1 - \text{IOU} + \frac{\rho^2(b, \hat{b})}{c^2} + \frac{16}{\pi^4}\frac{(\arctan(w\wedge/h\wedge) - \arctan(w/h))^4}{1 - \text{IOU} + (4/\pi^2)(\arctan(w\wedge/h\wedge) - \arctan(w/h))^2}, \quad (6)$$

The dataset contains three parts: training set, validation set, and test set, accounting for 40%, 10%, and 50% of the sample number, respectively. This paper focuses on the detection of small faces, which will be more difficult to detect. Therefore, the verification set and test set are divided into three difficulty levels: easy, medium, and hard. There are many small-scale faces in the hard subset, most of which are 10 pixels~50 pixels. Thus, benchmark is suitable to verify the effectiveness and performance in realistic scenes. The experimental environment configuration is shown in Table 3.

Training and testing ofSR-YOLOv5 Models

1.2. Training Model. The YOLOv5s code [11] is used as our basic framework, and we implement all the modifications as described above in PyTorch. We set the initial learning rate at 1E-2, and then we go down to 1E-5 with the decay rate of 5E-3. We set momentum at 0.8 in the first 20 epochs. After that, the momentum is 0.937. The precision-recall (PR) curves of our SR-YOLOv5 detector are shown in Figure 4.

1.3. Testing Model. The detection effect of our improved algorithm on the wider face dataset is shown in Figure 5. It can be seen that this method has good robustness and high accuracy for small faces in various complex scenes. (a) The figure can detect faces with slight occlusion. (b) The figure itself has a low resolution, but the detection result shows thatthe detection effect is still good. (c) The figure fully shows that numerous small faces can be well detected even in a high-density crowd.

1.4. Evaluation Index. In the evaluation of the effect of face detection, there are some relevant parameters: TP (true positives) means that the face is detected, and there are faces in the actual picture; TN (true negatives) means that no face is detected, and no face exists in the actual picture; FP (false positives) means that faces are detected when there is no facein the actual image. FN (false negatives) means that no face is detected, but there are faces in the actual image. The evaluation indexes of the model in this paper include recall rate R, accuracy rate P, and F_1 score. The recall rate is used to evaluate the proportion of faces detected to the total face price

https://doi.org/10.46335/IJIES.2024.9.9.7 e-ISSN: 2456-3463
Vol. 9, No. 9, 2024, PP. 33-38
International Journal of Innovations in Engineering and Science, www.ijies.net

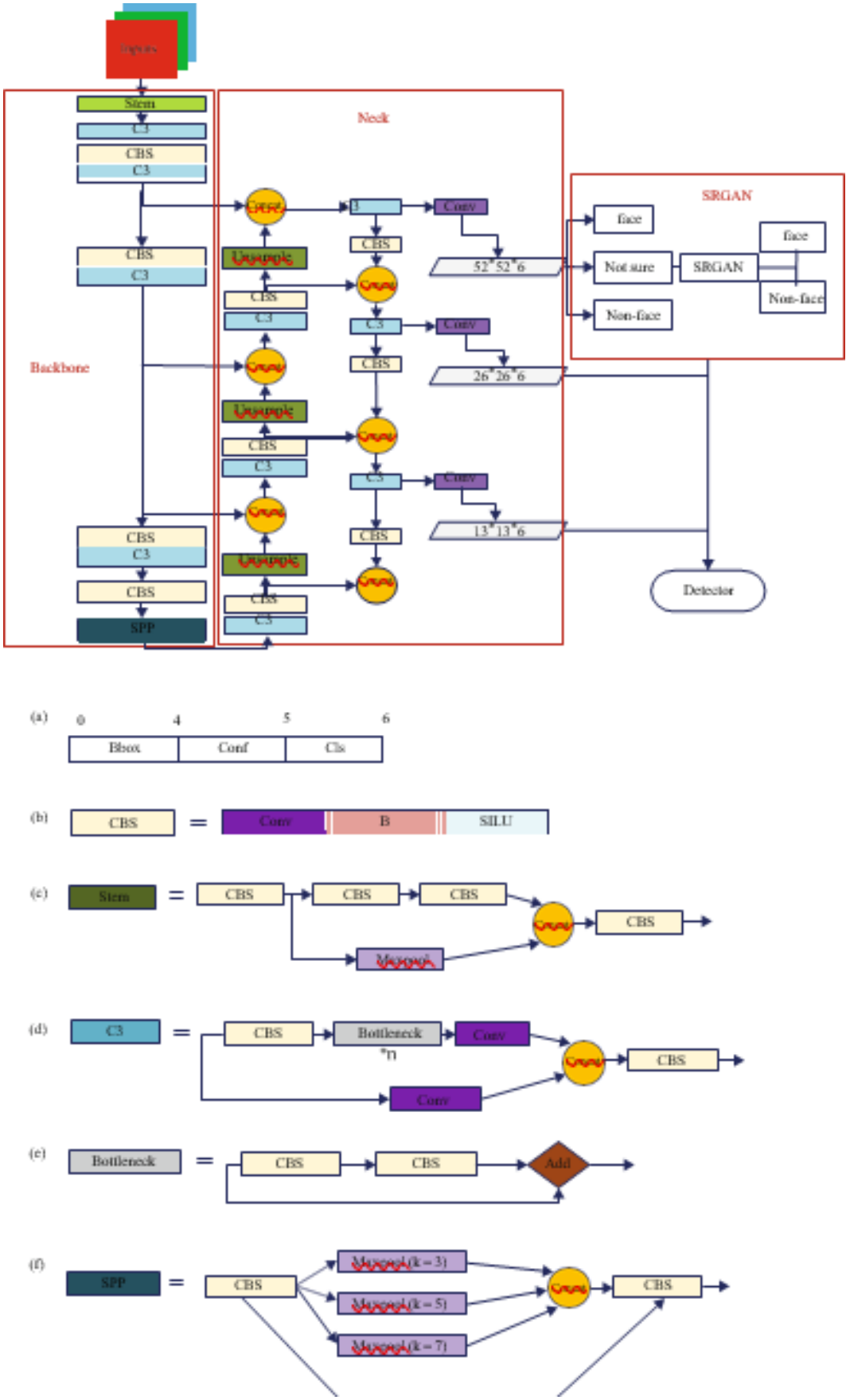


Figure 3: The architecture of improved SR-YOLOv5.

https://doi.org/10.46335/IJIES.2024.9.9.7 e-ISSN: 2456-3463
Vol. 9, No. 9, 2024, PP. 33-38
International Journal of Innovations in Engineering and Science, www.ijies.net

TABLE 2: Available datasets.

Datasets	Pictures	Faces
Wider face	32203	393703
AFW	205	473
FDDB	2845	5171
Pascal face	851	1341
IJB-A	24327	49759
MALF	5250	11931

TABLE 3: Experimental environment configuration.

Experimental environment	Configuration
Operating system	Linux 64
GPU	TITAN Xp
CPU	Intel(R)Core i7-3770CPU@
Deep learning framework	PyTorch

detected, When the two are close, refer to F_1 score, and the higher the score of F_1, the better the algorithm will be.

$$P = \frac{TP}{TP + FP}, \quad (7)$$

$$R = \frac{TP}{TP + FN}, \quad (8)$$

$$F_1 = \frac{2 \times P \times R}{P + R} \quad (9)$$

The trained model is verified on the validation set, and the recall rate R = 0.96, accuracy rate P = 0.975, and F1 =
96.75 were obtained from Equations (6), (8), and (9). From the point of view of the score, the proposed algorithm has better performance.

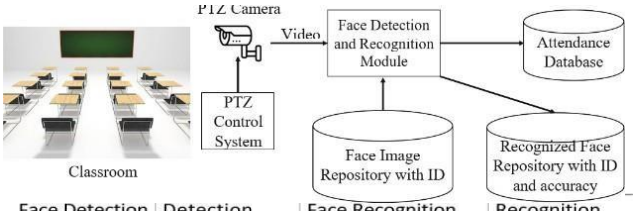
1.5. Model Performance Analysis. After the fusion of SRGAN in the YOLOv5 network, the rationality and effectiveness of the fused network should be verified first. We select 1000 pictures from the test set for network model test and com- parison. As shown in Table 4, compared with YOLOv3, the speed of the network after the fusion of superpartition reconstruction technology is reduced, because the network depth is increased when the new network is integrated. Compared with the HR using Resnet101 as the backbone network, the average detection accuracy of the improved network has been significantly improved, which is 2.3% higher than HR.

Comparison of Accuracy of Relevant Algorithms. To demonstrate the effectiveness of the algorithm, some excel- lent face detection algorithms are selected to test on the wider face dataset, and the results are analyzed. As shown in Table 5, all existing methods achieve mAP in a range of 85.1-95.6% on the easy subset, 82.0-94.3% on the medium subset, and 62.9-85.3% on the hard subset. The mean aver- age precision of the proposed algorithm on the easy, medium, and hard validation subsets are 96.3%, 94.9% and 88.2%, respectively, which is 0.7%, 0.6%, and 2.9% higher than the top one.

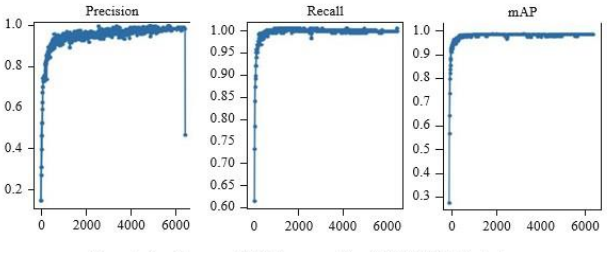
The SR-YOLOv5 proposed in this paper is improved onthe YOLOv5s network, and the image superresolution reconstruction technology is introduced for the secondary detection of small-scale fuzzy faces, deepening the network to make facial features easier to be detected, capturing smalltarget information, and making the network

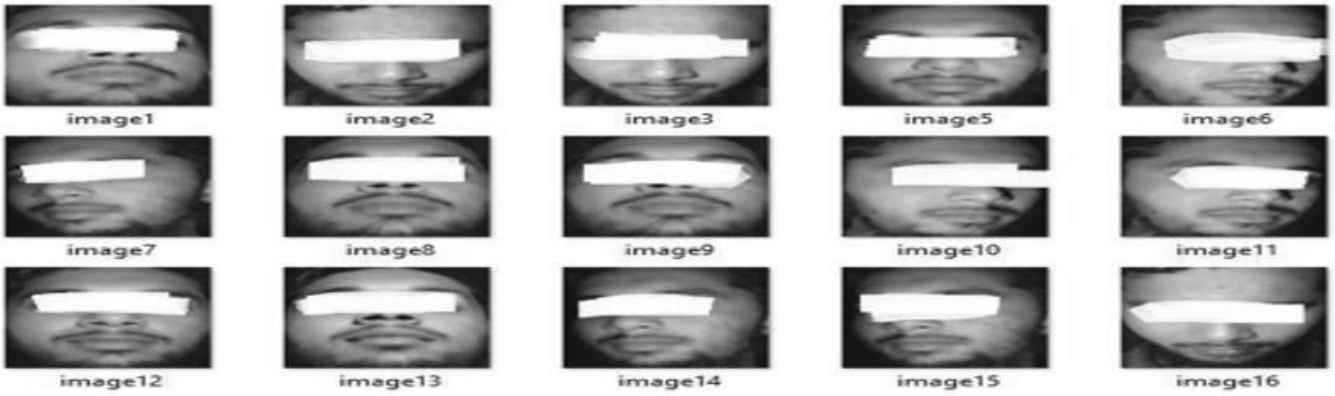
TABLE 4: Performance comparison using different models.

Model	Backbone	AP50	Time/ms
HR	Resnet101	57.5%	198
YOLOv3	Darknet53	57.9%	51
Ours	YOLOv5s-SRGAN	59.8%	75



Face Detection Model	Detection Accuracy (%)	Face Recognition Model	Recognition Accuracy (%)
HaarCascade	75	LBPH	66.67
SSD	85	FaceNet	78.44
MTCNN	95	DeepFace	86.23
YOLOv5	95	ArcFace	94.74





REFERENCES

[1] *Z. Zou, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: a survey," 2019, https://arxiv.org/abs/1905.05055.*

[2] *S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: towards real- time object detection with region proposal networks," Advances in Neural Information Processing Systems, vol. 28, pp. 91–99, 2015.*

[3] *S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in Proceedings of the IEEE*

[4] *conference on computer vision and pattern recognition, pp. 8759– 8768, Salt Lake City, UT, United States, 2018.*

[5] *K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 9, pp. 1904–1916, 2015.*

[6] *K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn.," in Proceedings of the IEEE international conference on computer vision, pp. 2961–2969, Venice, Italy, 2017.*

[7] *W. Liu, D. Anguelov, D. Erhan et al., "Ssd: single shot multibox detector," in Computer Vision – ECCV 2016, Lecture Notes in Computer Science, Springer, 2016.*

[8] *A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "Yolov4: optimal speed and accuracy of object detection," 2020, https:// arxiv.org/abs/2004.10934.*

[9] *C. Ding and D. Tao, "Trunk-branch ensemble convolutional neural networks for video-based face recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 4, pp. 1002–1014, 2018.*

[10] *Z. Tang, G. Zhao, and T. Ouyang, "Two-phase deep learning model for short-term wind direction forecasting," Renewable Energy, vol. 173, pp. 1005–1016, 2021.*

[11] *Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: BMVC (2015)*

[12] *Sun, Y., Chen, Y., Wang, X., Tang, X.: Deep learning face representation by joint identification- verification. In: NIPS (2014)*

[13] *Xu, R., Lin, H., Lu, K., Cao, L., Liu, Y.: A forest fire detection system based on ensemble learning. Forests 12(2),*

[14] *217 (2021). https://doi.org/10.3390/f12020217*

[15] *Liu, W., et al.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe,N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016).*

[16] *https://doi.org/10.1007/978-3-319- 46448-0 2*

[17] *Charan E.S., Khaja, S., Hussain, S.A., Shyam, A.: Student attendance monitoring system using face recognition (2021). https://ssrn.com/abstract=3851056, https:// doi.org/10.2139/ssrn.3851056*

[18] *Alhanaee, K., Alhammadi, M., Almenhali, N., Shatnawi, M.: face recognition smart attendance system using deep transfer learning. Procedia Comput. Sci. 192 (2021).*

[19] *https://doi.org/10.1016/j.procs.2021.09.184. ISSN 1877-0509*

[20] *Sutabri, T., Pamungkur, P., Kurniawan, A., Saragih, R.E.: Automatic attendance system for university student using face recognition based on deep learning. Int. J.Mach. Learn. Comput. 9(5), 668–674 (2019)*

[21] *Alon, A.S., Casuat, C.D., Malbog, M.A.F., Marasigan, R.I., Gulmatico, J.S.: A YOLOv3 inference approach for student attendance face recognition system. Int. J. Emerg. Trends Eng. Res. 8(2), 384–390 (2020). https://doi.org/10.30534/ijeter/ 2020/2482202*