Jiang *et al.*
EURASIP Journal on Image and Video Processing (2024) 2024:10
https://doi.org/10.1186/s13640-024-00625-4

EURASIP Journal on Image and Video Processing

RESEARCH Open Access

4AC-YOLOv5: an improved algorithm for small target face detection

Bin Jiang[1*], Hongbin Jiang[1], Huanlong Zhang[2], Qiuwen Zhang[3], Zuhe Li[3] and Lixun Huang[1]

*Correspondence:
jiangbin@zzuli.edu.cn

[1] School of Electronics and Information, Zhengzhou University of Light Industry, Zhengzhou 450001, China
[2] College of Electric and Information Engineering, Zhengzhou University of Light Industry, Zhengzhou 450001, China
[3] School of Computer Science and Technology, Zhengzhou University of Light Industry, Zhengzhou 450001, China

Abstract

In real scenes, small target faces often encounter various conditions, such as intricate background, occlusion and scale change, which leads to the problem of omission or misdetection of face detection results. To solve this puzzle, an improved algorithm of small target face detection 4AC-YOLOv5 is proposed. First, the algorithm by introducing a new layer to detect faces at a much smaller size, through the fusion of more shallow information, enhance the network perception of small objects, the accuracy of small target detection is improved; second, to improve the neck structure, to add the adaptive feature fusion network AFPN to replace FPN + PAN, to prevent the large information gap between non-adjacent Level to some extent, and to fully retain and integrate different scale characteristic information; and finally, improve the C3 module and propose a new multiscale residual module C3_MultiRes. Improving the expressive power of the network by introducing a multibranched structure and gradually increasing resolution somewhat reduces the complexity of the model calculation. The experimental results show that the precision of the improved model reached 94.54%, 93.08% and 84.98% in easy, medium and hard levels of WiderFace data set, respectively, and the results of detection are better than the original network. 4AC-YOLOv5 can meet the requirements of small target face detection in complex environment.

Keywords: Small target face, Face detection, YOLOv5 algorithm, Multi-scale

1 Introduction

In the last few years, object detection [1] has always been a core issue in the computer vision research field, and its task is to distinguish objects in an image or video from other regions of interest, determine whether the target exists, and to confirm target classes and positions. With the feature extraction technology based on deep neural networks being widely used in computer vision tasks, object detection technology has made a major breakthrough. At present, according to the rule of whether to generate candidate regions, algorithms of target detection are usually divided into two-stage algorithms [2] and one-stage algorithms [3]. The former mainly uses deep learning method and extracts the depth features of the corresponding regions by extracting candidate regions, representing the algorithms R-CNN [4], SPP-net [5], Fast R-CNN [6], Faster R-CNN [7] and R-FCN [8], etc. Although the two-stage algorithm maintains a large advantage in

accuracy, but, in the real-time scene detection, the speed of detection is often unsatisfactory. In contrast, the latter does not need to select the target candidate region, and directly regression the distribution probability and location coordinates of the target, which can obtain a great improvement in speed. Due to the advantages of high detection accuracy and less time consuming, the one-stage algorithm has gradually turn into a study hotspot in the domain of target detection, and the representative algorithms include RetinaNet [9], YOLO [10–16] series and SSD [17] series.

Among many target detection technologies, face detection [18] is the earliest technology put into practical application. For instance, in the domain of security, face detection can help the monitoring system effectively lock the criminal suspect; in terms of mobile payment and secure login, face detection has also become an important part of face unlocking technology, which can be seen that face detection not only facilitates people's lives, but also provides a guarantee for life safety. However, in real scenes, the problem of small target face detection [19] is more prominent. When the security camera takes a long-range picture or a crowd gathering place, there are often some small size, low resolution faces, their visualization information is less, it is difficult to let the detection model obtain more discriminating features, resulting in the problem of missing the target face. Moreover, when the background information around the small target face is similar to the color and texture of the face, detection error easily happened. The faces of small targets moving in outdoor scenes are more likely to be blocked by other objects, thus increasing the complexity of detection. In response to these difficulties, Zhu et al. [20] proposed a CMS–RCNN model based on Faster RCNN, which fuses context information into face detection and utilizes multi-scale PRN to make the model have stronger feature expression capability, but the model detection is relatively slow. Deng [21] et al. proposed a pixel-based Face positioning method RetinaFace, which adopted a multi-task learning tactics to predict the 3D position and correspondence of the face score, the face box, the key points of five people's faces, and each face pixel at the same time, and realized SOTA on the Wider Face dataset. Li et al. [22] proposed a DSFD algorithm, which uses FEM module to enhance receptive field and relies on progressive anchor loss function PAL to obtain high-resolution information. The improved anchor points can better match faces, but the model is larger and the calculation speed is slower. Qi et al. [23] proposed a new face finder model, YOLO5Face, on the basis of YOLOv5 model, meanwhile proposed different types of face finders according to the needs of different applications. The Focus module was replaced with the Stem module, which improved the model feature expressiveness and reduced the complexity of the calculation. The loss function is replaced with Wing Loss to achieve rapid convergence of the network, but this method is aimed at detecting large faces, and the detection performance of small faces needs to be further improved.

Although the above methods have good research results face detection field, they cannot maintain a balance between detection accuracy for small targets face and the calculation and parameters amount of model. For this problem, this paper improved the method based on YOLOv5s, and proposed a small target face detection model 4AC-YOLOv5. First, by introducing a new layer to detect faces at a much smaller size, the model can better adapt to targets of various sizes, and can more accurately identify and locate small targets in complex scenes with large size changes, the accuracy of small

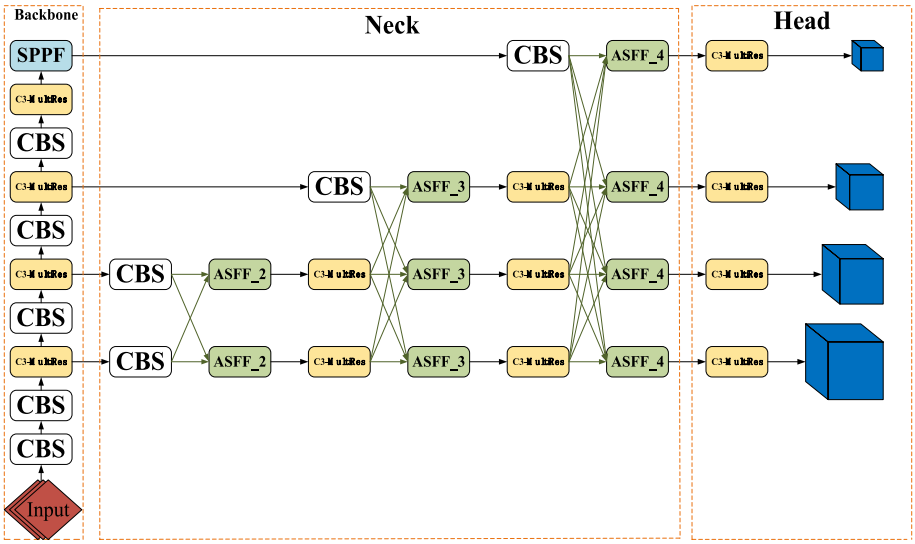

Fig. 1 Network structure of 4AC-YOLOv5

targets detection is improved. Second, the asymptotic feature pyramid network (AFPN) replaces FPN + PAN to prevent the large information gap between not adjacent level, which leads to information loss or degradation in multilevel transmission. Because only normal convolution components are introduced, the algorithm improves the detection precision rate, meanwhile reduces the calculation and parameters amount of model. Finally, a multi-scale residual module C3_MultiRes is proposed to process features of different scales through multiple parallel branch structures, so as to enlarge the receptive field and make the model better understand the semantic information of the whole image. The improved model reduces parameters amount compared with the benchmark model, the detection of small target face under complex background accuracy has also been improved. Figure 1 shows the improved 4AC-YOLOv5 network model.

2 YOLOv5

As one of the models widely recognized by researchers in object detection, YOLOv5 has made some improvements and optimizations based on yolov4, so that it can better balance detection accuracy and detection speed. Compared with YOLOv7 and YOLOv8 models, YOLOv5 has low performance requirements on devices and fast computing speed, and is more suitable for deployment on intelligent terminal devices. The YOLOv5 algorithm has versions of four different size models, YOLOv5s, YOLOv5l, YOLOv5x and YOLOv5m, among which, YOLOv5s is the smallest model in this series, which has the fewest layers and the smallest computational complexity, and performs best on devices with limited computing resources. Therefore, YOLOv5s algorithm is selected in this text for further research.

The main structure of YOLOv5s includes the following parts: input, backbone, neck, and head. The input is usually used to receive image data and preprocess it. It normalizes, scales and clips the original image to adapt to the input requirements of the model. The backbone network mainly adopts the CSPDarknet structure. It is responsible for extracting the features of the image, gradually transforming the image from low-level

pixel information to high-level semantic features through a series of convolutional layers and residual blocks. Backbone extracts rich feature information at different scales for subsequent target detection tasks; the neck network adopts FPN + PAN [24] structure to integrate feature maps of different scales and carry out feature fusion, compared with YOLOv4 neck structure, YOLOv5s adopts CSP2 structure, which further enhances the capability of network feature fusion. The function of the channel block squeeze (CBS) module is to enhance the feature extraction capability and improve the accuracy of target detection. Three detection layers of different sizes are designed in the detection head module, using GIoU as bounding box regression loss function, and the best target bounding box is found by enhanced NMS algorithm.

3 Related technologies

3.1 Design small target detection layer

The YOLOv5 model uses the classic FPN + PAN structure, and the features are further extracted and fused through the fusion and matching of rich location information and feature maps with different resolutions, thus improving the detection accuracy of targets. In terms of detection layer, its model is detected and output on the feature layer of large, medium and small sizes, respectively. However, as we all know, small targets usually have small size and low resolution. Compared with large targets, small targets are more likely to suffer from missed detection and false detection, so the detection layer needs to have higher positioning accuracy to find targets more accurately. For the problem that small targets are difficult to extract, 160×160 small target detection layer is introduced base on 80×80, 40×40, 20×20 prediction feature output layer of the benchmark network. Compared with the other three output layers, the small target detection layer fuses shallower information, making the location information relatively rich, enhancing the network's perception of small targets, and being able to more accurately identify and locate small targets in complex scenes with large target size variations, thus improving the detection accuracy. Meanwhile, because the traditional fixed-size anchor frame may not match the size of the target object, the small target may lead to missing or false detection, K-means clustering algorithm [25] is added to regenerate 12 different preset anchor frame sizes for the four detection layers, which can better adapt to the requirements of specific scenes and reduce the deviation between the target and the anchor frame. Improve the precision and accuracy of target detection.

3.2 Improved neck network

Multi-scale feature fusion algorithm [26] can grasp face information by using features at different scales in face detection, and then fuse the extracted face feature information to further optimized the algorithm performance. One of the classic networks is the combination of FPN (feature pyramid network) and PAN (path aggregation network) structures adopted in the Neck structure of YOLOv5. FPN can extract rich feature information at different scales through top-down path linking, while PAN is a structure that aggregates features at different scales through bottom-up path linking, so as to integrate feature maps with different resolutions. However, due to the increase in resolution, some details in the original feature map may not be fully recovered, which weakened information transfer and the retention of important features

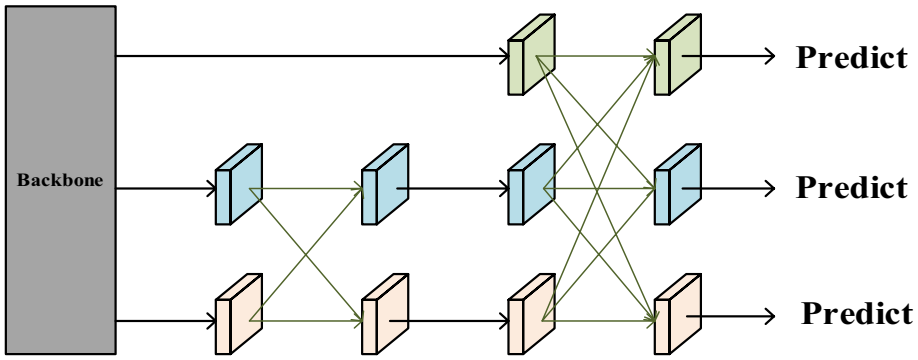

Fig. 2 AFPN network structure

Table 1 Sampling stride comparison data on Wider Face data set

Sampling stride	Easy	Medium	Hard	Params (M)	Flops (G)
4 × 4	94.60	93.05	83.14	**6.327**	**5.602**
8 × 8	**94.74**	**93.12**	**84.75**	7.139	7.291

Bold value is that it is the best value in the comparison result

between non-adjacent levels. Yang et al. [27] proposed a progressive feature pyramid network (AFPN) to facilitate semantic interaction between non-adjacent layers. In the AFPN structure, low-level features are mutually fusion with high-level features through up-sampling and residual connection, so that high-level features are gradually introduced into the fusion process; in this way, features of different scales can be fully utilized. Meanwhile, considering the possible multi-object information conflict in feature fusion process, AFPN also dynamically adjusts the weight of features through adaptive feature fusion operation, so that targets of different scales can get proper attention, thus improving the detection performance of small targets.

The AFPN structure is shown in Fig. 2. In Backbone network, a top-down feature extraction process is adopted to extract the last feature layer from each feature layer to generate a set of new features of different scales, which are expressed as {C2,C3,C4,C5}. In the subsequent feature interaction process, low-level features C2 and C3 are first input into the feature pyramid, and the semantic gap between them is reduced through adaptively spatial feature fusion (ASFF). Because C3 and C4 are adjacent hierarchical features, the semantic gap between non-adjacent C2 and C4 is reduced. Then, C4 is added for adaptive spatial feature fusion, and finally C5 is added. To align the fusion dimensions, 2×2 convolution with stride of 2, 4×4 convolution with stride of 4 and 8×8 convolution with Stride of 8 are applied to achieve 2, 4 and 8 downsampling, respectively; the same method was used for upsampling. In the YOLOv5 model, only three levels of features are used, this is shown in Table 1, this paper appends a new detection layer, and eight times of up-sampling and down-sampling are used to better improve the small targets detection performance.

In order to features of different scales can be fully utilized, the adaptive feature fusion method ASFF is adopted in AFPN [28]. As shown in Fig. 3, this method can not only retain details in low-level feature maps, but also deliver semantic information in

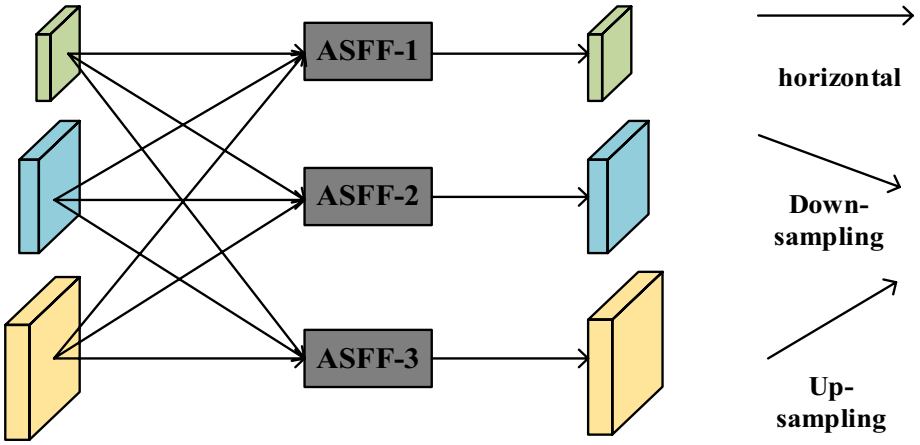

Fig. 3 Adaptive feature fusion operation

high-level feature maps; according to the feature quality of each level and its contribution to target detection, the weights of features of different levels are dynamically adjusted and thus better captures target details and context information. It can be seen from formula (1) that this method integrates the features of three levels. Let $c_{ij}^{n\to l}$ represent the eigenvector from Level n feature resize to Level l at position (i,j). The eigenvectors of these different layers are then multiplied by the weight parameters α_{ij}^{l}, β_{ij}^{l} and γ_{ij}^{l} and add them to get the resulting feature vector, which is represented as d_{ij}^{l}:

$$\mathrm{d}_{ij}^{l} = \alpha_{ij}^{l} \cdot c_{ij}^{1\to l} + \beta_{ij}^{l} \cdot c_{ij}^{2\to l} + \gamma_{ij}^{l} \cdot c_{ij}^{3\to l} \tag{1}$$

Among them, $\alpha_{ij}^{l} + \beta_{ij}^{l} + \gamma_{ij}^{l} = 1$. Since the addition method is adopted, the features of each Level are up-sampled or down-sampled, so that they have the same size as the features of other levels, and the channels number is adjusted by convolution operation, so that each Level also has the same number of feature channels. Considering the difference in the number of fusion features at each stage of AFPN, an adaptive spatial fusion module with the number of specific stages is realized. In the article, the feature fusion of 2, 3 and 4 levels is adopted, so that it can extract the features of small targets at a deeper level.

3.3 C3_MultiRes

The C3 module used in YOLOv5 consists of a series of convolutional layers, with having multiple branches and cross-layer connections, which mainly function is to extract the characteristics of input data and reduce the parameters amount and calculation amount of convolutional layers to improve the efficiency of the model. In 2021, Gao et al. [30] proposed an improved multi-scale residual module Res2Net based on the ResNet [29] model, whose structure was mainly improved by replacing the 3×3 convolution in the middle of the original ResNet model with a multi-scale residual convolution. Compared with ResNet, Res2Net obvious advantage is that it can effectively capture the different scales of space information. In ResNet, each convolutional layer can only capture

features within a certain range, while Res2Net can effectively expand the range of the receptor field by decomgenerating the convolution into multiple sub-modules and connecting these sub-modules. Thus, richer feature information can be captured. In addition, Res2Net can also improve network performance without increasing network depth, so it has certain advantages in some tasks with limited computing resources. In this paper, a multi-scale residual C3_MultiRes module is proposed by combining Res2Net module and C3 module. Figure 4 shows its structure. First, the input features undergo a 1×1 convolution to extract low-level features, and channel number is $n(n = s \times w)$, then the feature map is sent into s residual blocks, the number of channels in each residual block is w, and X_m is used to represent each residual block (because x is divided into four parts in the figure, so $s = 4$), and then, the 3×3 convolution of 3 w channels is used to replace the original single 3×3 convolution in the bottleneck. Where X_1 is divided into two branches after 3×3 convolution, one to Y_1 and one to X_2. X_2 combines with the information from it, and repeats the above operation to the end of X_3 after 3×3 convolution. To avoid the increase of parameters amount, X_4 is passed directly to Y_4 without any operation. After processing all sub-feature graphs, they are merged into 1×1 convolution to obtain multi-scale residual information. As shown in formula (2), assuming X_m, where $m \in \{1, 2, 3, 4\}$, K_m () represents the convolution of 3×3, and the output is represented as Y_m:

$$Y_{\mathrm{m}} = \begin{cases} K_m(X_m) & m = 1 \\ K_m(X_m + Y_{m-1}) & 1 < m \leq 3 \\ X_m & m = 4 \end{cases} \tag{2}$$

This multi-scale residual structure [31] improves the expressiveness of the model by introducing a multi-branch structure and progressively increasing resolution, which makes the network better able to process features of faces at different scales and resolutions. Meanwhile, it also ensures the effective transmission of features and the full use of

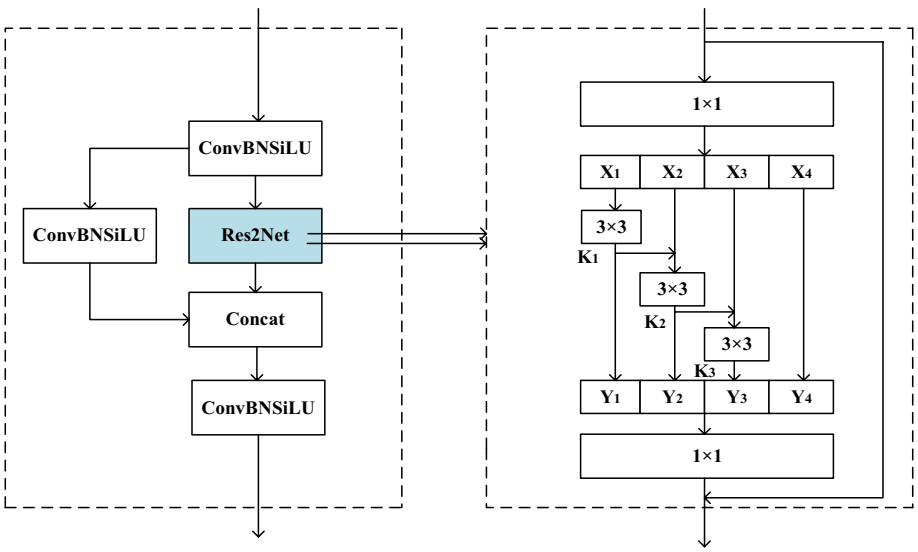

Fig. 4 C3_MultiRes module structure

information, which means that the model can better adapt to faces of different scales, gestures and complex backgrounds for small target face detection, thus improving the detection accuracy of small target faces.

4 Experimental results and analysis

4.1 Data set

4.1.1 Wider face data set

In the article, the experiment using Wider Face [32] data set for training and evaluation our model. As an authoritative face detection data set, Wider Face has a large amount of data and high scene complexity, including more than 32,000 images and nearly 400,000 high-precision labeled faces. Each face is marked with detailed information such as illumination, occlusion, and posture, and the data set is divided into 61 categories according to the type of event scene, and the images of each type of event are divided according to the proportion of training set 40%, verification set 10%, and test set 50%. Divide the face into three scales based on the height of the image: small (10–50 pixels), medium (50–300 pixels), large (more than 300 pixels). According to the detection rate of EdgeBox, the detection difficulty of WiderFace is set to three levels: easy, medium and hard. The difficult data set contains a large number of small target face image data, which is exactly in line with the requirements of this paper for small target face detection. It is undoubtedly challenging to achieve high detection accuracy on the difficult subset.

4.1.2 FDDB data set

FDDB (face detection data set and benchmark) [33] is an industry-recognized standard data set for face detection algorithm evaluation. The data set contains 5171 labeled face images involving various poses, expressions and lighting conditions, and the evaluation results can mirror the performance of the algorithm under different conditions.

4.2 Experimental environment

Operating system: Windows 11 Professional.

Hardware configuration: Intel(R) Core(TM) i5-11400@2.60 GHz CPU, NVIDIA GeForce RTX 3060Ti GPU.

The compilation environment is Ubuntu16.04.7, Python: 3.7.1, Pytorch1.12.0+cu113, CUDA11.3.

Experimental training parameter: img-size is 640×640, batch-size is 16 and epochs is 250.

4.3 Evaluation index

In the experiment, recall (R), precision (P), average precision (AP), mean average precision (mAP), parameter number (Params) and computational effort (FLOPs) were used to evaluate the detection performance of the model on the data set. The formulas are shown in formulas (3–6), respectively:

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{3}$$

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{4}$$

$$\mathrm{AP} = \int_0^1 P(R)\mathrm{d}R \tag{5}$$

$$m\mathrm{AP} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{AP}_i \tag{6}$$

where TP is that the model correctly detects the real positive example, FP is that the model incorrectly determines the non-existent positive example as the positive example, and FN is that the model incorrectly fails to detect the real positive example, n indicates target categories amount that need to be detected, mAP is the AP average of multiple categories, higher values indicate better detection performance. The number of parameters is used as the criterion to evaluate the complexity of the model. When parameters amount is small, it means the less computing resources are used, and the calculation amount can measure the computing time of the model.

4.4 Ablation experiment

In the same experimental environment and training parameters, the improvement effect of the 4AC-YOLOv5 compared with the benchmark model was verified by parameter number and calculation amount of the model and the precision of the WiderFace dataset at three levels: Easy, Medium and Hard. Table 2 shows the experimental results.

Table 2 shows that the detection accuracy of the initial YOLOv5 model in Easy, Medium and Hard levels is 94.43%, 92.82% and 83.08%, respectively. The number of parameters is 7.013 M, and the calculation amount is 5.909G. When the small target detection layer is added (this module is defined as a small layer), the detection accuracy of simple and medium levels is increased by 0.58% and 0.52%, respectively, but the detection accuracy of difficult levels is increased by 1.53%, which indicates that adding a small target detection layer can enhance the localization accuracy of small targets and extract the details of the small target more efficiently, so as to detect the detection performance of small target faces.

Table 2 Ablation study on Wider Face data set

Model	Small layer	AFPN	C3_MultiRes	Easy	Medium	Hard	Params (M)	Flops (G)
YOLOv5				94.43	92.82	83.08	7.013	5.909
Improvement one	√			95.01	93.34	84.61	7.156	6.935
Improvement two		√		94.60	93.05	83.14	6.327	5.602
Improvement three			√	94.32	92.73	83.32	6.496	5.413
Improvement four	√	√		94.74	93.12	84.75	7.139	7.291
Improvement five		√	√	94.59	92.94	83.29	5.983	5.229
Improvement six	√		√	94.09	92.67	84.51	6.628	6.348
4AC-YOLOv5	√	√	√	94.54	93.08	84.98	6.791	6.860

When the AFPN and C3_MultiRes improvements are made, respectively, although the detection accuracy was slightly improved, compared with the benchmark model, the parameters amount and calculation amount are reduced by 9.8% and 5.2%, respectively, for AFPN and 7.4% and 8.4%, respectively, for C3_MultiRes, indicating that the improvement point had improved the detection accuracy, meanwhile, the speed of model detection is significantly improved.

Finally, all the improvement points are combined, and compared with the baseline model, improves by 0.11%, 0.26% and 1.9% on the easy, medium and difficult levels, respectively, while the parameters amount is reduced by 3.1% compared with the models previously proposed, the 4AC-YOLOv5 model proposed in the article can reduce the model complexity and also can significantly increase the detection accuracy, enables both to achieve balance, and significantly improve the level of small face difficulty, meeting the requirements of small target face detection.

4.5 Contrast experiment

To fully confirmed that the model in the article is better able to detect small target faces, 4AC-YOLOv5 and current mainstream face recognition models were compared in WiderFace data set and the same training environment. Table 3 shows the comparison results.

As can seen in Table 3 is that compared with the two-stage algorithm CMS−RCNN, the improved model in the article has an improvement of 20.68% in Hard level, 4.34% and 5.68% in Easy level and Medium level, respectively. Compared with RetinaFace, ResNet-10GF [34] and DSFD, which are all one-stage algorithms, although the improvement of our model is not obvious at Easy level, it is significantly improved at Medium level and Hard level. In particular, the Hard level increased by 20.81%, 4.56% and 13.59%, respectively, indicating that the new model can play a better detection effect on small target faces, and the number of parameters decreased by 76.9%, 0.8% and 94.3%, and the calculation amount decreased by 81.7%, 32.6% and 97.3%, respectively.

For YOLOv5, YOLO5Face and YOLOv7-Tiny-Face, which belong to the same YOLO algorithm, the Hard level of the new model is increased by 1.9%, 1.69% and 2.80%, respectively, and the number of parameters decreased by 3.1%, 4.0% and 13.4%, respectively. Although the amount of computation is slightly increased, it still meets the requirements of small target face detection.

Table 3 Contrast study on Wider Face data set

Model	Easy	Medium	Hard	Params (M)	Flops (G)
CMS−RCNN [20]	90.2	87.4	64.3	/	/
RetinaFace [21]	94.92	91.90	64.17	29.50	37.59
ResNet-10GF [34]	94.69	92.90	80.42	6.85	10.18
DSFD [22]	94.29	91.47	71.39	120.06	259.55
YOLOv5	94.43	92.82	83.08	7.013	5.909
YOLO5Face [23]	93.86	92.04	83.29	7.075	5.791
YOLOv7-Tiny-Face	94.72	92.63	82.18	7.843	6.276
4AC-YOLOv5(ours)	94.54	93.08	84.98	6.791	6.860

In conclusion, the model proposed in the article not only has good property in model size and computation amount, but also achieves high detection accuracy in Medium level and Hard level of data set, meeting the requirements of light model, fast computation and high precision. Moreover, the improvement of detection accuracy at the Hard level proves that our model can solve the question that small target faces is difficult to extract in complex scenes.

Evaluation experiments on FDDB face detection data set further demonstrate the robustness of the improved model. First, the model obtained after training on the Wider Face data set was tested and evaluated on the FDDB data set, and the evaluation results were all true positive rate (TPR) when the false positive (FP) was 1000. As shown in Fig. 5a, it can be able to see that the TPR of the model in this paper reached 0.969 without training, which decreased by only 0.004 compared with the benchmark model. This is because the model in the article mainly targets small faces that are difficult to detect. However, the face targets in the FDDB data set are generally large in size, which is not very friendly to the model results we trained on the faces of small targets, so the improvement is not very obvious.

In view of this situation, the benchmark model and the 4AC-YOLOv5 model were used in the article to retrain and evaluate the FDDB data set. Figure 5b shows the evaluation results. It can be able to see that when the false positive (FP) of the trained 4AC-YOLOv5 model is 1000, the true positive rate (TPR) reaches 0.990, which increased by 0.002 compare with the benchmark YOLOv5 model, and still maintained a higher detection accuracy compared with the current mainstream model. This shows that the model in the article can still maintain a higher detection performance for large face targets with complex backgrounds and different illumination levels, as well as some small face targets. Therefore, it can also prove that the improved model has strong robustness and can adapt to face detection tasks in scenes with different target sizes and high background complexity.

To more intuitively reflect the effectiveness and efficiency of the improved algorithm, this paper compares the detection result graph of the improved 4AC-YOLOv5 algorithm with the benchmark model YOLOv5, as shown in Fig. 6.

In Fig. 6, the left side is the test result of the improved 4AC-YOLOv5 model, and the right side is the test result of the YOLOv5 model. As can be seen from the comparison

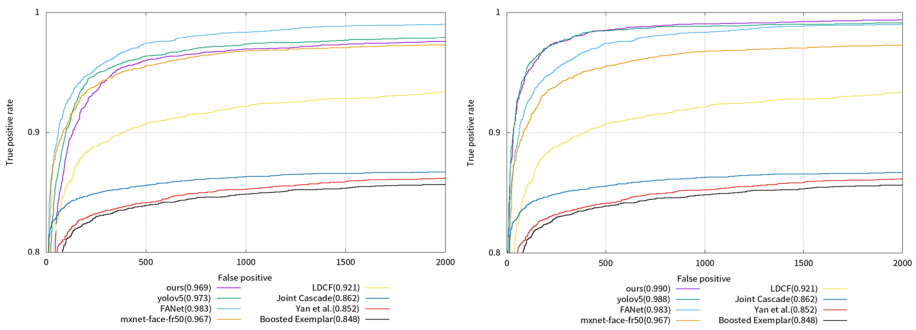

Fig. 5 Robustness study on FDDB data sets. Among them, **a** is a comparison diagram of evaluation results on FDDB data set after the model is trained on Wider Face data set, and **b** is a comparison diagram of the results of the model training and evaluation on the FDDB data set

Fig. 6 **a** is the comparison of omission detection effects when facing small target faces, **b** is the comparison of detection effects when facing partially occluded faces and complex background, and **c** is the comparison of detection effects when facing a large number of small target faces and complex background. And the left side is the test result of the improved 4AC-YOLOv5 model, and the right side is the test result of the YOLOv5 model

diagram in Fig. 6a, the improved 4AC-YOLOv5 model can effectively avoid the error of face omission detection; in Fig. 6b, facing the problem of face semi-occlusion and background blur, our model can still accurately detect faces; in Fig. 6c, when there are a large number of small target faces and complex backgrounds, our model can detect more and more accurate faces than the benchmark model. In summary, the 4AC-YOLOv5 algorithm has better comprehensive performance, and can better solve the problem of false detection and leakage detection in small target face detection under complex background compared with other algorithms.

5 Conclusion

To improve the detection accuracy of small face targets and reduce the false detection and omission of small face targets, the article proposes a small target face detection model 4AC-YOLOv5, which the algorithm introduces the small target detection layer base on the YOLOv5 algorithm. At the same time, adaptive feature fusion network AFPN and multi-scale residual module C3_MultiRes are introduced to obtain different scale feature information, and the small target face detection accuracy is further improved. Confirmed by experimental results, the detection accuracy of the improved algorithm reaches 94.54%, 93.08% and 84.98%, respectively, in the levels of Easy, Medium and Hard of Wider Face data set, which are all higher than the benchmark model, the parameters amount is also reduced. The model is simpler and has stronger comprehensive performance. The validation on FDDB data set further confirms that the algorithm in the article has strong robustness. Finally, in the field of multi-feature fusion and face recognition, a new multi-feature fusion and decomposition (MFD) framework for age-invariant face recognition has emerged. In the future,

we will consider applying this model to age-invariant face recognition and integrating it into embedded devices and carry out applications in more outdoor scenarios.

Author contribution
BJ and HJ conceived the algorithm, designed the experiments, analyzed the results, and wrote the paper; ZL and LH wrote the codes and performed the experiments; HZ and QZ were in charge of the overall research and contributed to the paper writing. The author(s) read and approved the final manuscript.

Funding
National Natural Science Foundation of China, 61702464, Bin Jiang, 62272423, Huanlong Zhang, 61702462, Zuhe Li, Henan Provincial Science and Technology Research Project, 222102210103, Bin Jiang, 232102211014, Qiuwen Zhang, 222102210010, Zuhe Li, 232102211006, Zuhe Li, 232102210044, Zuhe Li, Basic Research Projects of Education Department of Henan, 21zx003, Qiuwen Zhang, the Key 365 projects Natural Science Foundation of Henan, 232300421150, Qiuwen Zhang.

Data availability
Data of this study are available from the corresponding author on reasonable request.

Declarations

Competing interests
The authors declare that they do not have any competing interests.

Received: 17 January 2024 Accepted: 22 April 2024
Published online: 20 May 2024

References

1. Y. Liu, L. Geng, W. Zhang et al., Survey of video based small target detection. J Image Graph 9(4), 122–134 (2021)
2. L. Du, R. Zhang, X. Wang, Overview of two-stage object detection algorithms. J Physics Conf Ser 1544(1), 012033 (2020)
3. Y. Zhang, X Li, F. Wang, et al. A comprehensive review of one-stage networks for object detection. 2021 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC). IEEE, 2021: 1–6.
4. R. Girshick, J. Donahue, T. Darrell, et al. Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580–587.
5. K. He, X. Zhang, S. Ren et al., Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. 37(9), 1904–1916 (2015)
6. R. Girshick. "Fast r-cnn", Proceedings of the IEEE international conference on computer vision. 1440–1448 (2015)
7. S. Ren, K. He, R. Girshick, et al. Faster r-cnn: towards real-time object detection with region proposal networks. Adv. Neural Inf. Process Syst. 28 (2015).
8. J. Dai, Y. Li, K. He, et al. R-fcn: Object detection via region-based fully convolutional networks. Adv. Neural Inf. Proc. Syst. 29 (2016).
9. T.Y. Lin, P. Goyal, R. Girshick, et al. Focal loss for dense object detection. Proceedings of the IEEE international conference on computer vision. 2980–2988 (2017).
10. J. Redmon, S. Divvala, R. Girshick, et al. You only look once: unified, real-time object detection. Proceedings of the IEEE conference on computer vision and pattern recognition. 779–788 (2016).
11. J. Redmon, A. Farhadi. YOLO9000: better, faster, stronger. Proceedings of the IEEE conference on computer vision and pattern recognition. 7263–7271 (2017).
12. J. Redmon, A. Farhadi. Yolov3: an incremental improvement. arXiv preprint arXiv:1804.02767 (2018).
13. A. Bochkovskiy, C.Y. Wang, H.Y.M. Liao. Yolov4: optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020).
14. C. Li, L. Li, H. Jiang, et al. YOLOv6: a single-stage object detection framework for industrial applications. arXiv preprint arXiv:2209.02976 (2022).
15. C. Y. Wang, A. Bochkovskiy, H. Y. M. Liao. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 7464–7475 (2023).
16. Z. Ge, S. Liu, F. Wang, et al. Yolox: exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430 (2021).
17. W. Liu, D. Anguelov, D. Erhan, et al. Ssd: single shot multibox detector. Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer International Publishing, 2016: 21–37.
18. A. Kumar, A. Kaur, M. Kumar, Face detection techniques: a review. Artif. Intell. Rev. 52, 927–948 (2019)
19. D. Mamieva, A.B. Abdusalomov, M. Mukhiddinov et al., Improved face detection method via learning small faces on hard images based on a deep learning approach. Sensors 23(1), 502 (2023)
20. C. Zhu, Y. Zheng, K. Luu, et al. Cms-rcnn: contextual multi-scale region-based cnn for unconstrained face detection. Deep Learn. Biometric. 57–79 (2017).
21. J. Deng, J. Guo, Y. Zhou et al. Retinaface: single-stage dense face localisation in the wild. arXiv preprint arXiv:1905.00641 (2019).

22. J. Li, Y. Wang, C. Wang, et al. DSFD: dual shot face detector. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 5060–5069 (2019).
23. D. Qi, W. Tan, Q. Yao, et al. YOLO5Face: why reinventing a face detector. European Conference on Computer Vision. Cham: Springer Nature Switzerland. 228–244 (2022).
24. S. Liu, L. Qi, H. Qin et al. Path aggregation network for instance segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition. 8759–8768 (2018).
25. M. Ahmed, R. Seraj, S.M.S. Islam, The k-means algorithm: a comprehensive survey and performance evaluation. Electronics 9(8), 1295 (2020)
26. N. Zeng, P. Wu, Z. Wang et al., A small-sized object detection oriented multi-scale feature fusion approach with application to defect detection. IEEE Trans. Instrum. Meas. 71, 1–14 (2022)
27. G. Yang, J. Lei, Z. Zhu, et al. AFPN: asymptotic feature pyramid network for object detection. arXiv preprint arXiv:2306.15988 (2023).
28. S. Liu, D. Huang, Y. Wang. Learning spatial fusion for single-shot object detection. arXiv preprint arXiv:1911.09516 (2019).
29. K. He, X. Zhang, S. Ren, et al. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition. 770–778 (2016).
30. S.H. Gao, M.M. Cheng, K. Zhao et al., Res2net: a new multi-scale backbone architecture. IEEE Trans. Pattern Anal. Mach. Intell. 43(2), 652–662 (2019)
31. J. He, X. Song, Z. Feng et al., ETM-face: effective training sample selection and multi-scale feature learning for face detection. Multimed. Tools Appl. 82, 26595–26611 (2023)
32. S. Yang, P. Luo, C.C. Loy, et al. Wider face: a face detection benchmark. Proceedings of the IEEE conference on computer vision and pattern recognition. 5525–5533 (2016).
33. V. Jain, E. Learned-Miller. Fddb: a benchmark for face detection in unconstrained settings. UMass Amherst technical report (2010).
34. J. Guo, J. Deng, A. Lattas, et al. Sample and computation redistribution for efficient face detection. arXiv preprint arXiv:2105.04714 (2021).

Publisher's Note