applied sciences

MDPI

Article

PBA-YOLOv7: An Object Detection Method Based on an Improved YOLOv7 Network

Yang Sun *, Yi Li *, Song Li, Zehao Duan, Haonan Ning and Yuhang Zhang

College of Mechanical and Equipment Engineering, Hebei University of Engineering, Handan 056038, China; lisong1422@163.com (S.L.); 15153292207@163.com (Z.D.); 18617720631@163.com (H.N.); 17831953480@163.com (Y.Z.)

* Correspondence: sungcdx@foxmail.com (Y.S.); liyimuzili711@163.com (Y.L.)

Abstract: Deep learning-based object detection methods address the problem of how to trade off the object detection accuracy and detection speed of the model. This paper proposes the PBA-YOLOv7 network algorithm, which is based on the YOLOv7 network, and first introduces the PConv, which lightens the ELAN module in the backbone network structure and reduces the number of parameters to improve the detection speed of the network and then designs and introduces the BiFusionNet network, which better aggregates the high-level semantic features and the low-level semantic features; and finally, on this basis, the coordinate attention mechanism is introduced to make the network focus on more critical features without increasing the model complexity. The coordinate attention mechanism is introduced to make the network focus more on important feature information and improve the feature expression ability of the network without increasing the model complexity. Experiments on the publicly available KITTI's dataset show that the PBA-YOLOv7 network model significantly improves both detection accuracy and detection speed compared to the original YOLOv7 model, with 4% and 7.8% improvement in mAP0.5 and mAP0.5:0.95, respectively, and six frames improvement in FPS. The improved algorithm in this paper weighs the model's detection accuracy and detection speed in the detection task. It performs well compared to other algorithms, such as YOLOv7 and YOLOv5l.

Keywords: YOLOV7 network model; PConv convolution; BiFusionNet; coordinate attention

Citation: Sun, Y.; Li, Y.; Li, S.; Duan, Z.; Ning, H.; Zhang, Y. PBA-YOLOv7: An Object Detection Method Based on an Improved YOLOv7 Network. *Appl. Sci.* **2023**, *13*, 10436. https://doi.org/10.3390/app131810436

Academic Editors: Chuan-Ming Liu and Wei-Shinn Ku

Received: 8 August 2023
Revised: 13 September 2023
Accepted: 16 September 2023
Published: 18 September 2023

1. Introduction

Accompanied by the arrival and application of the Internet of Things and 5G technology, intelligent vehicles have entered a new stage of development, and environment sensing based on target detection is the basis for realizing autonomous driving technology [1]. In the complex street environment, it is of great significance to accurately and efficiently identify the target information, reduce the misjudgment rate of the self-driving vehicle in driving, and reduce the occurrence of traffic accidents. Traditional target detection methods utilize machine learning algorithms that mainly rely on a sliding window selection of candidate regions to extract features in images, such as Scale-Invariant Feature Transform (SIFT) [2], Histogram of Oriented Gradients (HOG) [3], and Deformable Part Models (DPM) [4], etc., and then the obtained semantic features are classified and regressed by Support Vector Machine (SVM) [5]. The traditional algorithms need better mobility and more generalization ability, and the process of manually extracting features is cumbersome and complex, which leads to the limitations of the algorithms' applications and makes it difficult to cope with today's complex traffic scenarios.

With the rapid development of computers and deep learning [6,7], many researchers have employed sophisticated deep learning models to implement object detection in sensed images. Object detection algorithms based on deep learning neural networks can be broadly categorized into two main types: two-stage detection algorithms and one-stage detection algorithms.

Two-stage detection algorithms concern two-stage processing of target frames or images; representative algorithms are R-CNN (Region-based Convolutional Neural Network) [8], Fast R-CNN, Faster R-CNN, and Mask R-CNN, which as the pioneering work of two-stage detection, uses selective search algorithms to propose a series of Regions of Interest (RoI) in the first stage, and uses CNNs to extract deep features within the regions in the second stage and classify and localize them accordingly. Fast R-CNN [9] advances the position of convolutional layers at each position and classifies and localizes them accordingly. In the Region of Interest (RoI) in the first and second stages, CNN is used to extract the deep features in the region and classify and localize them accordingly. Fast R-CNN advances the position of the convolutional layer and computes a set of feature vectors at each position, which reduces the computational effort in the second stage to compute the feature vectors at the corresponding positions through the positional mapping relationship between the image and the features and reduces the computational effort in the second stage. Faster R-CNN [10] proposes a Region Proposal Network (RPN) to replace the selective search algorithm so that part of the convolutional layer directly generates region proposals and intra-region detection features, simplifying the first stage's computation process. Fast R-CNN and Faster R-CNN reduce the computation time of the first and second stages, respectively, based on R-CNN. However, the computation for many redundant region proposals inevitably lowers the detection efficiency. Mask R-CNN [11] expands the target instance segmentation function based on Faster R-CNN. In Faster R-CNN, the main focus is on target detection, i.e., determining the location and class of the target. Mask R-CNN further introduces pixel-level instance segmentation on this basis, i.e., generating an accurate mask for each target.

The principle of the two-stage detection algorithm is to first extract a series of regions of interest that may contain the target using a selective search method, and then use the CNN model to extract the deep features of the candidate regions one by one, and finally classify and regress the extracted features. Although the two-stage detection algorithm has high detection accuracy, its many redundant operations increase its time cost, resulting in slow detection speed, which cannot meet the real-time requirements and is difficult to deploy on self-driving cars [12,13].

The second category is single-stage detection algorithms. The main algorithms for single-stage detection algorithms are the YOLO (You Only Look Once) series [14] and SSD (Single Shot MultiBox Detector) [15]. The SSD algorithm is another deep learning-based target detection algorithm whose core idea is to transform the target detection into a single forward pass regression problem—being able to accomplish both target localization and classification in a single forward propagation. The SSD algorithm performs target detection by applying a series of predefined anchor boxes on feature maps at different scales, sliding these anchor boxes on the feature map, and using convolutional operations to classify and bounding box regression on them to obtain the final target detection results.

YOLO is a commonly used single-stage target detection algorithm with fast speed and high accuracy [16,17]. Its core idea is to transform target detection into a single forward-pass regression problem. A comparison table of the YOLO series versions is shown in Table 1. YOLOv1 [16] is the initial version of the YOLO series. It implements real-time target detection by dividing the input image into grid cells and predicting each cell's bounding box and category probabilities.YOLOv1 uses convolutional neural networks for feature extraction and a fully connected layer for target prediction.YOLOv2 [18] introduces several significant improvements over YOLOv1, including using the Darknet-19 architecture, multiscale training and testing, anchor frames, scale clustering, and fine-grained features.YOLOv3 [19] added more improvements to YOLOv2, including using a Residual Network (ResNet) as a feature extractor, adopting multiscale prediction, and using a feature pyramid network for features of different scales.YOLOv4 [20] introduced some critical improvements in the YOLOv3, introduced CSPDarknet53 as a feature extractor, used the SPP structure, applied multiscale inference, and the PAnet network. YOLOv5 makes some modifications and optimizations based on YOLOv4, including using a lightweight net-

work structure, introducing model distillation techniques, and adding data enhancement strategies. YOLOX, like YOLOv5, uses the Focus network structure in the backbone part, obtains four independent feature layers, and then makes the four independent feature layers stacked. At this time, the width and height information is concentrated on the channel information, and the input channel is expanded four times [21]. The YOLOv8 model in adaptive multi-sample matching, concerning the YOLOX, adopts the Anchor-Free method and introduces a dynamic TaskAlignedAssigner matching strategy. The YOLO detection algorithm performs satisfactorily in detecting small and occluded targets in complex field environments and has better detection speed than other deep learning algorithms [22,23]. The single-stage detection algorithm deals with detection as a regression problem, compared to the two-stage detection algorithm that does not require a candidate region and can directly generate information such as category probabilities for the target. Although the single-stage detection method is faster in detection, the accuracy is relatively low.

Table 1. YOLO Series Version Comparison Table.

Model	Size	Train	mAP@0.5	FPS
YOLOv1	416 ×416	VOC2007+2012	63.4%	45
YOLOv2	416 × 416	VOC2007+2012	76.8%	67
YOLOv3	416 × 416	MS COCO	55.3%	35
YOLOv4	608 × 608	MS COCO	65.7%	62
YOLOv5-l	640 × 640	MS COCO	66.9%	73
YOLOX-l	640 × 640	MS COCO	68.5%	69
YOLOv7	640 × 640	MS COCO	69.7%	161

In deep learning network training, the more sufficient samples of the model, the stronger the generalization and the higher the robustness of the trained network model. However, for simple datasets, overfitting, weak generalization ability, and low robustness tend to occur when training the model due to too few samples. Fang presents an inpainting strategy called comparative sample augmentation, which enhances the quality of the training set by filtering irrelevant images and constructing additional images using information about the surrounding regions of the target image. This strategy managed to augment the datasets [24]. The datasets augmented by this strategy significantly reduce the probability of model overfitting during model training and effectively improve the model's generalization ability. While acquiring the dataset, all the image data inevitably have noise, such as granular speckles and discoloration. Zheng proposes a hybrid denoising CNN (HDCNN) [25]. HDCNN improves the quality of the image and makes the image more straightforward and more prosperous in detail. In many image processing and computer vision tasks, such as image recognition, target detection, etc., removing noise improves the accuracy and performance of the algorithm. Reducing noise reduces false detections and misjudgments and improves the algorithm's ability to understand and analyze the image. Ahmad presented a technique called CBIR-similarity measure via artificial neural network interpolation (CBIR-SMANN) [26]. CBIR-SMANN measures the similarity between an interpolated image and a target image by using an artificial neural network and then calculating the similarity between the interpolated image and the target image. The image can be interpolated from low resolution to high resolution, thus improving the quality and detail of the image. This interpolation technique can be used to generate a high-resolution image similar to the target image by learning the features of the image through a neural network and applying them to the target image. This ultimately improves the accuracy and performance of the algorithm in image recognition or target detection tasks.

The application of intelligent vehicles is to be realized in complex urban environments. The target detection technology in the field of automatic driving environment perception not only requires high-precision recognition ability but also needs to be able to respond in real-time to the complex and changing traffic scene to ensure that the driving system can make timely and accurate instructions. Although the two-stage detection algorithm

has high accuracy, its high time cost leads to slow detection speed which cannot meet the real-time requirements. The single-stage detection method is faster, but the accuracy could be higher, and security needs to be improved.

Synthesizing the above network model problems, this paper proposes the PBA-YOLOv7 network model algorithm based on the YOLOv7 model, and the main contributions are as follows:

1. Firstly, lightweight convolutional Partial Convolution (PConv)is introduced to optimize the ELAN module in the backbone to reduce the number of parameters and the number of visits in the network.
2. Introduce BiFusionNet network to replace the original PANet network of YOLOv7 in the feature fusion module of Neck to enhance the localization signal without increasing the computational burden.
3. Embedding the coordinate attention mechanism module before the representative convolution module in the Neck network to improve the detection accuracy of the network model while ensuring that the network is sufficiently lightweight.

The structure of this paper consists of four main parts. The first part details the key issues and challenges of target detection, mainly as applied to intelligent vehicle environment sensing detection, and methods to address these issues. The second part describes the structural components of YOLOv7, the structure of PBA-YOLOv7, and other principles and proposes in detail the improvement methods for each module. The third part is the experimental analysis, which analyzes the detection performance of each improved module in detail, evaluates and analyzes the experimental results, and compares them with similar methods. Finally, the fourth part summarizes the thesis and gives an outlook on the future research direction.

2. Methodology

YOLOv7 is a single-stage target detection algorithm that transforms the detection task into a regression problem [27]. Compared with other detection algorithms, YOLOv7 has faster detection speed and higher accuracy, which meets the requirements of real-time detection and recognition of targets in moving self-driving vehicles.

In this paper, we balance the detection speed and accuracy of the network model and choose the YOLOv7 network as the fundamental network model.

2.1. YOLOv7 Network Architecture

The YOLOv7 network mainly contains Input, Backbone, Neck, and Head [27]. Firstly, the image is pre-processed by Input and then sent to Backbone for feature extraction; then, the extracted features are processed by Neck feature fusion to obtain features of three sizes: large, medium, and small; and finally, the fused features are sent to the Head and output the result after detection.

The Backbone consists of multiple convolutions involving an Efficient Layer Aggregation Network (ELAN) module and a Maximum Pooling Convolution (MPConv) module. In the MPConv module, the MaxPool operation expands the sensory field of the current feature layer. Then, it fuses it with the feature information after the normal convolution process, which improves the network's generalization. The SPPCSPC module adds multiple MaxPool operations to the convolution in parallel to avoid distortion caused by image processing operations. The input image is subjected to feature extraction in Backbone to obtain three adequate feature layers for the following network construction.

The neck part of the three valid feature layers obtained from Backbone is used for feature fusion in this part using PANet. PANet network is added on top of FPN with an additional bottom-up path to shorten the information path of the low-level and top-level features, which helps to propagate accurate signals from the low-level features [28]. This part combines the feature information at different scales. It continues to extract features from the valid feature layers already obtained, up-sampling the features for feature fusion, and down-sampling the features again for feature fusion. Finally, RepConv is used to design

a heavily parameterized convolutional architecture that provides more gradient diversity for feature maps at different scales [29]. It increased the training time and improved the inference effect [30].

The Head part selects IDetect Head with three target sizes: large, medium, and trim. The Head is used as a classifier and regressor, and through the Backbone and Neck, three enhanced adequate feature layers are obtained, which are inputted into the Head for decoupling of feature information and outputting the position, confidence, and target type. The network structure of YOLOv7 is shown in Figure 1.

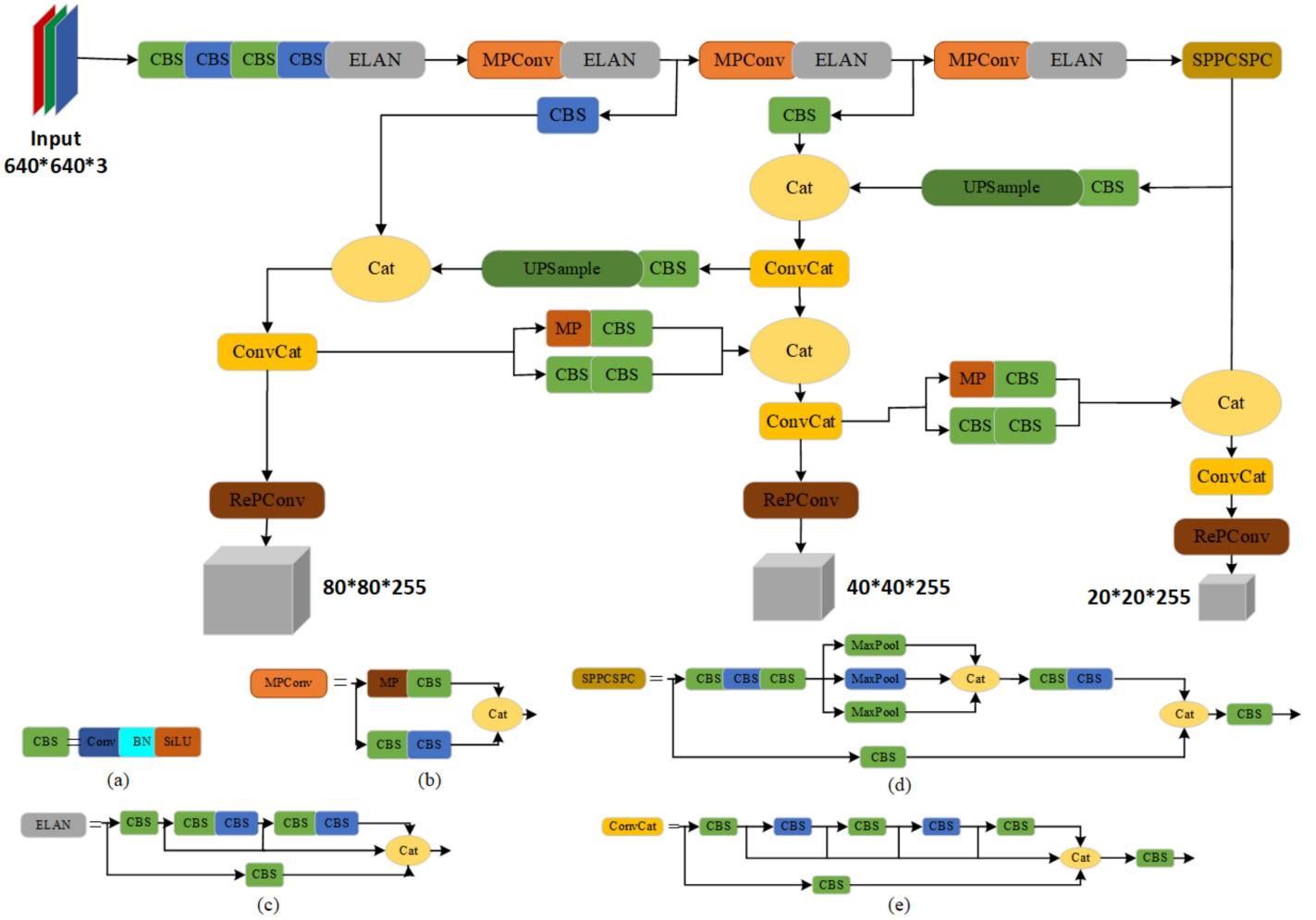


Figure 1. Network structure diagram of YOLOv7.

2.2. PBA-YOLOv7: Improved Algorithm for YOLOv7

In this paper, the algorithm is based on YOLOv7. Firstly, Partial Convolution (PConv) [31] is introduced in Backbone to optimize the Efficient Layer Aggregation Network (ELAN) module [32] in the Backbone extraction network to alleviate the number of parameters and the number of visits to the network and improve the detection speed of the network. Then, in the feature fusion part of Neck, we design the BiFusionNet network to optimize the feature pyramid network of YOLOv7 to better aggregate high-level semantic features and low-level semantic features, which improves the detection accuracy. Finally, the Coordinate Attention (CA) mechanism [33] is introduced to improve the detection accuracy of the network model while ensuring that the network is lightweight enough. The improved YOLOv7 network structure is shown in Figure 2.

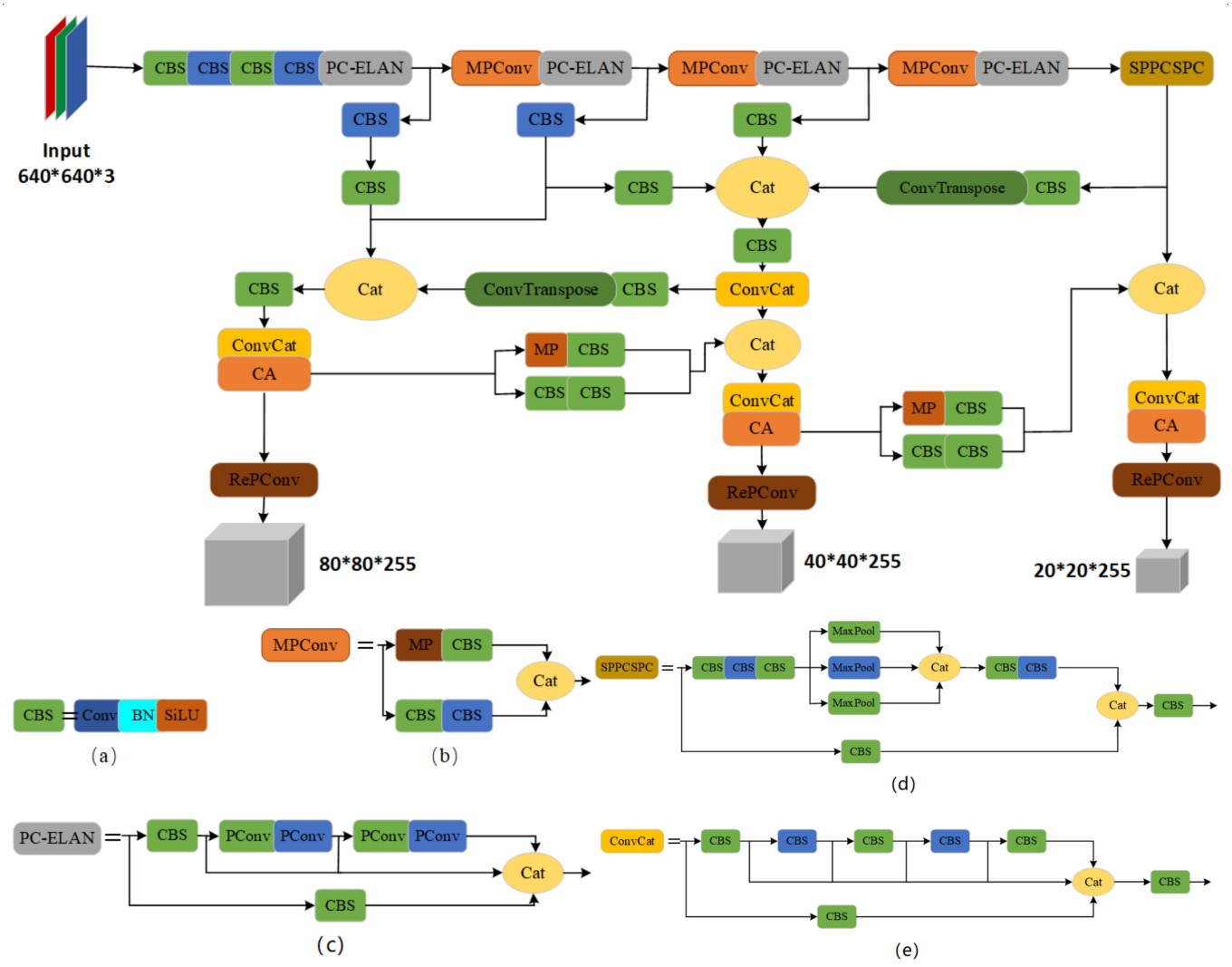


Figure 2. Improved YOLOv7 structure. (**a**) Structure diagram of CBS; (**b**) Structure diagram of MPConv; (**c**) Structure diagram of PC-ELAN; (**d**) Structure diagram of SPPCSPC; (**e**) Structure diagram of ConvCat.

2.2.1. Optimization of the ELAN Module

ELAN [32], mainly composed of VoVNet [34] and CSPNet [35], is an efficient layer aggregation network that optimizes the gradient length of the entire network using the stacking structure in the computational blocks. The network is designed to avoid the problems of using too many transition layers and the rapid lengthening of the shortest gradient paths in the whole network. The emergence of ELAN solves the problem of deterioration of the overall accuracy of the network due to reduced accuracy gains and even deterioration of the network convergence that occurs when the model scales, i.e., when the model reaches a certain depth and the stacking of the computational blocks is continued. However, the optimization of the ELAN network in terms of the number of parameters and the amount of computation is suboptimal, so in this paper, while ensuring the structural integrity of the ELAN network, we introduce the PConv, construct the PC-ELAN network module, and replace the convolutional kernel of 3×3 convolutional layers in the ELAN network by using the PConv. The structure of the PC-ELAN network is shown in Figure 1(c). Partial Convolution (PConv) is a new type of simple convolution to reduce computational redundancy while reducing memory access. Its working principle is shown in Figure 3.

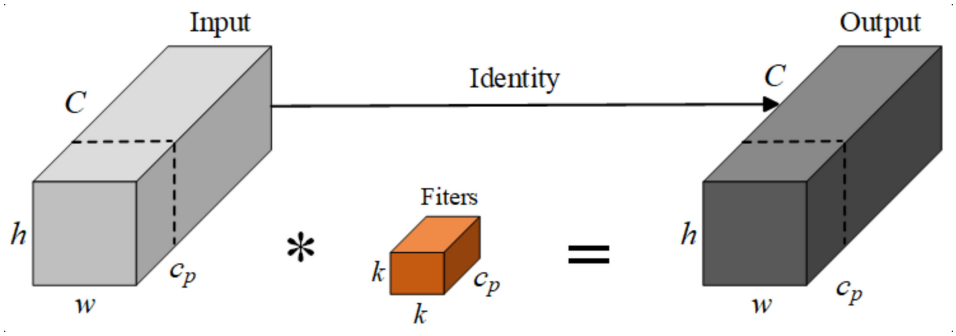


Figure 3. Partial convolution structure diagram.

PConv applies regular convolution for spatial feature extraction on only a portion of the input channels and keeps the rest of the channels unchanged, and for consecutive or regular memory accesses, computes the first or last consecutive channel as if it were representative of the entire feature map. The input and output feature maps have the same number of channels without loss of generality.

$$r = \frac{c_p}{C} \tag{1}$$

In the formula, $\boldsymbol{C}$ is the number of regular convolution channels and c_p is the number of PConv channels.

$$h \times \omega \times k^2 \times c_p^2 \tag{2}$$

The formula is the FLOPs calculation formula, where $\boldsymbol{h}$ is the height of the feature map, $\boldsymbol{w}$ is the width of the feature map, $\boldsymbol{k}$ is the width and height of the convolution kernel, and c_p is the number of PConv channels.

$$h \times \omega \times 2c_p + k^2 \times c_p^2 \approx h \times \omega \times 2c_p \tag{3}$$

The formula is the FLOPs calculation formula, where $\boldsymbol{h}$ is the height of the feature map, $\boldsymbol{w}$ is the width of the feature map, $\boldsymbol{k}$ is the width and height of the convolution kernel, and c_p is the number of PConv channels.

In this paper, $\boldsymbol{r} = 1/4$ is chosen, and from Equation (2), the computational amount of PConv convolution is reduced by 15/16 compared to conventional convolution, and from Equation (3), the memory access of PConv is reduced by 3/4 compared to conventional Conv.

Both the computational and memory accesses of the PC-ELAN network compared to the ELAN network have been drastically reduced. The overall structure of the network is effectively optimized, and the inference speed of the network is improved.

2.2.2. Building a Bidirectional Fusion Network (BiFusionNet) Module

It has been shown in many experiments that multi-scale feature aggregation is a critical component of target detection, and feature pyramid networks provide more accurate localization by aggregating high-level semantic features and low-level semantic features through top-down paths.

The YOLOv7 feature fusion network uses the PANet network structure [28]. PANet is an additional bottom-up path added on top of FPN. The network feature fusion better understands the contextual semantics of the target to shorten the information paths of both low-level and high-level features, which helps to propagate accurate localization signals from the low-level features. YOLOv6 [36] designed an enhanced feature fusion module, the BiC-Bidirectional Connection Module, which aggregates the effective feature maps of three neighboring layers to achieve enhanced localization signals without increasing the computational burden. In this paper, BiFusionNet is constructed based on the BiC bi-directional connectivity module principle. It better aggregates the three adequate feature

layers in the backbone extraction network and the depth feature layer in the neck. It retains more accurate localization signals compared to the PANet. A comparison of the structure of the PANet and BiFusionNet is shown in Figure 4.

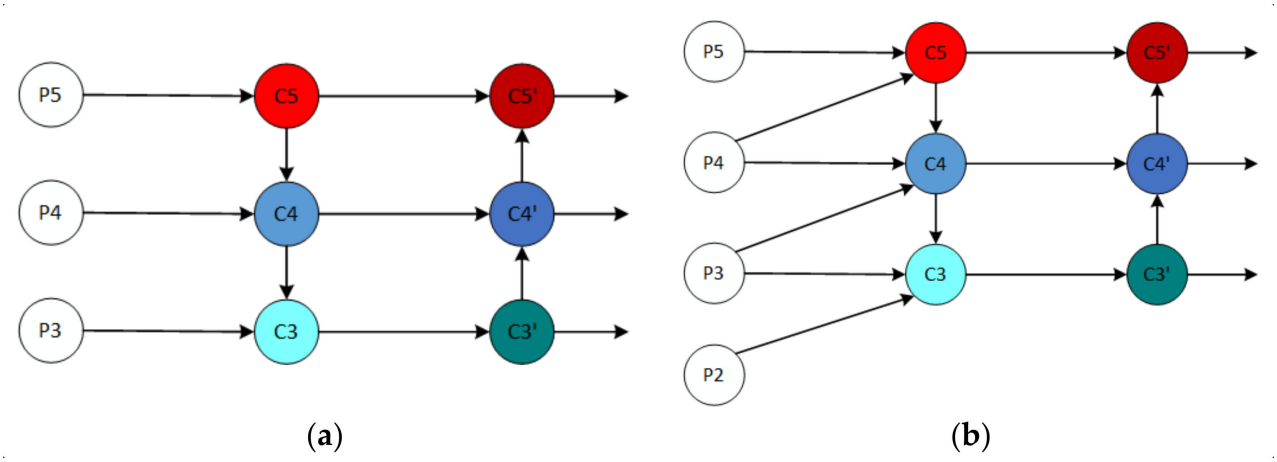


Figure 4. (**a**) Network structure of PANet; (**b**) Network structure of BiFusionNet. ***P*** is the low-level feature obtained by the main trunk network, and ***C*** is the high-level feature obtained by the neck network.

The BiFusionNet connection process is shown in Figure 5, where transposed convolution is introduced instead of up-sampling for feature recovery and image expansion to make the model more robust. Firstly, the same scale feature maps are downscaled using convolution with kernel 1. Next, the large-scale feature map is down-sampled using a convolution with convolution kernel 1 and down-sampled using a convolution with convolution kernel 3 and step size 2. Then, the small-scale feature map is up-sampled using a transposed convolution with convolution kernel 2. Finally, the feature maps obtained from the three parts are spliced and down-sampled again using a convolution with convolution kernel 1.

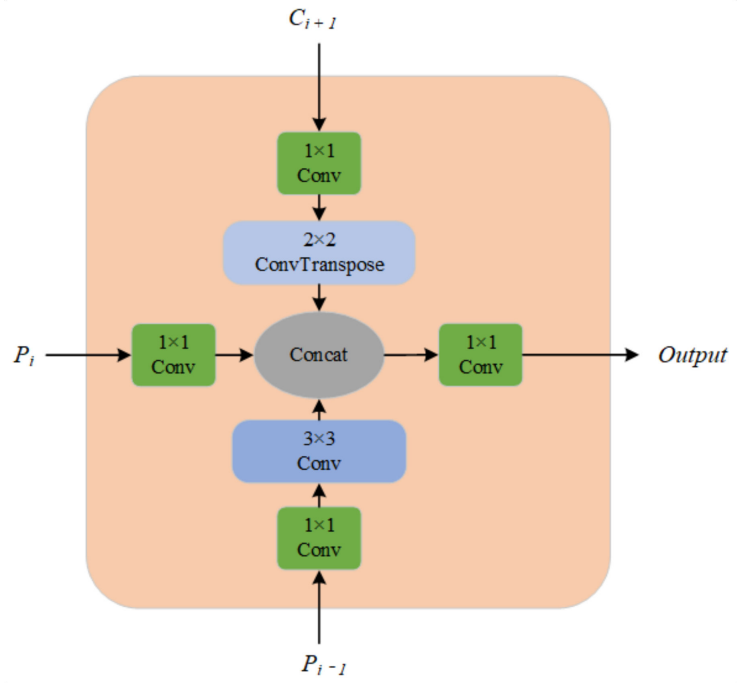


Figure 5. BiFusionNet Module connection flowchart. ***P*** is the low-level feature obtained by the main trunk network, and ***C*** is the high-level feature obtained by the neck network.

2.2.3. Coordinate Attention Mechanism

As the network structure becomes deeper and deeper and the layers are superimposed, although richer semantic information can be obtained, the resolution of the feature map decreases, resulting in the loss of some target location information. The attention mechanism is a standard data processing method widely used in machine learning tasks in various fields [37]. To ensure the network can extract rich semantic information while obtaining accurate location-aware information, the coordinate attention mechanism module is embedded before the representative convolution module in the Neck network.

The coordinate attention mechanism [33] can capture the associations and dependencies between different locations in an image by computing attention on the spatial coordinates of the features. This allows the model to understand better the spatial structure information in the image, which improves the understanding and perception of the target. The coordinate attention mechanism focuses on specific locations in the image and spatial relationships at different scales. This allows the model to adapt to targets and scenes at different scales, with better robustness to size, scale, and rotation changes in the image. The coordinate attention mechanism also allows the network model to better model the spatial relationships between different locations and more accurately model and understand the details and local structures in the image.

The structure of the coordinate attention mechanism is shown in Figure 6, which enhances the sensitivity to attentional information by decomposing the channel attention into a process of one-dimensional feature encoding performed in parallel to form a set of feature maps sensitive to both direction and position dimensions simultaneously.

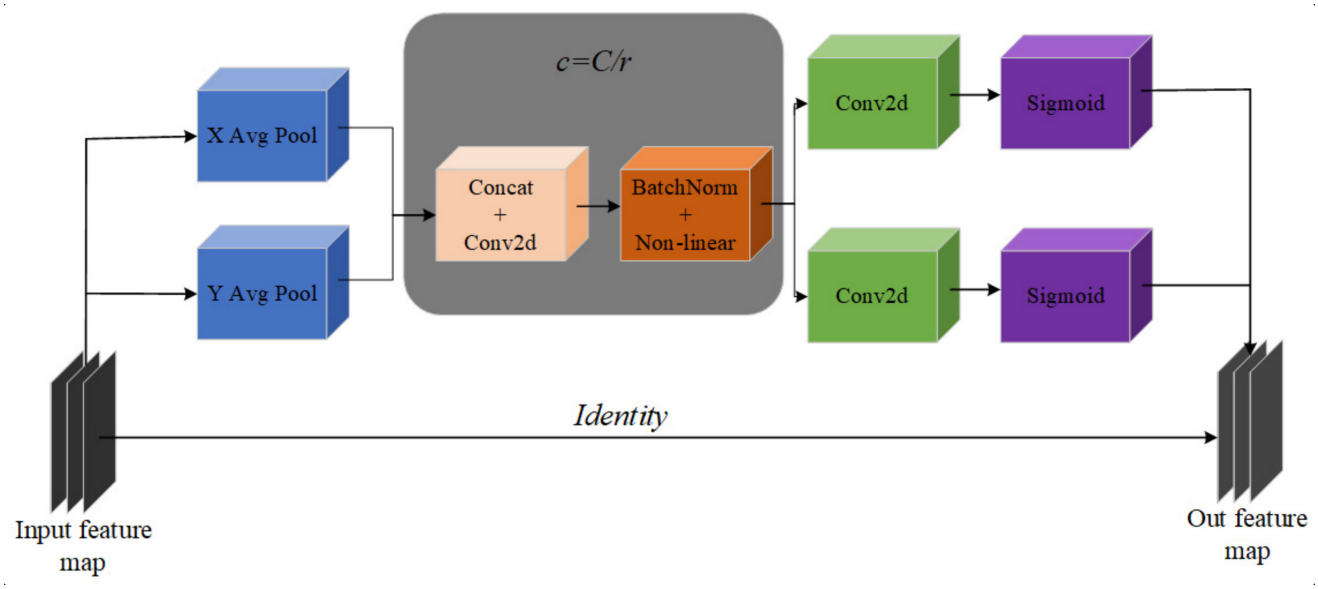


Figure 6. Structure diagram of coordinate attention mechanism.

The coordinate attention mechanism module decomposes the coordinate information into a set of one-dimensional feature codes by global pooling according to Equation (4).

$$Z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_c(i,j) \tag{4}$$

In the equation, Z_c is the output associated with the C channel, H and W are the height and width of the input feature map, respectively, and $x_c\ (i, j)$ is the input feature map of the C channel.

For an input tensor X, each channel is encoded in the horizontal coordinate direction using a pooling kernel of dimension (***H***, 1), and the output of the ***C***channel with height ***H*** can be represented as

$$Z_C^H(h) = \frac{1}{W} \sum_{0 \le i \le W} x_c(H, i) \tag{5}$$

In Equation $Z_C^H(h)$ is the output of the ***C***th channel with height ***H***. ***W*** is the width of the input feature map.

Encoding each channel in the vertical coordinate direction using a pooling kernel of dimension (1, ***W***), the output of the *C* channel of width ***W*** can be expressed as

$$Z_C^W(w) = \frac{1}{H} \sum_{0 \le \mathrm{j} \le H} x_c(j, W) \tag{6}$$

After Equations (5) and (6) are used to aggregate features in horizontal and vertical directions on the input, a pair of direction-sensitive attention features Z_H and Z_W are outputted. This step enables the coordinate attention module to acquire remote dependence in one spatial direction while preserving precise position information along another spatial direction, which helps the network to better acquire the global sensory field and encode more precise position information.

In the coordinate attention generation phase, the Z_H and Z_W attentional representations generated in the previous layer are first utilized in a cascade operation.

$$\mathrm{f} = \delta(\mathrm{F}[Z_H, Z_W]) \tag{7}$$

F is the transform function and δ is the h_swish() activation function. Applying Equation (7) yields $\mathrm{f} \in RC/r \times (H + W)$ as the intermediate feature mapping for encoding the coordinate features both horizontally and vertically, and *r* is the scaling parameter for down-sampling to reduce the number of channels in f.

3. Experimental Results and Analysis

3.1. Experimental Environment and Parameter Setting

The network experimental environment is ubuntu20.04, python3.7.0, pytorch1.13.0, and the related hardware configuration and model parameters are shown in Table 2.

Table 2. Hardware configuration and model parameters related to the experiment.

Parameter	Configuration	Parameter	Configuration
		Size of the picture	1242 × 375
GPU	RTX3080Ti	Learning rate	0.01
CPU	Core^{TM}i9-10900X	Batch size	8
CUDA	11.3	Workers	8
CUDNN	11.3	Optimizer	Adam
		Epochs	100

Due to the YOLOv7 algorithm model layers being more profound, the number of parameters is more significant, so the processor with higher computing power is used to maximize the GPU utilization and set the batch size to eight. and workers to eight. The model uses the Pytorch framework and the learning rate is 0.01. After the number of training iterations of 100, the optimal weights file of the model is obtained.

3.2. KITTI's Dataset

The KITTI dataset was co-founded in 2012 by the Karlsruhe Institute of Technology (KIT) in Karlsruhe, Germany, and by the Toyota Technological Institute at Chicago (TTI-C) in Chicago, U.S.A., and it is one of the most commonly used international datasets for evaluating computer vision algorithms in autonomous driving scenarios.

This paper conducts experiments on the KITTI's dataset, a commonly used and publicly available target detection dataset, for training and validation. A total of 7481 images are selected, which contain three categories: Car, Pedestrian, and Cyclist. This total of 7481 images are randomly divided into training and validation sets according to a 9:1 ratio and finally imported into the test sets with KITTI's dataset. Care is taken to ensure there are no duplicate images between the training set, validation set, and test set to prevent overfitting of the model [38]. There are 6733 images in the training set, 748 in the validation set, and 7518 in the test set.

3.3. Object Detection Network Comparison Experiment Results

In selecting the object detection model, we apply the existing popular object detection model to the KITTI's dataset for training and testing. We also compare the average accuracy mAP@0.5 and detection speed FPS as evaluation indicators. Finally, YOLOv7 was selected as the target detection model for subsequent experiments. The experimental results are shown in Table 3.

Table 3. Comparison of object detection algorithms.

Model	mAP@0.5	FPS
SSD	44.03%	39
Faster R-CNN	55.2%	17
YOLOv5l	87.4%	40
YOLOv7-tiny	85.3%	48
YOLOv7	90.5%	37

As can be seen from Table 3, the detection performance of YOLOv7 is better than other detection algorithms. For example, the mAP@0.5 of the YOLOv7 algorithm is 46.47% higher than that of the SSD algorithm. Although the FPS of the YOLOv7-tin algorithm is higher than that of YOLOv7, the mAP@0.5 is significantly lower than that of YOLOv7 by 5.2%. To better weigh the detection accuracy and detection speed of the model, we chose YOLOv7 as the target detection algorithm used in the experiment.

3.4. Evaluation Metrics

To accurately evaluate the superiority of the algorithm, the evaluation metrics used in this study are mean Average Precision (mAP), Frames Per Second (FPS), model size, and number of parameters.

1. mAP: reflects the mean of the detection accuracy of all target categories in the dataset, calculated as follows.

$$\mathrm{m}AP = \frac{1}{C}\sum_{i=1}^{C} AP_{\mathrm{i}} \tag{8}$$

   In the formula $\boldsymbol{C}$ denotes the number of all categories in the dataset, $\boldsymbol{i}$ denotes the number of detections, and $\boldsymbol{AP}$ denotes the average of single-target detection accuracies.
2. FPS: The number of frames per second transmitted by the model, reflecting the processing speed of the model.
3. Params: the number of parameters that reflect the model's memory footprint.
4. Model Size: reflects the size of the memory occupied by the model in M.

3.5. Ablation Experiment and Analysis

To verify the improvement effect of the algorithm in this paper, ablation experiments are used to verify each improved scheme under the condition that the environment and parameter settings are kept uniform. Improvement point ablation experiments are carried out in seven groups, with YOLOv7 as the baseline model, "$\surd$" indicates that the corresponding improvement point is selected, and each improvement point is first added to

the original model of YOLOv7 individually and sequentially to carry out the ablation experiments. Model 6 is the model of this paper's algorithm (PFA-YOLOv7) and the results of the experiments are shown in Table 4. The comparison diagram of the model of the ablation experiments is shown in Figure 7.

Table 4. Improved point ablation experiment.

Model	YOLOV7	PConv	BiFusionNet	CA	Parameters	Size	mAP@0.5	mAP@0.5:0.95	FPS
1	√				37,208 K	74.8 M	90.5%	60.5%	37
2	√	√			32,699 K	66.5 M	91.8%	62.4%	45
3	√		√		37,113 K	73.6 M	92.8%	65.3%	41
4	√			√	37,253 K	75.2 M	94.4%	69.1%	32
5	√	√	√		32,604 K	65.5 M	92.0%	63.4%	47
6	√	√	√	√	32,640 K	65.7 M	94.5%	68.3%	43

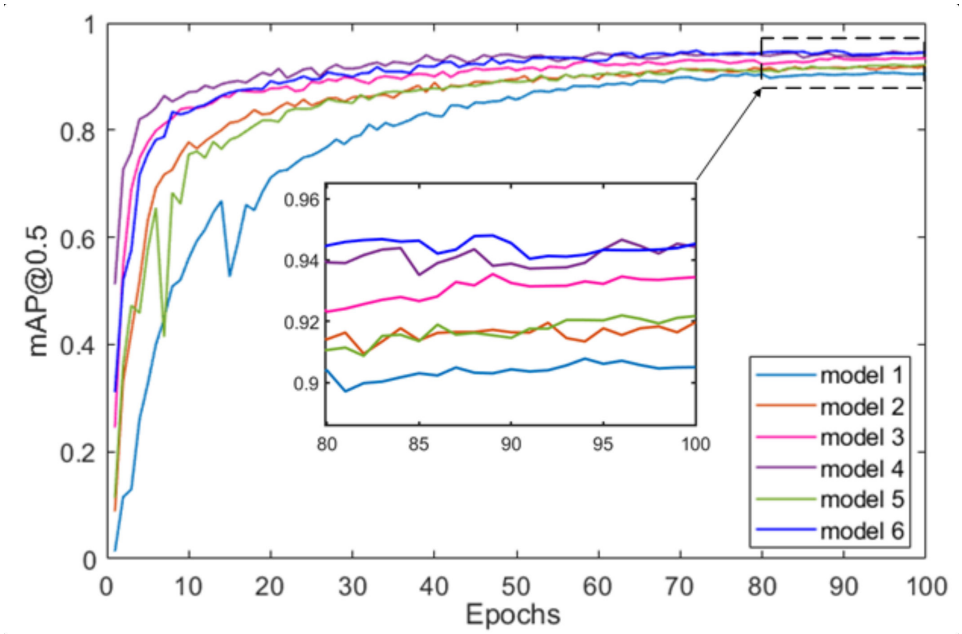


Figure 7. Comparison of experimental data of improved point ablation mAP@0.5.

The following conclusions can be drawn from Table 4:

1. Model 1 results from the original YOLOv7 algorithm experiment, a comparison benchmark for the following sets of experiments. Its parameter count is 37,208 K, model size is 74.8 M, mean average precision mAP@0.5 is 90.5%, mAP@0.5:0.95 is 60.5%, frames precision second (FPS) is 37.
2. Model 2 is the ELAN module that introduces a partial convolution to improve the backbone network based on model 1, which not only reduces the number of parameters and the size of the model relative to model 1, but also improves the mean average precision and frames precision second based on this model.
3. Model 3 is based on model 1, wherein only the BiFusionNet module is replaced, and this model keeps the number of parameters unchanged, based on which the mean average precision and frames precision second is substantially improved.
4. Model 4 is based on model 1 and introduces the coordinate attention mechanism. This model improves the mean average precision while reducing the frame's precision by a second.
5. Model 5 is based on model 1, which sequentially introduces the ELAN module for partial convolutional improvement of the backbone network and introduces the BiFusionNet module in the enhanced feature extraction network. This model

improves the evaluation indexes relative to Model 1 and Model 2 and improves the frame's precision second relative to Model 3 while reducing the number and size of the parameters of the model to ensure that the mean average precision does not decrease significantly.

6. Model 6 is the algorithm of this paper, which introduces the ELAN module of the partially convolutional improved backbone network, BiFusionNet module, and coordinate attention mechanism for the enhanced feature extraction network which is based on model 1.

Relative to model 1, the parameter of parameters is reduced by 4568 K, the size of the model is reduced by 9.1 M, the mAP@0.5 is significantly improved by 4%, the mAP@0.5:0.95 is significantly improved by 7.8%, and six frames significantly increase the FPS. The experiments show that the algorithm in this paper substantially improves all evaluation indexes compared with the original YOLOV7 algorithm, which not only balances the model detection speed and detection accuracy to meet the demand of real-time detection but also minimizes the parameters and size of the model.

Relative to Model 2, the number of parameters is reduced by 59 K, the model size is reduced by 0.8 M, the mAP@0.5 is improved by 2.7%, the mAP@0.5:0.95 value is improved by 5.9%, and two frames reduce the FPS. Model 6 is based on Model 2, first replacing the PANet module with the BiFusionNet module as a path-combining network and later introducing the coordinate attention mechanism. Although the complexity of the model computation is increased, the inspection accuracy is significantly improved with the reduction of the detection speed FPS, and the number of parameters and the model size are also reduced to different degrees.

Relative to Model 3, the parameter quantity was reduced by 4473 K, the model size was reduced by 7.8 M, the mAP@0.5 was increased by 1.7%, the mAP@0.5:0.95 value was increased by 3%, and two frames increased the FPS. Model 6, based on model 3, has introduced part of the convolution to improve the ELAN module of the backbone network and the coordinate attention mechanism, which on the whole reduces the computational complexity of the model as well as makes the model obtain richer semantic information about the network. The experiments have shown that the various evaluation indexes of the model have been improved, which proves the necessity of the experimental improvement.

Relative to Model 4, the number of parameters is reduced by 4613 K, the model size is reduced by 9.5 M, the mAP@0.5 is increased by 0.1%, the mAP@0.5:0.95 value is reduced by 0.8%, and nine frames increase the FPS. Model 6, based on model 4, also successively introduces the ELAN module and BiFusionNet module of a partially convolutional improved backbone network as an enhanced feature extraction network, which reduces the complexity of the model and makes the model better aggregate high-level and low-level semantic information. Experiments show that under the premise of guaranteeing detection accuracy, the detection speed of the model is significantly improved, and the model size and number of parameters are minimized. The model size and the number of parameters are minimized to meet the real-time demand of model detection.

Compared with Model 5, the number of parameters increased by 36 K, the model size increased by 0.2 M, the mAP@0.5 increased by 2.5%, the mAP@0.5:0.95 value increased by 4.9%, and the FPS decreased by four frames. Model 6 introduces the coordinate attention mechanism based on model 5, which increases the complexity of the model computation, delays the detection speed, and consequently increases the number of parameters and model size slightly, but the detection accuracy is significantly improved.

The algorithm balances detection accuracy and detection speed and maximizes the detection accuracy to meet the real-time demand.

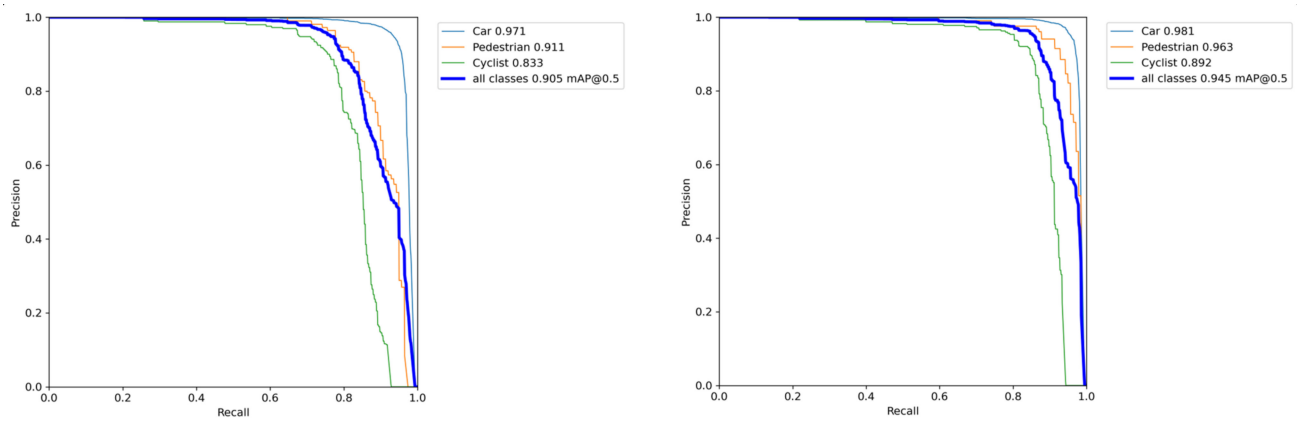
3.6. Comparative Experiments and Analysis of PBA-YOLOv7 and Baseline Model YOLOv7

The performance metrics of this paper's model PBA-YOLOv7 and the baseline model YOLOv7 are compared on KITTI's dataset, as shown in Table 5.

Table 5. PBA-YOLOv7 and baseline model YOLOv7 comparison experiments.

Model	Parameters	Size	mAP@0.5	mAP@0.5:0.95	FPS
YOLOv7	37,208 K	74.8 M	90.5%	60.5%	37
PBA-YOLOv7	32,640 K	65.7 M	94.5%	68.3%	43

From Table 5, it can be seen that the PBA-YOLOv7 model has better parameters, model size, mAP@0.5, mAP@0.5:0.95, and FPS than the YOLOv7 model where the parameters and model size are reduced by 4568 K and 9.1 M, respectively, and mAP@0.5, mAP@0.5:0.95, and FPS are higher by 4%, 7.8%, and six FPS, respectively. The Precision-Recall curve of the two models is shown in Figure 8 and the test results are shown in Figure 9.



(**a**) (**b**)

Figure 8. Precision-Recall curve for the model. (**a**) YOLOv7; (**b**) PBA-YOLOv7.

(**a**)

(**b**)

Figure 9. Comparison chart of test results on the KITTI dataset. (**a**) YOLOv7; (**b**) PBA-YOLOv7.

As can be seen from the detection results in Figure 9, compared with the baseline model YOLOv7, the introduction of BiFusionNet instead of PANet as the feature fusion network better aggregates the effective feature layer in the backbone extraction network and the depth feature layer in the neck to enhance the localization signals in the target detection, and finally the introduction of the coordinate attention mechanism significantly improves the model detection accuracy, and also effectively reduces the leakage detection rate.

In order to verify the superiority of the model performance in this paper, we used the PBA-YOLOv7 and baseline model YOLOv7 algorithms to train, validate, and test the VOC2012 dataset under the same environment and experimental conditions. A total of 21,503 images which contained 20 categories were selected. The 21,503 images were randomly divided into the training set, validation set, and test set according to a ratio of 7:2:1, where the training set, validation set, and test set are 15,052, 4301, and 2150, respectively. The experimental results are shown in Table 6.

Table 6. PBA-YOLOv7 and baseline model YOLOv7 comparison experiments.

Model	Parameters	Size	mAP@0.5	mAP@0.5:0.95	FPS
YOLOv7	37,208 K	75.1 M	77.3%	46.7%	35
PBA-YOLOv7	32,990 K	66.9 M	85.7%	53.6%	44

The data in Table 6 shows that the PBA-YOLOv7 model is 8.4% and 6.9% higher than the original YOLOv7 model with mAP@0.5 and mAP@0.5:0.95, and the detection speed is nine FPS higher than the original model. The parameters and the model sizes have been reduced by 4218 K and 8.2 M, respectively. The results of the PBA-YOLOv7 and the YOLOv7 algorithms on VOC2012 are shown in Figure 10.

(**a**)

(**b**)

Figure 10. Comparison chart of test results on the VOC2012 dataset. (**a**) YOLOv7; (**b**) PBA-YOLOv7.

3.7. Experimental Comparison of Different Models

In order to verify the superiority of this paper's algorithm (PFA-YOLOv7), the proposed algorithm is compared and experimented with other popular network algorithms to ensure the identical configuration environment and training parameters. The single-stage detection algorithms SSD, YOLOv5l, YOLOv7-tiny, YOLOX-l, YOLOv8n, and the two-stage detection algorithm Faster R-CNN are selected as the comparison algorithms. The experimental results are shown in Table 7.

Table 7. Comparative experiment.

Model	mAP@0.5	FPS
SSD	44.03%	39
Faster R-CNN	55.2%	17
YOLOv5-l	87.4%	40
YOLOv7-tiny	85.3%	48
YOLOX-l	82.9%	31
YOLOv8n	89.3%	68
Algorithm of this paper	94.5%	43

The Faster R-CNN algorithm uses a region proposal network (RPN) to select candidate regions from the features extracted by the backbone network. Then, it extracts the information of the candidate regions for detection in the second stage. However, the mAP and FPS are lower than the algorithm proposed in this paper (PFA-YOLOv7), and it is worth mentioning that the FPS of Faster R-CNN is only 17 frames.

Compared with the Faster R-CNN algorithm, the SSD algorithm does not need to select candidate regions and can simultaneously localize and classify targets in one forward propagation by predicting targets at different scales on different feature scales. Although the FPS is fast enough, its mAP0.5 is relatively low.

For the YOLOv5l algorithm, the method uses PANet as a path-binding network to realize the fusion of feature maps at different scales, and the use of a lightweight network structure makes the detection fast enough. Although the FPS is 40, it is reduced by 7.1% compared to the mAP@0.5 algorithm in this paper.

YOLOv7-tiny uses a more lightweight network structure, and despite an FPS of 48, the mAP0.5 is significantly lower.

YOLOX-l, like YOLOv5, uses the Focus network structure in the backbone part, which reduces the number of parameters to be computed and improves the model's speed. However, its mAP0.5 is then 3% lower relative to YOLOv7-tiny. It is significantly lower by 11.6% compared to the modeling algorithm in this paper.

YOLOv8, currently the newest detection method in the YOLO family, is fast enough. However, its average accuracy could be more significant, with a mAP0.5 that is 5.2% lower relative to the model algorithm PBA-YOLOv7 in this paper.

In conclusion, the algorithm in this paper has a more significant performance in target detection compared to other algorithms.

4. Conclusions

Deep learning-based target detection methods address the problem of how to trade off the target detection accuracy and detection speed of the model. In this paper, we take the YOLOv7 network as the baseline model and propose an improved YOLOv7 network model. Firstly, we introduce PConv to optimize and improve the ELAN module in the backbone network, in order to reduce the model size and parameter counts significantly and improve the detection speed of the network under the guarantee of the model detection accuracy. In the path aggregation network, the BiFusionNet network is designed to replace the PAnet network, which better fuses and utilizes the shallow and deep information of the network, as well as retaining the rich feature information and more accurate localization information of the shallow network targets, which further improves the detection accuracy

of the network. Finally, the coordinate attention mechanism is introduced to ensure that the network can extract the rich semantic information and at the same time obtain accurate location-aware information, which makes the detection accuracy of the network further improved. Experimental results show that the improved YOLOv7 algorithm exhibits good robustness compared to the original YOLOv7 model and other comparative models in the public KITTI's dataset test, thus proving the effectiveness and superiority of the algorithm proposed in this paper.

In the future, the network structure algorithm will continue to be optimized to meet the complex and changing environment of automated vehicle driving. Furthermore, when a new domain dataset is used, the model in this paper can be used as a pre-trained model to efficiently and quickly adapt to the new dataset, thus improving the detection performance of new domain targets.

Author Contributions: Conceptualization, Y.S.; methodology, Y.S.; software, S.L. and Y.Z.; validation, Y.L.; investigation, H.N. and Y.Z.; data curation, Y.Z.; writing—original draft preparation, Y.L.; writing—review and editing, Y.L. and Y.S.; visualization, Z.D. and Y.L.; supervision, Y.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the Natural Science Foundation of Hebei Province (Grant No. F 2021402011).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used and analyzed during the current study are publicly available datasets that can be downloaded independently or obtained from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Liu, W.G. The challenges of autonomous driving. *Intell. Connect. Cars* **2019**, *1*, 58.
2. Lowed, D.G. Distinctive image features from scale-invariant key points. *Int. J. Comput. Vis.* **2004**, *60*, 91–110.
3. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–26 June 2005; Volume 1, pp. 886–893.
4. Felzenszwalb, P.; McAllester, D.; Ramanan, D. A discriminatively trained, multiscale, deformable part model. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.
5. Hearst, M.A.; Dumais, S.T.; Osuna, E.; Platt, J.; Scholkopf, B. Support vector machines. *IEEE Intell. Syst. Appl.* **1998**, *13*, 18–28. [CrossRef]
6. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Springer International Publishing: Cham, Switzerland, 2014; pp. 740–755.
7. Wang, P.; Bayram, B.; Sertel, E. A comprehensive review on deep learning based remote sensing image super-resolution methods. *Earth-Sci. Rev.* **2022**, *232*, 104110. [CrossRef]
8. Bharati, P.; Pramanik, A. Deep learning techniques—R-CNN to mask R-CNN: A survey. In *Computational Intelligence in Pattern Recognition: Proceedings of CIPR 2019*; 2020; pp. 657–668; Available online: https://www.researchgate.net/publication/338302306_Computational_Intelligence_in_Pattern_Recognition_Proceedings_of_CIPR_2019_Proceedings_of_CIPR_2019 (accessed on 13 September 2023).
9. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
10. Ren, S.; He, K.; Girshick, R. Faster r-cnn: Towards real-time object detection with region proposal networks. *In Advances in Neural Information Processing Systems*. 2015, Volume 28. Available online: https://arxiv.org/abs/1506.01497 (accessed on 13 September 2023).
11. He, K.; Gkioxari, G.; Dollár, P. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
12. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.

13. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef] [PubMed]
14. Jiang, P.; Ergu, D.; Liu, F.; Cai, Y.; Ma, B. A Review of Yolo algorithm developments. *Procedia Comput. Sci.* **2022**, *199*, 1066–1073. [CrossRef]
15. Liu, W.; Anguelov, D.; Erhan, D. Ssd: Single shot multibox detector. In Proceedings of the Computer Vision–ECCV, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
16. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
17. Li, G.; Suo, R.; Zhao, G.A.; Gao, C.Q.; Fu, L.S.; Shi, F.X.; Dhupia, J.; Li, R.; Cui, Y.J. Real-time detection of kiwifruit flower and bud simultaneously in orchard using YOLOv4 for robotic pollination. *Comput. Electron. Agric.* **2022**, *193*, 106641. [CrossRef]
18. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
19. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804-02767.
20. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
21. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:2107.08430.
22. Yan, B.; Fan, P.; Lei, X.Y.; Liu, Z.J.; Yang, F.Z. A Real-Time Apple Targets Detection Method for Picking Robot Based on Improved YOLOv5. *Remote Sens.* **2021**, *13*, 1619. [CrossRef]
23. Lu, S.Y.; Wang, B.Z.; Wang, H.J.; Chen, L.H.; Ma, L.J.; Zhang, X.Y. A real-time object detection algorithm for video. *Comput. Electr. Eng.* **2019**, *77*, 398–408. [CrossRef]
24. Fang, B.; Jiang, M.; Shen, J.; Stenger, B. Deep generative inpainting with comparative sample augmentation. *J. Comput. Cogn. Eng.* **2022**, *1*, 174–180. [CrossRef]
25. Zheng, M.; Zhi, K.; Zeng, J.; Tian, C.; You, L. A hybrid CNN for image denoising. *J. Artif. Intell. Technol.* **2022**, *2*, 93–99. [CrossRef]
26. Ahmad, F. Deep image retrieval using artificial neural network interpolation and indexing based on similarity measurement. *CAAI Trans. Intell. Technol.* **2022**, *7*, 200–218. [CrossRef]
27. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 7464–7475.
28. Wang, K.; Liew, J.H.; Zou, Y. Panet: Few-shot image semantic segmentation with prototype alignment. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9197–9206.
29. Ding, X.H.; Zhang, X.Y.; Ma, N.N.; Han, J.G.; Ding, G.G.; Sun, J. RepVGG: Making VGG-style ConvNets Great Again. *arXiv* **2021**, arXiv:2101.03697.
30. Ding, X.H.; Hao, T.X.; Tan, J.C.; Liu, J.; Han, J.G.; Guo, Y.C.; Ding, G.G. ResRep: Lossless CNN Pruning via Decoupling Remembering. *arXiv* **2021**, arXiv:2007.03260.
31. Chen, J.; Kao, S.; He, H. Run, Don't Walk: Chasing Higher FLOPS for Faster Neural Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 12021–12031.
32. Zhang, X.; Zeng, H.; Guo, S. Efficient long-range attention network for image super-resolution. In *Computer Vision–ECCV*; Springer Nature: Cham, Switzerland, 2022; pp. 649–667.
33. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13713–13722.
34. Lee, Y.; Hwang, J.; Lee, S. An energy and GPU-computation efficient backbone network for real-time object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–17 June 2019.
35. Wang, C.Y.; Liao, H.Y.M.; Wu, Y.H. CSPNet: A new backbone that can enhance learning capability of CNN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 13–19 June 2020; pp. 390–391.
36. Li, C.; Li, L.; Geng, Y. YOLOv6 v3. 0: A Full-Scale Reloading. *arXiv* **2023**, arXiv:2301-05586.
37. Niu, Z.; Zhong, G.; Yu, H. A review on the attention mechanism of deep learning. *Neurocomputing* **2021**, *452*, 48–62. [CrossRef]
38. Gu, Y.; Wang, S.C.; Yan, Y.; Tang, S.J.; Zhao, S.D. Identification and Analysis of Emergency Behavior of Cage-Reared Laying Ducks Based on YoloV5. *Agriculture* **2022**, *12*, 485.