Li *et al.*
EURASIP Journal on Advances in Signal Processing (2024) 2024:6
https://doi.org/10.1186/s13634-023-01105-z

EURASIP Journal on Advances in Signal Processing

RESEARCH Open Access

ABYOLOv4: improved YOLOv4 human object detection based on enhanced multi-scale feature fusion

Rui Li[1,2], Xin Zeng[1], Shiqiang Yang[1*], Qi Li[1], An Yan[1] and Dexin Li[1]

*Correspondence: yangsq@126.com

[1] School of Mechanical and Precision Instrument Engineering, Xi'an University of Technology, Xian 710048, China
[2] Xi'an People's Hospital, Xi'an, China

Abstract

The purpose of human object detection is to obtain the number of people and their position in images, which is one of the core problems in the field of machine vision. However, the high missing detection rate from small- and medium-sized human bodies due to the large variety of human scale in human object detection tasks still influences the performance of human object detection. To solve the above problem, this paper proposed an improved ASPP_BiFPN_YOLOv4 (ABYOLOv4) method to detect human object detection. In detail, Atrous Spatial Pyramid Pooling (ASPP) module was used to replace the original Spatial Pyramid Pooling module to increase the receptive field level of the network and improve the perception ability of multi-scale targets. Then, the original Path Aggregation Network (PANet) multi-scale fusion module was replaced by the self-built bi-layer bidirectional feature pyramid network (Bi-FPN). Meanwhile, a new feature was imported into the proposed model to reuse the mid- and low-level features, which could enhance the ability of the network to express the characteristics of small- and medium-sized targets. Finally, the standard convolution in Bi-FPN was replaced by depth-separable convolution to make the network achieve the balance of accuracy and the number of parameters. To identify the performance of the proposed ABYOLOv4 model, the human object detection experiment is carried out by using the public data set of VOC2007 and VOC2012, the improved YOLOv4 algorithm is 0.5% higher than the original AP algorithm, and the weight file size of the model is reduced by 45.3 M. The experimental results demonstrated that the proposed ABYOLOv4 network has higher accuracy and lower computational cost for human target detection.

Keywords: Deep learning, Human object detection, YOLOv4, ASPP, Bi-FPN

1 Introduction

Human object detection has become a crucial and fundamental technology in the field of computer vision in recent years. Its ultimate goal is to obtain the positions and quantities of humans from images or videos. Object detection serves as the foundation for many other computer vision tasks, such as instance segmentation [1–4], image captioning [5–7], and object tracking [8].

Traditional methods for human detection include segmentation-based and matching-based methods, gradient-based methods, and statistical learning-based methods. Segmentation-based methods for human object detection accurately determine the position of the target human object by employing background subtraction or establishing statistical models. For instance, Rother et al. [9] proposed the GrabCut algorithm based on improved Graph-cuts, which converts target segmentation into energy minimization problems for human object detection in static images. To enhance the speed and accuracy of target segmentation detection, Han et al. [10] combined the GrabCut algorithm with a fusion of local linear estimation in a multi-scale framework. Gradient-based methods for object detection primarily utilize the gradient variations in images to detect the target objects. Dalal et al. [11] introduced the Histogram of Oriented Gradients (HOG), which focuses on the gradient direction changes in the image to detect the target objects. Statistical learning-based methods mostly employ support vector machines (SVMs). Ronfard et al. [12] performed segmentation and annotation of human body parts in images and utilized SVM and correlation vector machines for classification learning of the segmented body parts. This approach resulted in a body part classification model that accurately classifies body parts and provides their position information. These traditional methods for human detection have laid the foundation for this field. However, most of them face challenges such as slow detection speed, poor real-time performance, and subpar performance in complex scenes. They have not fully utilized the advantages of big data and exhibit limited robustness in handling variations in human scales.

The development of deep learning techniques has brought revolutionary progress to object detection. In the field of human object detection, deep learning-based methods outperform traditional approaches, as they can autonomously learn features at different levels from the training dataset. These methods can be mainly categorized into two types. The first type is two-stage object detection algorithms based on region proposals. Representative examples include the region-based convolutional neural network (R-CNN) framework proposed in 2014 [13], along with its improved versions, such as Fast R-CNN [14] and Faster R-CNN [15]. Although two-stage object detection methods achieve higher accuracy, they suffer from slow detection speeds, making it challenging to meet real-time requirements. The second type is one-stage object detection algorithms based on regression. Examples of such methods include YOLO [16], RetinaNet [17], and SSD [18]. These methods eliminate the need for region proposal networks and formulate object detection as a regression problem. They achieve faster detection speed while ensuring detection accuracy and improving robustness. Among the YOLO series algorithms, YOLOv4 [19], as a classic model in the YOLO series of algorithms, has been optimized in many aspects to achieve a compromise between speed and accuracy [20]. It has a simple structure, low requirements on basic equipment, and is easy to deploy.

Since YOLOv4 is a universal target detector, it is necessary to adjust the network in order to make it suitable for the single-class detection task of human target detection. This paper focuses on addressing the issues of low detection accuracy and significant missed detections of small- and medium-sized human targets in complex real-world visual scenes and proposes improvements to the original YOLOv4 algorithm.

The contributions of this work can be summarized as follows:

(1) To enhance the perception of multi-scale targets, this paper intends to replace the original SPP (spatial pyramid pooling) [21] module with the ASPP (Atrous Spatial Pyramid Pooling) [22] module, which increases the receptive field hierarchy of the network.
(2) To enhance the feature representation capability of the network for small- and medium-sized targets, a custom-built two-layer Bi-FPN (Bidirectional Feature Pyramid Network) [23] is employed to replace the original PANet (path aggregation network) [24] multi-scale fusion module. Additionally, a new feature input is introduced to reuse mid–low-level features and reduce the missed detection rate.
(3) To achieve a balance between accuracy and parameter efficiency, the standard convolutions in Bi-FPN are replaced with depth-wise separable convolutions [25], resulting in a complete ABYOLOv4 algorithm. Finally, the performance of the two different improved algorithms proposed in this paper is evaluated using the VOC2012 test set. The improved YOLOv4 algorithm shows a 0.5% increase in average precision (AP) compared to the original algorithm, while reducing the model's weight file size by 45.3 M, achieving a balance between accuracy and parameter efficiency.

The structure of the following sections is as follows: In the Method section of Chapter 2, a brief overview of the overall structure of YOLOv4 will be provided, followed by detailed explanations of the improvements made to the single-category YOLOv4 human detection model, including the introduction of the ASPP module and the construction of the two-layer Bi-FPN module. In the Experimental Results and Analysis section of Chapter 3, the proposed ABYOLOv4 algorithm will be validated, compared, and analyzed on the publicly available VOC2007 and VOC2012 datasets. The Conclusion section of Chapter 4 will provide a summary of the proposed algorithm and present prospects for future research.

2 Materials and methods

2.1 YOLOv4 object detection algorithm

YOLOv4 is a one-stage object detector that can determine the positions of the target objects in given images or videos, which incorporates numerous optimization techniques based on previous algorithms, including improvements in backbone networks, activation functions, loss functions, network training, and data processing. YOLOv4 consists of three main components: the backbone network (CSPDarkNet53) for feature extraction, the Spatial Pyramid Pooling (SPP) module to enhance high-level semantic features, a Path Aggregation Network (PANet) for multi-scale feature fusion, and a YOLO Head for predicting object positions. The detection process of YOLOv4 is illustrated in Fig. 1.

The backbone network used in YOLOv4 is CSPDarkNet53, which is an improvement based on DarkNet53. It incorporates residual connections to accelerate the training process and employs the Mish activation function. Compared to the traditional ReLU with hard zero boundaries, the Mish function enhances the network's generalization and accuracy. For feature aggregation, YOLOv4 adopts the PANet, a path aggregation network, at its neck. It introduces a bottom-up feature pyramid, aggregating different

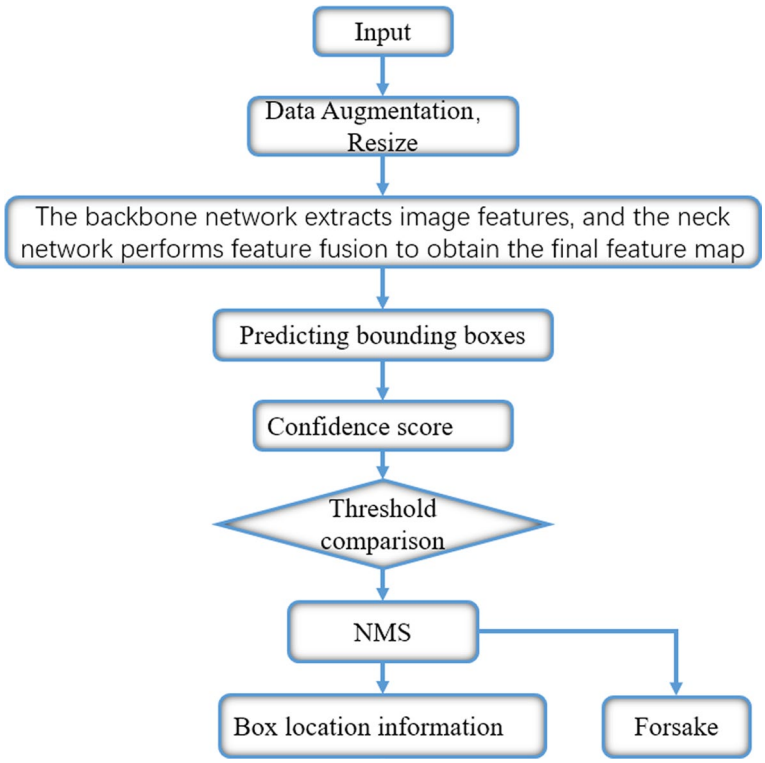

Fig. 1 Flow chart of YOLOv4 object recognition

features from various layers of the backbone network to enhance feature extraction capabilities. Additionally, YOLOv4 incorporates the SPP (Spatial Pyramid Pooling) as an additional module to enlarge the network's receptive field, enabling the extraction of important contextual feature information. The network structure is illustrated in Fig. 2.

The head of the YOLOv4 network remains the same as YOLOv3. However, in terms of the loss function, YOLOv4 no longer utilizes mean square error (MSE) as the regression box prediction error. Instead, it adopts the CIOU (complete intersection over union) error. The CIOU error takes into account the distance, overlap ratio, and scaling between the target and anchor boxes. Compared to the IOU loss, the CIOU loss provides a more stable regression for the bounding boxes. The specific formulas for calculating the CIOU error are given by Eqs. (1–3):

$$L_{\mathrm{CIOU}} = 1 - \mathrm{IOU}_{(a,b)} + \frac{\rho^2(a_{\mathrm{ctr}}, b_{\mathrm{ctr}})}{d^2} + \alpha v \tag{1}$$

$$\alpha = \frac{v}{\left(1 - \mathrm{IOU}_{(a,b)}\right) + v} \tag{2}$$

$$v = \frac{4}{\pi^2}\left(\frac{\arctan w_{gt}}{h_{gt}} - \arctan \frac{w}{h}\right)^2 \tag{3}$$

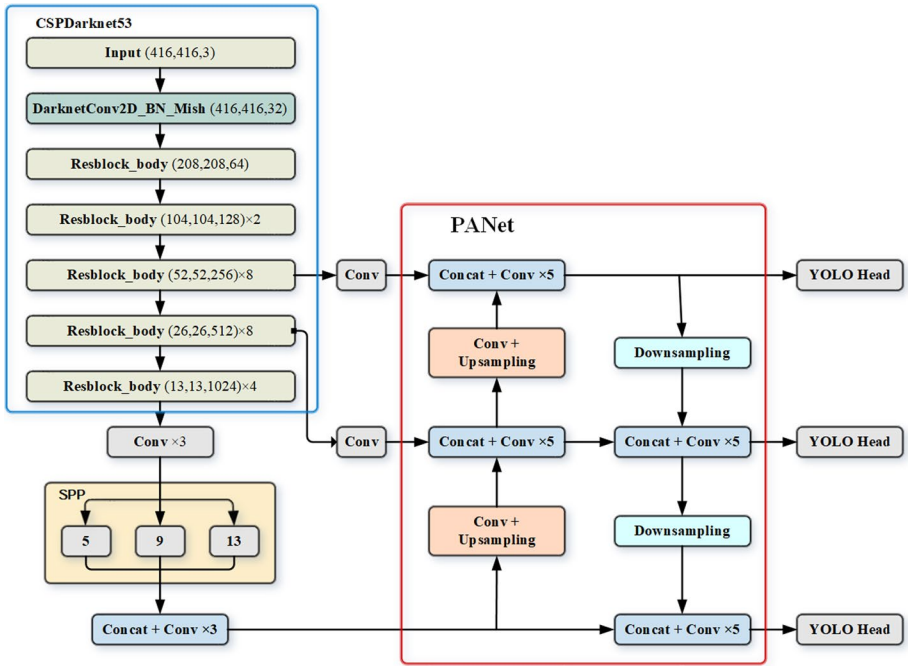

Fig. 2 YOLO4 framework

In this context, $\mathrm{IOU}_{(a,b)}$ refers to the intersection of union between the ground truth box and the predicted box. $\rho^2(a_{\mathrm{ctr}}, b_{\mathrm{ctr}})$ represents the Euclidean distance between the centers of the ground truth box and the predicted box, while d denotes the diagonal distance of the minimum bounding box that encloses both boxes. w_{gt}, h_{gt}, w, h are the width and height of the real box and the predicted box, respectively.

2.2 Category adjustment for detection

The original YOLOv4 is a multi-class universal detector. However, single-class detection is advantageous for saving computational resources and accelerating model inference speed. In the class adjustment for YOLOv4, a single-class object detection model is trained specifically for detecting human bodies. YOLOv4 employs a multi-scale prediction approach, where it performs detection on three different scales: $(52 \times 52 \times 64)$, $(26 \times 26 \times 256)$, and $(13 \times 13 \times 1024)$. Each scale is responsible for detecting large, medium, and small objects , respectively, and the total number of predicted pixels is the sum of the pixels across the three scales. The original YOLOv4 defines a set of anchor boxes with different sizes for each scale, resulting in a total number of anchor boxes of:

$$N_{\mathrm{box}} = N_{\mathrm{pixel}} N_{\mathrm{anchor}} = 52 \times 52 + 26 \times 25 + 13 \times 13 \times 3 = 10647 \tag{4}$$

YOLOv4 obtained the size of the prediction box by predicting the horizontal and vertical offset of each anchor box and the width–height difference $\widehat{t_x}$, $\widehat{t_y}$, $\widehat{t_w}$, $\widehat{t_h}$, and each prediction box contained predicted target confidence c to judge whether there were targets in the prediction box, and 20 target categories confidence p to predict the probability of each category of the target number field in the box. Therefore, the total number of prediction parameters for YOLOv4 is:

$$N_{\text{parm}} = N_{\text{box}} \times (4 + 1 + 20) = 266175 \tag{5}$$

To obtain a single-class human object detector, we can modify the 20 class confidence scores to only include the "person" class. According to Eq. (5), the total number of predicted parameters for a single-class human detector is:

$$N_{\text{parm}} = N_{\text{box}} \times (4 + 1 + 1) = 63882 \tag{6}$$

By combining Eqs. (5) and (6), we can conclude that the parameter count of the YOLOv4 human object detector, after adjusting the detection classes, accounts for only 24% of the original YOLOv4 object detection algorithm. This undoubtedly reduces the computational burden and speeds up the model's inference rate. Therefore, adjusting the detection classes is necessary.

2.3 ABYOLOv4 human object detection network

After adjusting the detection categories of YOLOv4, further modifications were made to the model's structure to achieve faster and better human object detection. Firstly, to address the issue of low detection rate for small-scale human objects, the SPP module was replaced with the ASPP module to increase the network's receptive field hierarchy and improve its ability to perceive multi-scale objects. Secondly, a custom-built dual-layer Bi-FPN was introduced to replace the original PANet for multi-scale feature fusion, along with the addition of a new feature input to reuse middle and low-level features, enhancing the network's feature representation for medium and small-scale objects. Lastly, to strike a balance between accuracy and parameter count, standard convolutions in Bi-FPN were replaced with depth-wise separable convolutions, The improved network is named ABYOLOv4. The following is a detailed explanation of the improved module and the overall network structure of ABYOLOv4.

2.3.1 ASPP feature enhancement module

ASPP (Atrous Spatial Pyramid Pooling) is a multi-scale high-level spatial feature extraction module. It consists of a 1×1 convolution, three 3×3 convolutions with different dilation rates, and an adaptive global pooling layer. The dilation rates of the Atrous convolutions can be customized, allowing for flexible multi-scale feature extraction. Compared to the SPP module in the original YOLOv4 that uses pooling layers, ASPP replaces the pooling layer with three sets of dilated convolutions to reduce information loss during pooling. It divides the input features into five branches, each with a different receptive field. Compared to the SPP module with only three receptive fields, ASPP refines the spacing between receptive fields, enlarging the receptive field range and enhancing the network's ability to perform semantic discrimination on high-level features. The structure of ASPP is illustrated in Fig. 3.

The receptive field of a convolutional layer represents the information processing range of the convolution. As the network deepens, the receptive field expands, allowing the network to transition from local perception to global perception. To better align the ASPP module with the objectives of this paper, an analysis of the composition of the receptive field in the module is necessary. The receptive field can be represented by Eq. (7).

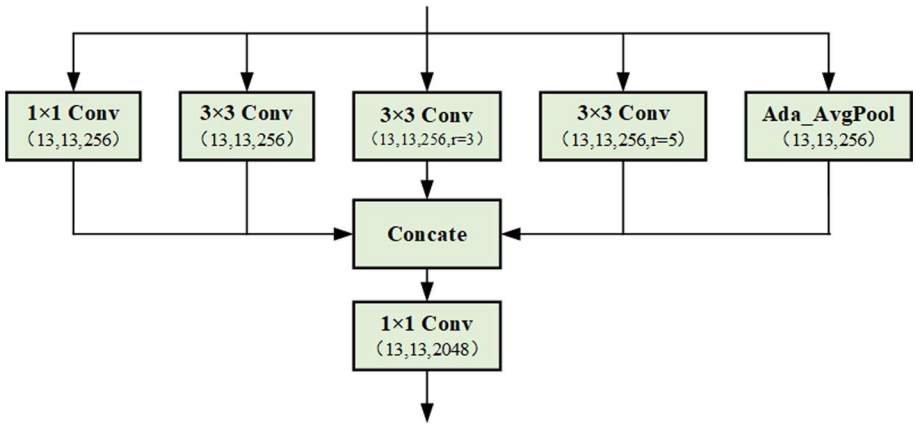

Fig. 3 ASPP structure

Table 1 Comparison table of dilated rate and receptive field

Convolution kernel	Dilation rate	Receptive field
3 × 3	1	3
3 × 3	2	5
3 × 3	3	7
3 × 3	4	9
3 × 3	5	11

$$R = (k - 1) \times (r - 1) + k \tag{7}$$

The formula indicates that R represents the receptive field size, k denotes the kernel size, and r is the dilation rate. Based on this equation, the correspondence between the convolutional receptive field and the dilation rate is shown in Table 1.

To enhance the detection performance of the network for small- and medium-sized human targets, this paper chooses a small receptive field size of 3. To ensure a balanced increase in the receptive field hierarchy, receptive fields 7 and 11 are selected. This corresponds to using 3×3 convolutions with dilation rates of 1, 3, and 5, along with the 1×1 convolution and global pooling layers in the ASPP module. The chosen receptive field composition for the ASPP module in this paper is 1, 3, 7, 11, and 13. This structure has two additional receptive field sizes compared to the SPP module in the original YOLOv4, and the receptive fields are relatively small. This allows for better extraction of multi-scale features and enhances the network's feature representation capability.

2.3.2 Bi-FPN multi-scale feature fusion module

Bi-FPN short for Bidirectional Feature Pyramid Network is a weighted bidirectional (top-down + bottom-up) feature pyramid network. Its comparison with the PANet network structure in YOLOv4 is shown in Fig. 4. In the figure, black arrows represent convolution operations, red arrows represent upsampling, blue arrows represent downsampling, and purple arrows represent cross-stage connections. Different-colored dots

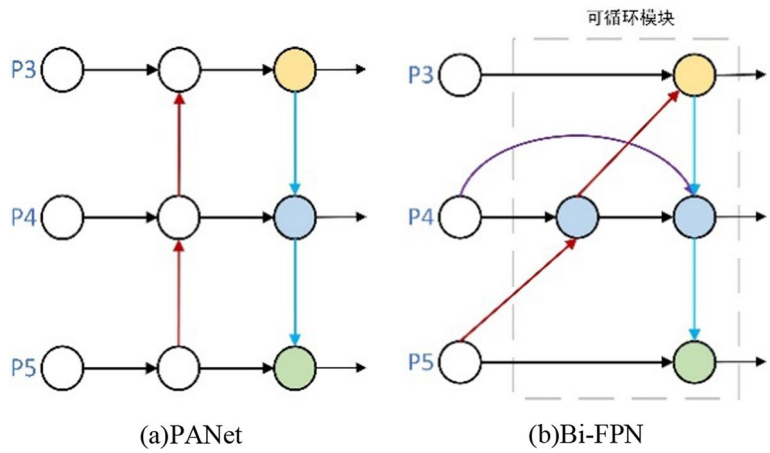

Fig. 4 Structure comparison of PANet and Bi-FPN

represent features from different levels, while dots of the same color represent features from the same level.

(1) Eliminating nodes with only one feature input: The design principle of Bi-FPN is to fuse features from different levels. If a node has only one input edge without feature fusion, its contribution to the feature network aimed at integrating different features will be minimal. Removing such nodes has little impact on the network and simplifies it.
(2) Preserving the original input features: When there is no upsampling or downsampling operation between input and output nodes, a cross-stage connection is added. This is done to prevent the loss of original information caused by convolutional operations and to increase the richness of feature sources. It is beneficial for the network to learn the relationships between different features.
(3) Recursive Structure Design: Unlike PANet, which only has one bidirectional feature fusion, the bidirectional feature fusion flow in Bi-FPN is a recursive structure. It can adapt to the backbone network by changing the number of iterations according to different network designs. This recursive structure also facilitates the realization of deeper feature fusion in the network.
(4) Weighted Feature Fusion: This approach involves learning the importance of different input features and performing a discriminative fusion. Traditional feature fusion methods often simply overlay or add feature maps, such as using concatenation or shortcut connections, without distinguishing between the simultaneously added feature maps. However, different input feature maps have different resolutions and contribute differently to the fused input feature map. Therefore, simply adding or overlaying them is not the optimal operation. This paper adopts the fast normalized fusion method to achieve weighted feature fusion. The specific expression is as follows:

$$O = \sum_i \frac{w_i \times I_i}{\in + \sum_j w_j} \tag{8}$$

The O represents the weighted output, I_i denotes the i-th feature input and w_i represents its corresponding weight.

Due to the stackability of Bi-FPN, this study conducted experiments on the stacking effect of different layers of Bi-FPN with YOLOv4. The experimental results are shown in Table 2.

As can be seen from Table 2, when the number of superpositions increases from 1 to 2, the AP_{person} index rises to 92.0%, and when the number of layers is increased, the network begins to degrade. Therefore, this paper chooses a two-layer Bi-FPN to merge with the original YOLOv4, in order to achieve a balance between the precision and the number of parameters in the merged network. Standard convolution in Bi-FPN is replaced by depth-separable convolution to reduce the number of parameters in a multi-scale fusion module.

2.3.3 Human object detection based on ABYOLOv4 algorithm

After completing the construction of ASPP and Bi-FPN, to enhance the performance of the multi-scale features in Bi-FPN, an additional adjustment input from the mid–low layers was added on top of the original network's three feature inputs. This modification transformed the neck part into a structure with four inputs and three outputs, aiming to reuse mid–low layers for small- and medium-scale targets. The final constructed ABYOLOv4 network structure is illustrated in Fig. 5, comprising the CSPDarknet53 backbone network, the feature pyramid ASPP structure, and the Bi-FPN multi-scale feature module. The red arrows represent the newly added mid–low feature input, the red box indicates the replaced Bi-FPN multi-scale feature module, and the green box signifies the replacement of the SPP module with the ASPP module.

When utilizing the improved ABYOLOv4 human object detection network for object detection, the specific workflow is as follows: When an image is input, it is first resized to (416, 416, 3)and then fed into the CSPDarknet53 backbone feature extraction network. Subsequently, the feature pyramid ASPP is employed for multi-scale feature enhancement. Following this, the bidirectional feature pyramid Bi-FPN combines the mid-low features extracted by the backbone network at three additional scales and the high-level semantic features enhanced by ASPP. Finally, the combined features are input into YOLOHead for prediction, yielding the positional information of human targets.

3 Experiment and result analysis

3.1 Model training parameter

The human detection model proposed in this paper was trained and tested on the VOC2007 and VOC2012 datasets. The experiments were conducted using a single P106-100 GPU with 6 GB of memory, an Intel(R) Core(TM) i5-4460 CPU @ 3.20 GHz, and

Table 2 Comparison table of different layers of Bi-FPN fusion effect

Bi-FPN layers	AP_{person} (%)
1	91.2
2	92.0
3	91.5

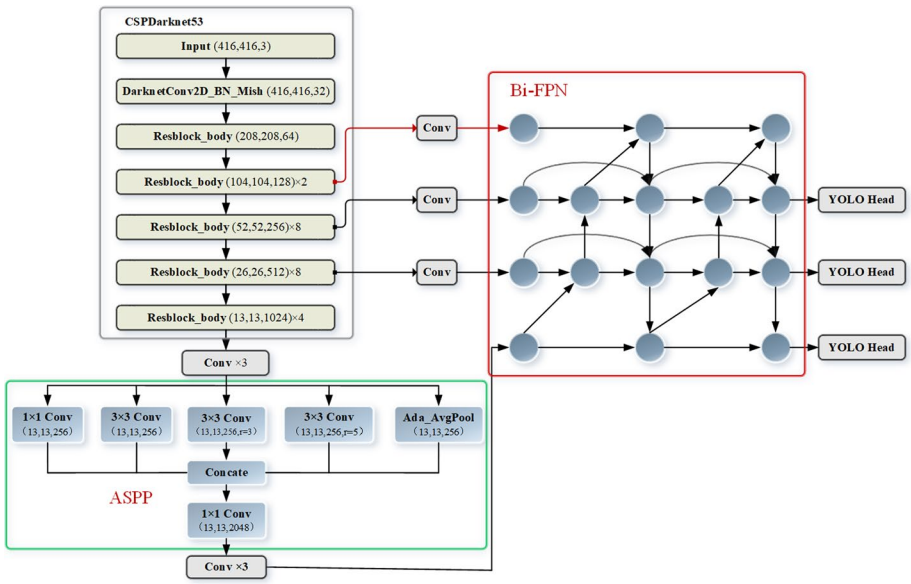

Fig. 5 ABYOLOv4 framework

Table 3 Experimental training environment

Environment configuration	
System	Windos10
GPU	P106-100
Memory size	6 GB
CPU	Intel(R) Core(TM) i5-4460 CPU @3.20GHZ
Python	3.7
Torch	1.2.0

the software environment consisted of PyTorch 1.2.0, CUDA 10.0, and cuDNN 7.4.0. Table 3 presents the experimental training environment, Adam optimizer was used in the model, and other parameters were the default values. The training process consisted of two stages: the first stage involved frozen training, which preserved the training speed while preventing the network parameters from being disrupted in the early stages of training. The second stage involved unfrozen training. The total number of epochs was set to 100. The first 50 epochs were dedicated to frozen training with a batch size of 8 and a learning rate of $1 \times e^{-3}$. The remaining 50 epochs were allocated for unfrozen training, with a batch size of 4 and a learning rate of $1 \times e^{-4}$. The learning rate decay coefficient was set to 0.92.

3.2 Model evaluation indicator

In this paper, precision, recall, average precision (AP_{person}), and frame rate (FPS) were used to compare target detection networks. Precision refers to the ratio of the number of correctly predicted positive samples to the number of all predicted positive samples. Recall refers to the ratio of the number of correctly predicted positive samples to the

total number of true-positive samples; As shown in Formula 10, the value of mAP in target detection is equal to the average of AP values for each category. ABYOLOv4 is a single classification human object detection model, that is $n = 1$, so AP is equivalent to mean average precision (mAP). Therefore, this paper uses AP as the evaluation index; FPS refers to how many images can be detected per second; the purpose is to compare the speed of object detection in the object detection model, which is an indispensable indicator.

$$AP_{\text{person}} = \frac{\sum \frac{N(TP)_{\text{person}}}{N(\text{object})_{\text{person}}}}{N(\text{images})_{\text{person}}} \tag{9}$$

AP_{person} is the average precision metric specifically used for pedestrian detection. A higher AP_{person} indicates fewer false positives and false negatives, indicating a better overall performance of the model. $N(TP)_{\text{person}}$ represents the number of correctly detected pedestrians, $N(\text{object})_{\text{person}}$ represents the actual number of pedestrians in the image, and $N(\text{images})_{\text{person}}$ represents the total number of pedestrians present in the test images.

$$mAP = \frac{1}{n}\sum_{i=1}^{i=n} AP_i \tag{10}$$

where n represents the number of classes, and AP_i represents the average accuracy of each class.

3.3 Model training and testing results

The training loss curve and PR curve for the ABYOLOv4 model are shown in Figs. 6 and 7; it can be observed that due to the use of pre-trained weights, the model's loss decreases rapidly and converges to the optimum around 50 epochs. The corresponding AP_{person} is 92.3%.

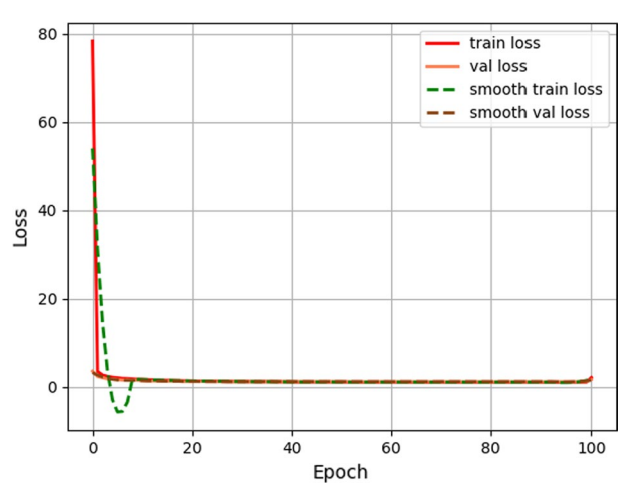

Fig. 6 Loss value curve

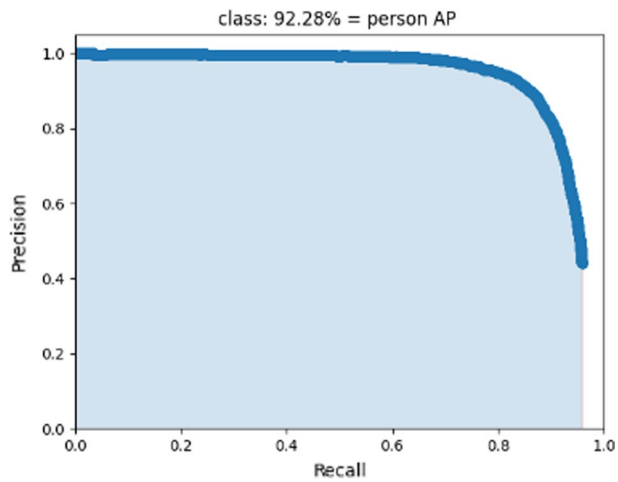

Fig. 7 PR curve

Table 4 Ablation experiment result

Model	Precision (%)	Recall (%)	AP_{person}(%)	FPS	Weight file size (mb)
YOLOv4	93.2	82.6	91.8	24	244.29
YOLOv4 + Bi-FPN	90.2	85.0	92.0	31	178.47
YOLOv4 + ASPP	92.7	83.7	92.1	23	275.32
ABYOLOv4	91.9	84.3	92.3	29	199.00

In Table 4, the results of the human detection experiment using distillation on the dataset are shown for a Score_threshold of 0.5. It can be observed that when ASPP and two-layer Bi-FPN are incorporated into the original YOLOv4, the AP_{person} of the network improves by 0.2% and 0.3%, respectively. The use of depth-wise separable convolutions in the self-built Bi-FPN reduces the size of the weight files by 24%. After adding ASPP, the size of the weight file increases by 12.5% due to the addition of convolution operation in the ASPP module. Compared with the original YOLOv4, the AP_{person} of the ABYOLOv4 network with Bi-FPN and ASPP modules added at the same time has an increase of 0.5%, and the recall rate and frame rate FPS of the ABYOLOv4 model have increased by 1.7% and 5, respectively, indicating that the model can detect more targets and faster detection speed. It achieves the purpose of reducing the missing rate of small- and medium-scale targets, and the weight file size is also reduced by 18%, achieving the balance between precision and parameter number.

In order to further verify the detection effect of the ABYOLOv4 network and the original YOLOv4 network on small and medium target human bodies, this paper selected 4 groups of pictures, respectively, to compare and verify the model detection effect. The experimental results are shown in Fig. 8. In a–h of Fig. 8, the detection effect of the original YOLOv4 network is listed on the left. On the right side are the improved ABYOLOv4 network detection results. Figure a and b are the pictures selected from the VOC2007 test set, Figure c, and Figure d are the pictures downloaded from the network,

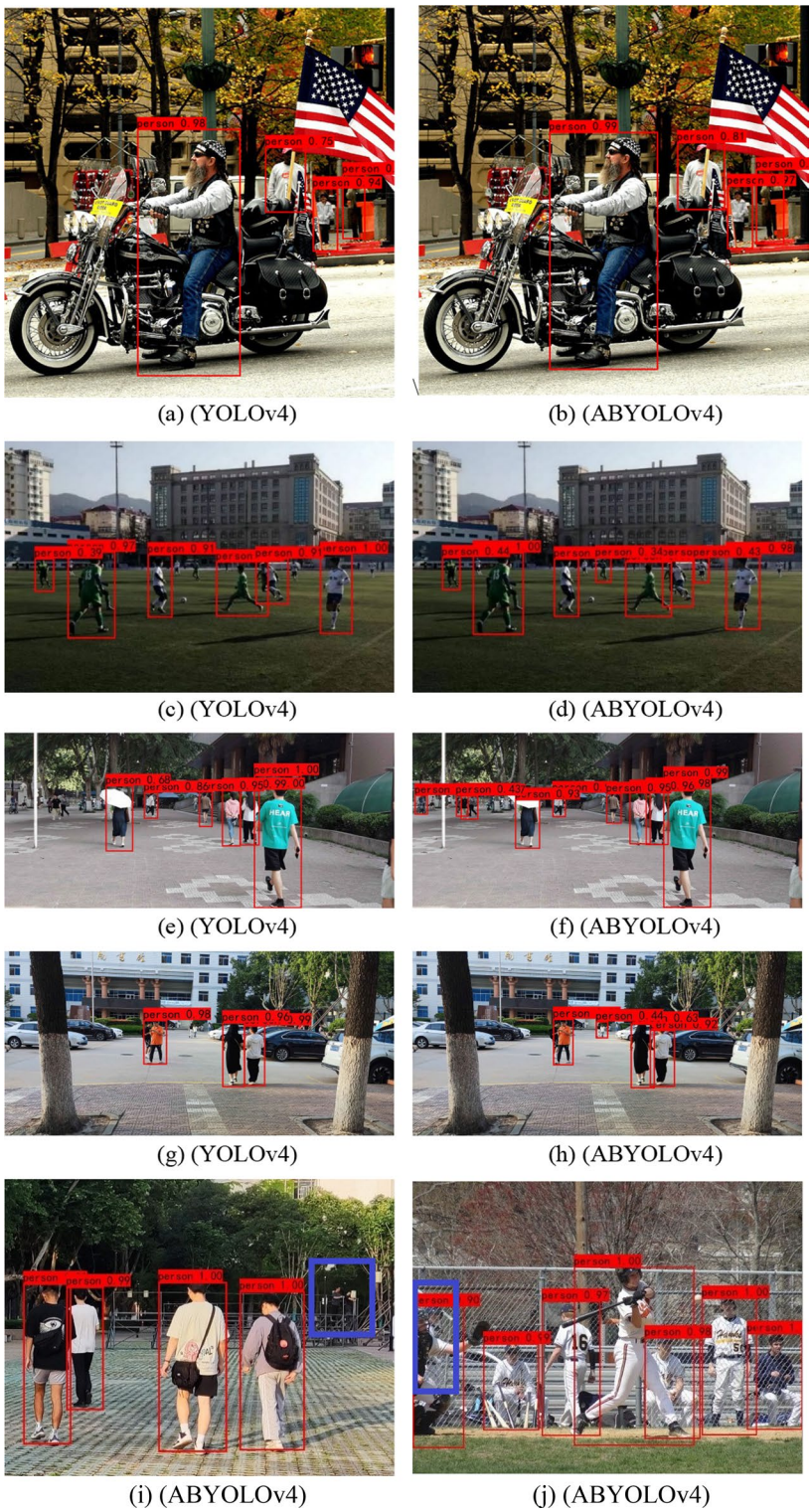

Fig. 8 Comparison of detection effect of YOLOv4 and ABYOLOv4

and Figure e,f,g, and h are the pictures collected by mobile phones on campus. It can be seen from the results of a and b that the two networks have a good effect on the image detection of the selected test set. There are no missed cases, but the classification confidence of ABYOLOv4 is higher than that of the original YOLOv4. According to the comparison of c,d,e,f,g, and h, it can be seen that when there are more human bodies in the picture and the scale is small, the original YOLOv4 has seriously missed detection, and only 6 human frames are detected, while ABYOLOv4 has detected 8 human frames, and the missed detection rate is lower than that of the original YOLOv4. The confidence of ABYOLOv4 is improved. It can be seen from the actual scene experiment results that the original YOLOv4 still has a high detection rate for small- and medium-sized human bodies, and the improved model has a better effect.

It is worth noting that, under the influence of background, occlusion, and other conditions, the proposed model will also have the phenomenon of missing detection. The detection results in Figure I and Figure J show that the human object missed in the picture is in the blue box. By analyzing the reasons for missing detection, Figure i is on the one hand because the human object is small in scale and slightly blocked. On the other hand, the person's wearing color and the background color are very similar, so the human object is not detected. In Figure j, the human body in the blue box on the left is seriously shielded, so there is a phenomenon of missing detection. In addition to blending into the background, too small a target severe occlusion, and serious interference, the human object missed by the original YOLOv4 network can be detected by the improved ABYOLOv4 network. The improved YOLOv4 network with enhanced integration of multi-scale features also has a better application than the original network in real scenes.

4 Discussion

To address the issue of high omission rates for small- and medium-sized human object detection caused by distractions, complex backgrounds, and varying human scales in human object detection tasks, this study builds upon the YOLOv4 network. Firstly, the SPP module is replaced with the ASPP module to increase the network's receptive field hierarchy and enhance its ability to perceive multi-scale targets. Secondly, the original PANet multi-scale fusion module is replaced with a self-built two-layer Bi-FPN, which introduces a new feature input to reuse medium- and small-scale features to enhance the network's ability to express features of small- and medium-sized targets. Finally, to achieve the balance between precision and parameter number, the standard convolution in Bi-FPN is replaced by deep separable convolution, and the ABYOLOv4 network model is established. The performance of the ABYOLOv4 network is trained on the VOC2007 and VOC2012 datasets and tested through distillation experiments. When compared to the original YOLOv4 network, the proposed improvement modules show corresponding enhancements, reducing the impact of varying human scales on human object detection. Model detection tests conducted on selected test sets and downloaded images validate that ABYOLOv4 outperforms the original YOLOv4 network in detecting small- and medium-sized human targets. In terms of the VOC dataset, compared with the original YOLOv4, the weight file size of the ABYOLOv4 model is reduced by 45.3 M, the AP_{person} performance index is improved by 0.5%, the recall rate is increased

by 1.7%, and the frame rate increase indicates that the detection speed is also improved. It reflects the effectiveness of the ABYOLOv4 network model, achieves the balance of precision and parameter number, and can realize more accurate human object detection.

Abbreviations

YOLO	You only look once
ABYOLOv4	ASPP BiFPN YOLOv4
VOC	Visual object classes
SPP	Spatial pyramid pooling
ASPP	Atrous Spatial Pyramid Pooling
PANet	Path aggregation network
Bi-FPN	Bidirectional feature pyramid network
HOG	Histogram of oriented gradients
SVM	Support vector machines
MSE	Mean square error
IOU	Intersection over union
R-CNN	Region convolutional neural network

Acknowledgements
The authors would like to thank the anonymous reviewers for their valuable comments and suggestions that helped improve the quality of the manuscript.

Author contributions
All authors take part in the discussion of the work described in this paper. These authors contributed equally to this work.

Funding
This work was supported by Shaanxi Provincial Department of Education 2022 General Special Research Program Projects (No.22JK0471) and China Postdoctoral Foundation Project (No.2022MD723841).

Availability of data and materials
The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 29 July 2023 Accepted: 20 December 2023
Published online: 04 January 2024

References

1. B. Hariharan, P. Arbelaez, R. Girshick, J. Malik, Simultaneous detection and segmentation. In Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, Proceedings, Part VII 13, 297–312 (2014). https://doi.org/10.5220/0009142905550561
2. B. Hariharan, P. Arbeláez, R. Girshick, J. Malik, Hypercolumns for object segmentation and fine-grained localization, In Proceedings of the IEEE conference on computer vision and pattern recognition. 447–456 (2015). https://doi.org/10.1109/cvpr.2015.7298642
3. J. Dai, K. He, and J. Sun, Instance-aware semantic segmentation via multi-task network cascades. In Proceedings of the IEEE conference on computer vision and pattern recognition, 3150–3158 (2016). https://doi.org/10.1109/cvpr.2016.343
4. K. He, G. Gkioxari, P. Dollar, and R. Girshick, Mask r-cnn. In Proceedings of the IEEE international conference on computer vision, 2961–2969 (2017). https://doi.org/10.48550/arXiv.1703.06870
5. A. Karpathy, L. Fei-Fei, Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE conference on computer vision and pattern recognition, 3128–3137 (2015). https://doi.org/10.1109/cvpr.2015.7298932
6. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov et al., Show, attend and tell: Neural image caption generation with visual attention. In International conference on machine learning, 2048–2057 (2015). https://doi.org/10.1109/cvpr.2015.7298935

7. Q. Wu, C. Shen, P. Wang, A. Dick, A. van den Hengel, Image captioning and visual question answering based on attributes and external knowledge. IEEE Trans. Pattern Anal. **40**(6), 1367–1381 (2018). https://doi.org/10.1109/tpami.2017.2708709
8. K. Kang, H. Li, J. Yan, X. Zeng, B. Yang, T. Xiao et al., T-cnn: Tubelets with convolutional neural networks for object detection from videos. IEEE Trans. Circ. Syst. Vid. **28**(10), 2896–2907 (2017). https://doi.org/10.1109/cvpr.2016.95
9. C. Rother, V. Kolmogorov, A. Blake, "GrabCut": interactive foreground extraction using iterated graph cuts. CM Trans. Graph. **3**(3), 309–314 (2004). https://doi.org/10.1145/1015706.1015720
10. S.D. Han, W.B. Tao, X.L. Wu, X.C. Tai, T.J. Wang, Fast image segmentation based on multilevel banded closed-form method. Pattern Recogn. Lett. **31**(3), 216–225 (2010). https://doi.org/10.1016/j.patrec.2009.10.005
11. N. Dalal, B. Triggs, Histograms of oriented gradients for human detection. In 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), **1**, 886–893 (2005). https://doi.org/10.1109/cvpr.2005.177
12. R. Ronfard, C. Schmid, B. Triggs, Learning to parse pictures of people. Lect. Notes Comput. Sci. 700–714 (2002). https://doi.org/10.1007/3-540-47979-1_47
13. R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition. 580–587 (2014). https://doi.org/10.1109/cvpr.2014.81
14. R. Girshick, Fast r-cnn. In Proceedings of the IEEE international conference on computer vision. 1440–1448 (2015). https://doi.org/10.1109/iccv.2015.169
15. S.Q. Ren, K.M. He, R. Girshick, J. Sun, Faster r-cnn: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. **39**(6), 1137–1149 (2017). https://doi.org/10.1109/tpami.2016.2577031
16. J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition. 779–788 (2016). https://doi.org/10.1109/cvpr.2016.91
17. T.Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision. 2980–2988 (2017). https://doi.org/10.1109/iccv.2017.324
18. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.Y. Fu, et al. SSD: single shot multibox detector. In Computer Vision–ECCV 2016: 14th European Conference. Part I 14 (21–37) (2016). https://doi.org/10.1007/978-3-319-46448-02
19. A. Bochkovskiy, C.Y. Wang, H.Y. M. Liao, Yolov4: optimal speed and accuracy of object detection, (2020). arXiv:2004.10934
20. D. Wu, S. Lv, M. Jiang, H. Song, Using channel pruning-based YOLO v4 deep learning algorithm for the real-time and accurate detection of apple flowers in natural environments. Pattern Recogn. Lett. **178**, 105742 (2020). https://doi.org/10.1016/j.compag.2020.105742
21. K. He, X. Zhang, S. Ren, J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. Pattern. Anal. **37**(9), 1904–1916 (2015). https://doi.org/10.1109/tpami.2015.2389824
22. L.C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. **40**(4), 834–848 (2017). https://doi.org/10.1109/tpami.2017.2699184
23. M. Tan, R. Pang, Q.V. Le, Efficientdet: scalable and efficient object detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 10781–10790 (2020). https://doi.org/10.1109/cvpr42600.2020.01079
24. S. Liu, L. Qi, H. Qin, J. Shi, J. Jia, Path aggregation network for instance segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition. 8759–8768 (2018). https://doi.org/10.1109/cvpr.2018.00913
25. A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, Weyand, T. et al., Mobilenets: Efficient convolutional neural networks for mobile vision applications. (2017). arXiv:1704.04861

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.