Hindawi
Computational and Mathematical Methods in Medicine
Volume 2021, Article ID 7748350, 9 pages
https://doi.org/10.1155/2021/7748350

Research Article

Effective Face Detector Based on YOLOv5 and Superresolution Reconstruction

Qingqing Xu, Zhiyu Zhu, Huilin Ge, Zheqing Zhang, and Xu Zang

School of Electronic Information, Jiangsu University of Science and Technology, Zhenjiang 212003, China

Correspondence should be addressed to Zhiyu Zhu; zzydzz@163.com

Received 3 September 2021; Revised 18 September 2021; Accepted 28 September 2021; Published 16 November 2021

Academic Editor: Kelvin Wong

The application of face detection and recognition technology in security monitoring systems has made a huge contribution to public security. Face detection is an essential first step in many face analysis systems. In complex scenes, the accuracy of face detection would be limited because of the missing and false detection of small faces, due to image quality, face scale, light, and other factors. In this paper, a two-level face detection model called SR-YOLOv5 is proposed to address some problems of dense small faces in actual scenarios. The research first optimized the backbone and loss function of YOLOv5, which is aimed at achieving better performance in terms of mean average precision (mAP) and speed. Then, to improve face detection in blurred scenes or low-resolution situations, we integrated image superresolution technology on the detection head. In addition, some representative deep-learning algorithm based on face detection is discussed by grouping them into a few major categories, and the popular face detection benchmarks are enumerated in detail. Finally, the wider face dataset is used to train and test the SR-YOLOv5 model. Compared with multitask convolutional neural network (MTCNN), Contextual Multi-Scale Region-based CNN (CMS-RCNN), Finding Tiny Faces (HR), Single Shot Scale-invariant Face Detector (S3FD), and TinaFace algorithms, it is verified that the proposed model has higher detection precision, which is 0.7%, 0.6%, and 2.9% higher than the top one. SR-YOLOv5 can effectively use face information to accurately detect hard-to-detect face targets in complex scenes.

1. 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-CNN [2], PANet [3], SPPNet [4], and Mask R-CNN [5]. The second is one-stage detectors, derived from SSD [6], YOLOv1 to YOLOv5 [7–11], and RetinaNet [12]. The former has higher detection accuracy, but its detection speed is slower, while the latter improves the detection speed and maintains performance. At the same time, the design of face detector gain has achieved the state-of-the-art (SOTA) architecture of general object detectors.

We consider face detector as a special task of general objector detection. General target detection is aimed at multiple categories, while face detection is a dichotomous problem that only detects the face category. In this paper, we design a face detector based on YOLOv5 [11] which has been verified for its superior performance in general target detection tasks. To resolve the challenge of multiscale, small faces, low-light, and dense scenes, we optimized the model with some practical tricks. We also use superresolution reconstruction technology [13] for processing false detection of fuzzy small-scale faces, contributing to richer

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.

2. 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.

2.1. 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 [6], YOLO series [7–11], FPN [23], MMDetection [24], EfficientDet [25], transformer (DETR) [26], and Centernet [22].

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].

2.2. 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 superresolution reconstruction technology to enrich the detailed information of low-resolution images and improve the expression ability of images. Currently, superresolution reconstruction technology [13] based on deep learning is widely used. Among them, the superresolution 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 superresolution image [28]. Based on SRResNet, 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-resolution image of the real sample or the superresolution image of the fake sample.

2.3. YOLO. In the past five years, the YOLO algorithm has been transformed into the fifth version with many innovative ideas from the object detection community. The first three versions including YOLOv1 [7], YOLOv2 [8], and YOLOv3 [9] were all proposed by the author of the original YOLO algorithm, and YOLOv3 [9] is recognized as a milestone with big improvements in performance and speed. We can find multiscale features (FPN) [23], a better backbone network (Darknet53), and replacing the soft-max loss with the binary cross-entropy loss in this algorithm.

YOLOv4 [10] was released by a different research team in early 2020. The team explored a lot of options in almost all aspects of the YOLOv3 [9] algorithm, including the backbone, and what they call bags of freebies and bags of specials. One month later, the YOLOv5 [11] was released by another different research team which significantly reduced size, increased in speed [10], and had a full implementation in Python (PyTorch). It is welcome by the object detection community until now.

YOLOv5 using CSPDarknet as a network of feature extraction, target information is extracted from the input image. The combination of CSP and Darknet formed the CSPDarknet. Figure 2 shows the structure of CSPDarknet. For the input tensor, CSP divides it into two parts in the channel, one part is convoluted once, the other part is convolution-residuals multiple times. The tensor is obtained by multiple convolution-residual operations, and the tensor obtained by one convolution of the previous part is spliced in channel dimensions. CSP makes the output graph retain more network gradient information and maintains the performance of the network while reducing the computational effort.

In the operation, the features of the previous stage can be used as the input of the next stage for up-sampling or down-sampling, and at the same time, the CONCAT with the feature map of the same size in the main part. This pyramid structure makes the high-level feature map integrate the accurate position information of the low level [30] and improves the accuracy of regression.

During detection, the input tensor is divided into $S \times S$ grids, and any one of the grids will be responsible for detecting the target if the center point of the target is located in it. For each grid, there will be B anchors. Specifically, for each anchor frame, $(5 + C)$ values are predicted, with the first 5 values used to regress anchor's center point position, the size of the anchor frame, then to determine whether there is a

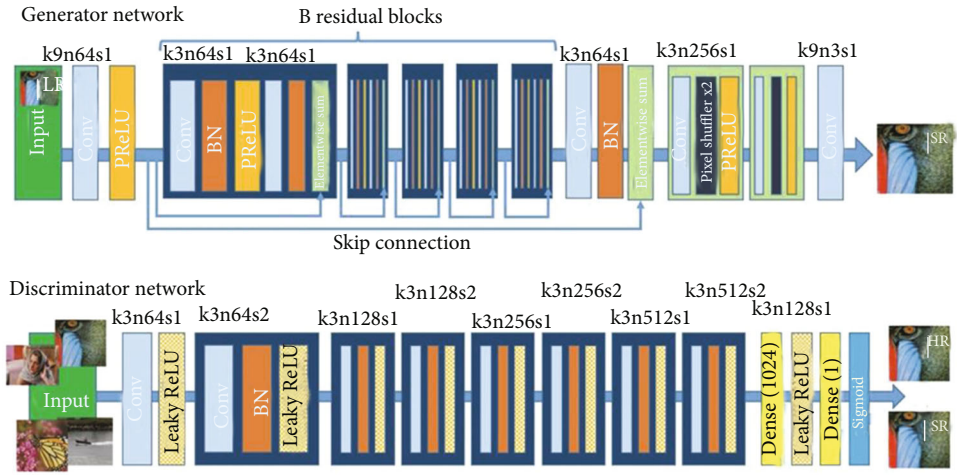


Figure 1: SRGAN network model.

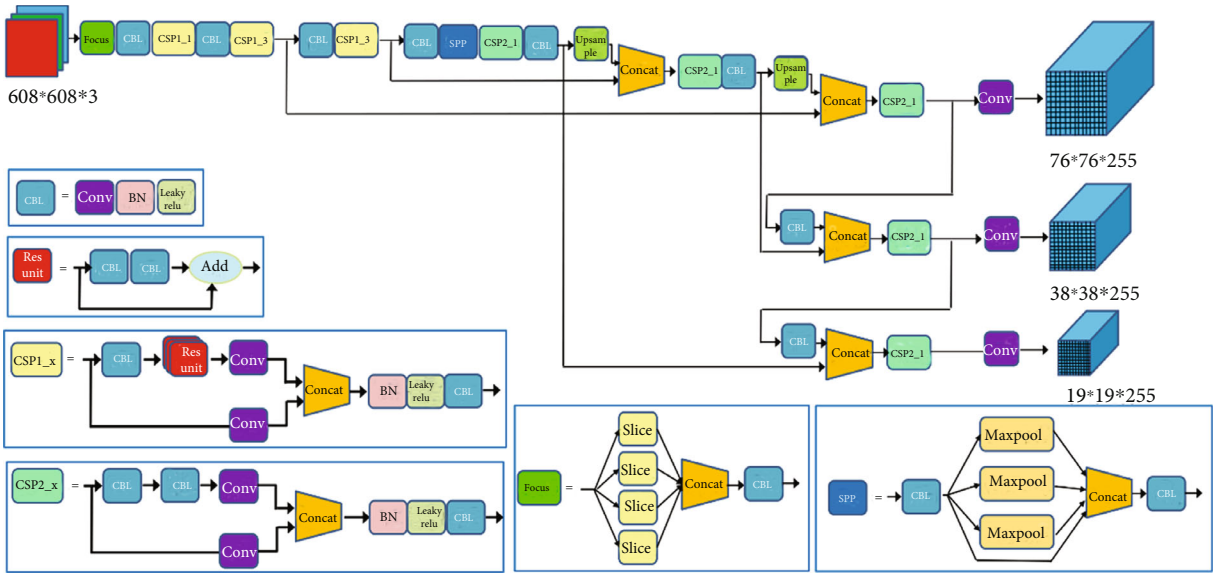


Figure 2: Structures of YOLOv5s.

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:

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

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.

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

$$\mathrm{loss} = l_{\mathrm{box}} + l_{\mathrm{cls}} + l_{\mathrm{obj}}, \tag{2}$$

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

$$l_{\mathrm{box}} = \lambda_{\mathrm{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{i,j}^{\mathrm{ojb}} bj(2 - w_i \times h_i) \left[\left(x_i - x\wedge_i^j\right)^2 + \left(y_i - y\wedge_i^j\right)^2 + \left(w_i - w\wedge_i^j\right)^2 + \left(h_i - h\wedge_i^j\right)^2 \right]. \tag{3}$$

The classification loss function is defined as

$$l_{\text{cls}} = \lambda_{\text{class}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{i,j}^{\text{obj}} \sum_{C \in \text{classes}} p_i(c) \log\left(\widehat{p}_l(c)\right). \tag{4}$$

The confidence loss function is defined as

$$l_{\text{obj}} = \lambda_{\text{noobj}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{i,j}^{\text{noobj}} \left(c_i - c\wedge_l\right)^2 + \lambda_{\text{obj}} \sum_{i=0}^{s^2} \sum_{j=0}^{B} I_{i,j}^{\text{obj}} \left(c_i - c\wedge_l\right)^2, \tag{5}$$

where λ_{coord} is the position loss coefficient, λ_{class} is the category loss coefficient, $\widehat{x}$, $\widehat{y}$is the true central coordinate of the target, and $\widehat{w}$, $\widehat{h}$ is the width and height of the target.

If the anchor box at (i, j) contains targets, then the value $I_{i,j}^{\text{obj}}$ is 1; otherwise, the value is 0. $p_i(c)$ represents the category probability of the target, and $\widehat{p}_l(c)$ is the true value of the category. The length of the two is equal to the total number of categories C.

3. Method

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 superresolution technology on the head and improve the loss function to ensure efficient detection speed.

3.1. SR-YOLOv5

3.1.1. 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 anchor point frames in the high complexity and high recall rate. The results of anchor boxes clustered by the algorithm are shown in Table 1.

3.1.2. Network Architecture

(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.

Table 1: Results of anchor boxes of the training set.

Feature map	Size	Anchor
Predict one	13×13	(43, 59) (76, 89) (178, 234)
Predict two	26×26	(30, 36) (22, 26) (15, 18)
Predict three	52×52	(10, 12) (7, 9) (5, 6)

(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 modules, such as the CS block and the CBS block we proposed.

(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 superresolution reconstruction and face detection again through its coordinate information. Finally, the output of the two-stage face detector is integrated and output.

3.2. 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 box and 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

$$l'_{\text{box}} = 1 - \text{IOU} + \frac{\rho^2\left(b, \widehat{b}\right)}{c^2} + \frac{16}{\pi^4} \frac{\left(\arctan\left(w\wedge/h\wedge\right) - \arctan\left(w/h\right)\right)^4}{1 - \text{IOU} + \left(4/\pi^2\right)\left(\arctan\left(w\wedge/h\wedge\right) - \arctan\left(w/h\right)\right)^2}, \tag{6}$$

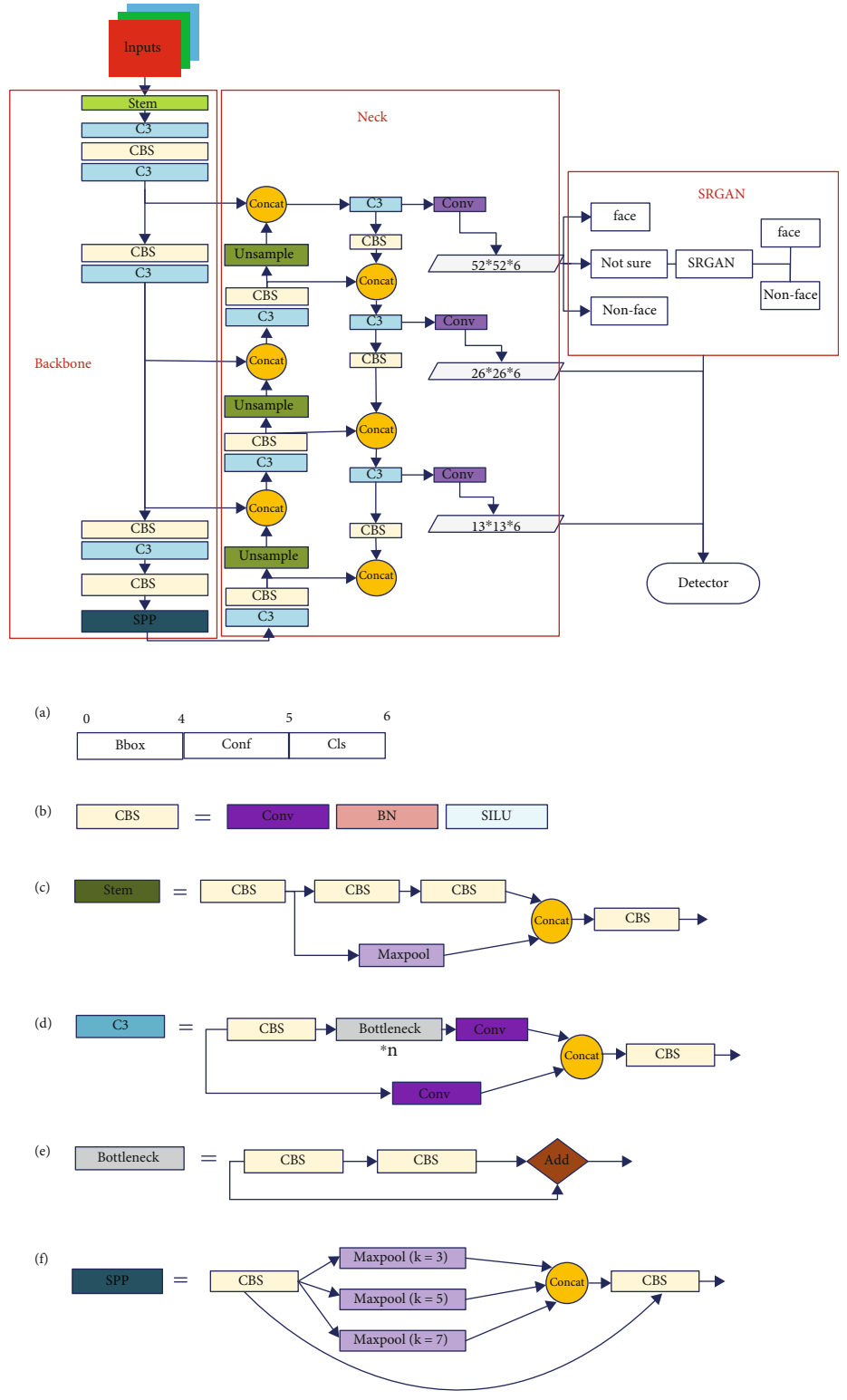


Figure 3: The architecture of improved SR-YOLOv5.

where b, b' represents the center point of the box, ρ represents the Euclidean distance, c represents the diagonal distance of the minimum enclosing rectangle, and $\widehat{w}$, $\widehat{h}$ is the width and height of the target.

In surveillance video images [36], face targets are not only numerous but also stacked, which leads to more than one target in each grid. However, judging by a single threshold often leads to a low recall rate [37]. Therefore, through

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.

4. Experiments

4.1. 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. 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, this benchmark is suitable to verify the effectiveness and performance in realistic scenes. The experimental environment configuration is shown in Table 3.

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

4.2. Training and Testing of SR-YOLOv5 Models

4.2.1. 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.

4.2.2. 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 that the detection effect is still good. (c) The figure fully shows that numerous small faces can be well detected even in a high-density crowd.

4.3. 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 face in 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 in the sample. The accuracy rate is used to evaluate the proportion of the correct face detected in the total face 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.

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{7}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{8}$$

$$F_1 = \frac{2 \times P \times R}{P + R}. \tag{9}$$

The trained model is verified on the validation set, and the recall rate $R = 0.96$, accuracy rate $P = 0.975$, and $F_1 = 96.75$ were obtained from Equations (6), (8), and (9). From the point of view of the score, the proposed algorithm has better performance.

4.4. 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 comparison. 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.

4.5. Comparison of Accuracy of Relevant Algorithms. To demonstrate the effectiveness of the algorithm, some excellent 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 average precision of the proposed algorithm on the easy, medium, and hard validation subsets are 96.3%, 94.9%,

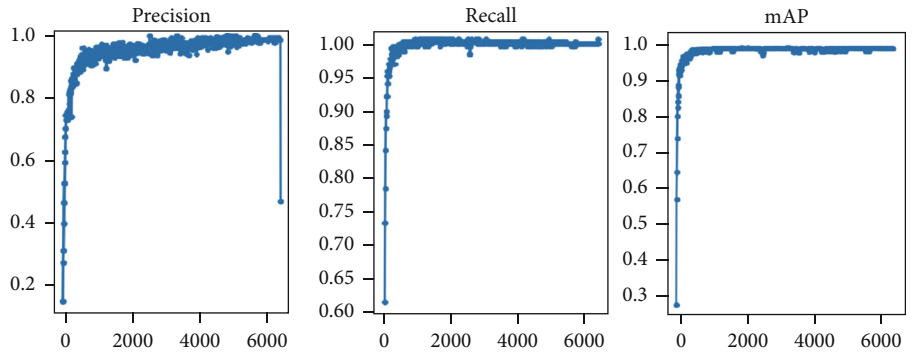
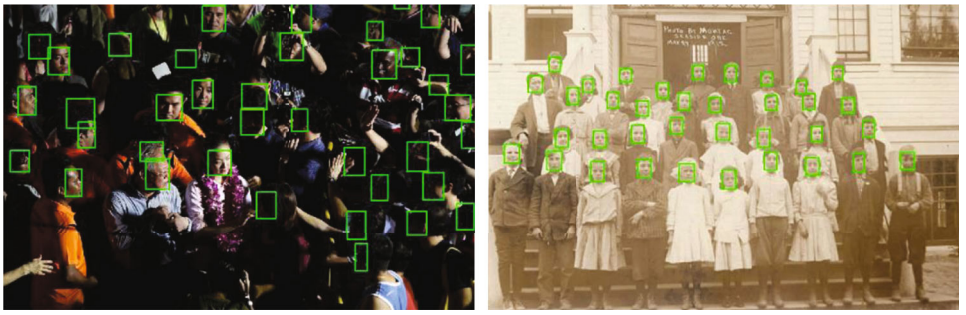


Figure 4: Precision-recall (PR) curves of our SR-YOLOv5 detector.

(a) Blocking faces



(b) Fuzzy scene

(c) Dense small faces

Figure 5: Part of the test results.

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

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 on the 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 small target information, and making the network more accurate when processing complex face and nonface classification

Table 5: Comparison of mAP using different face detection algorithms.

Face detection algorithms	Easy	Medium	Hard
MTCNN	85.1%	82.0%	62.9%
CMS-RCNN	90.2%	87.4%	64.3%
HR	92.5%	91.0%	81.9%
S3FD	93.7%	92.5%	85.9%
TinaFace	95.6%	94.3%	85.3%
Ours	96.3%	94.9%	88.2%

and detection. Through the comparative experiment on the wider face dataset, it is verified that the method used in this paper has higher detection accuracy and better robustness, especially in the hard subset, it has more outstanding performance.

5. Conclusion

To improve the face detection rate of security surveillance scenes with diverse scales in dense face images, this paper proposes a small face detection algorithm suitable for complex scenes. We integrate the image superresolution reconstruction technology into the network structure of the target detection algorithm YOLOv5s. YOLOv5s has a fast detection speed, but its detection accuracy is reduced compared with other SOTA detection algorithms. SRGAN is used to improve the performance of the detection head and then improve the detection accuracy of small-scale fuzzy faces in complex scenes. In the same environment with other face detection algorithms, using the same dataset to carry out comparative experiments, the results confirm the feasibility and superiority of the proposed method.

Data Availability

The data (wider face dataset) used in this research is cited in the references.

Conflicts of Interest

We declare that there are no conflicts of interest to report regarding the present study.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62006102).

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

[4] 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.

[5] 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.

[6] 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.

[7] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, Las Vegas, NV, United States, 2016.

[8] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7263–7271, Honolulu, HI, United States, 2017.

[9] J. Redmon and A. Farhadi, "Yolov3: an incremental improvement," 2018, https://arxiv.org/abs/1804.02767.

[10] 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.

[11] Ultralytics, "Yolov5," 2021, February 2021, https://github.com/ultralytics/yolov5.

[12] J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, and S. Zafeiriou, "Retinaface: single-stage dense face localization in the wild," 2019, https://arxiv.org/abs/1905.00641.

[13] C. Ledig, L. Theis, F. Huszár et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4681–4690, Honolulu, HI, United States, 2017.

[14] C. Xu, Z. Gao, H. Zhang, S. Li, and V. H. C. de Albuquerque, "Video salient object detection using dual-stream spatiotemporal attention," *Applied Soft Computing*, vol. 108, p. 107433, 2021.

[15] D. Qi, W. Tan, Q. Yao, and J. Liu, "YOLO5Face: why reinventing a face detector," 2021, https://arxiv.org/abs/2105.12931.

[16] R. Lienhart and J. Maydt, "An extended set of haar-like features for rapid object detection," in *Proceedings. Iinternational conference on image processing*, Rochester, NY, United States, 2002, September.

[17] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, pp. 886–893, San Diego, CA, United States, 2005, June.

[18] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: application to face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 2037–2041, 2006.

[19] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, Boston, MA, United States, 2015.

[20] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: a benchmark," in *Proceedings of the IEEE international conference on computer vision*, pp. 1116–1124, Santiago, Chile, 2015.

[21] S. Hwang, J. Park, N. Kim, Y. Choi, and I. So Kweon, "Multispectral pedestrian detection: benchmark dataset and baseline," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1037–1045, Boston, MA, United States, 2015.

[22] Z. Cai and N. Vasconcelos, "Cascade r-cnn: delving into high quality object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6154–6162, Salt Lake City, UT, United States, 2018.

[23] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, Honolulu, HI, United States, 2017.

[24] K. Chen, J. Wang, J. Pang et al., "MMDetection: open mmlab detection toolbox and benchmark.," 2019, https://arxiv.org/abs/1906.07155.

[25] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: scalable and efficient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10781–10790, Virtual, Online, United States, 2020.

[26] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Computer Vision – ECCV 202*, Springer, 2020.

[27] S. Yang, P. Luo, C. C. Loy, and X. Tang, "Wider face: a face detection benchmark," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5525–5533, Las Vegas, NV, United States, 2016.

[28] C. Wang, S. Dong, X. Zhao, G. Papanastasiou, H. Zhang, and G. Yang, "SaliencyGAN: deep learning semi-supervised salient object detection in the fog of IoTC," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 4, pp. 2667–2676, 2016.

[29] T. Jiang and J. Cheng, "Target recognition based on CNN with LeakyReLU and PReLU activation functions," in *2019 International Conference on Sensing, Diagnostics, Prognostics, and Control (SDPC)*, pp. 718–722, Beijing, China, 2019, August.

[30] S. Liu, K. Han, Z. Song, and M. Li, "Texture characteristic extraction of medical images based on pyramid structure wavelet transform," in *2010 International Conference on Computer Design and Applications*, pp. 342–345, Qinhuangdao, Hebei, China, 2010, June.

[31] C. Huang, Y. Zong, J. Chen, W. Liu, J. Lloret, and M. Mukherjee, "A deep segmentation network of stent Structs based on IoT for interventional cardiovascular diagnosis," *IEEE Wireless Communications*, vol. 28, no. 3, pp. 36–43, 2021.

[32] S. Elfwing, E. Uchibe, and K. Doya, "Sigmoid-weighted linear units for neural network function approximation in reinforcement learning," *Neural Networks*, vol. 107, pp. 3–11, 2018.

[33] H. Bai, J. Cheng, X. Huang, S. Liu, and C. Deng, "HCANet: a hierarchical context aggregation network for semantic segmentation of high-resolution remote sensing images," *IEEE Geoscience and Remote Sensing Letters.*, pp. 1–5, 2021.

[34] B. Yu and D. Tao, "Anchor cascade for efficient face detection," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2490–2501, 2019.

[35] C. Chen, X. Yang, R. Huang et al., "Region proposal network with graph prior and IoU-balance loss for landmark detection in 3D ultrasound," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5, Iowa City, IA, USA, 2020, April.

[36] 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.

[37] 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.