electronics

MDPI

Article

A Multi-Scale Traffic Object Detection Algorithm for Road Scenes Based on Improved YOLOv5

Ang Li [1,†], Shijie Sun [1,†], Zhaoyang Zhang [1,*], Mingtao Feng [2,3], Chengzhong Wu [3] and Wang Li [4]

[1] School of Information Engineering, Chang'an University, Xi'an 710064, China
[2] School of Computer Science and Technology, Xidian University, Xi'an 710000, China
[3] National Engineering Laboratory of Robot Visual Perception and Control Technology, Hunan University, Changsha 410000, China
[4] CRRC Zhuzhou Electric Locomotive Co., Ltd., Zhuzhou 412000, China
* Correspondence: zhaoyangzhang@chd.edu.cn; Tel.: +86-137-5993-9057
† These authors contributed equally to this work.

Abstract: Object detection in road scenes is a task that has recently become popular and it is also an important part of intelligent transportation systems. Due to the different locations of cameras in the road scenes, the size of the traffic objects captured varies greatly, which imposes a burden on the network optimization. In addition, in some dense traffic scenes, the size of the traffic objects captured is extremely small and it is easy to miss detection and to encounter false detection. In this paper, we propose an improved multi-scale YOLOv5s algorithm based on the YOLOv5s algorithm. In detail, we add a detection head for extremely small objects to the original YOLOv5s model, which significantly improves the accuracy in detecting extremely small traffic objects. A content-aware reassembly of features (CARAFE) module is introduced in the feature fusion part to enhance the feature fusion. A new SPD-Conv CNN Module is introduced instead of the original convolutional structure to enhance the overall computational efficiency of the model. Finally, the normalization-based attention module (NAM) is introduced, allowing the model to focus on more useful information during training and significantly improving detection accuracy. The experimental results demonstrate that compared with the original YOLOv5s algorithm, the detection accuracy of the multi-scale YOLOv5s model proposed in this paper is improved by 7.1% on the constructed diverse traffic scene datasets. The improved multi-scale YOLOv5s algorithm also maintains the highest detection accuracy among the current mainstream object detection algorithms and is superior in accomplishing the task of detecting traffic objects in complex road scenes.

Keywords: road scenes; object detection; YOLOv5; multi-scale; attention mechanism

Citation: Li, A.; Sun, S.; Zhang, Z.; Feng, M.; Wu, C.; Li, W. A Multi-Scale Traffic Object Detection Algorithm for Road Scenes Based on Improved YOLOv5. *Electronics* **2023**, *12*, 878. https://doi.org/10.3390/electronics12040878

Academic Editor: Dah-Jye Lee

Received: 10 January 2023
Revised: 4 February 2023
Accepted: 6 February 2023
Published: 9 February 2023

1. Introduction

The detection of traffic objects in road scenes is a critical part of intelligent transport systems and a key technology in the achievement of autonomous driving. Good real-time traffic object detection and recognition is essential for environment awareness in road scenes. Traffic object detection in intelligent transportation systems is usually divided into four categories: vehicle detection, pedestrian detection, traffic sign detection and other obstacle detection. Due to the rapid development of deep learning methods in recent years, object detection methods can be broadly classified into two main categories: traditional object detection methods and deep learning-based object detection methods.

The core idea of traditional object detection methods is to generate the corresponding artificial feature information from the image based on the characteristics of the target itself and then use these features for object detection. Objects in traffic scenes often contain a large number of regular features, such as the color and model of a car, the posture and limb structure of a pedestrian, the shape of a traffic sign, etc. This rule has given rise to a number of object detection algorithms based on edge feature information. Matthews et al. [1]

detect distinct vertical edges from the image and combine this with under-vehicle shadow detection to determine the left, right and lower boundaries of the vehicle for vehicle detection and recognition. You et al. [2] used HOG features and CIE-LUV histograms as low-level features and proposed an extended filter channel framework using the concept of filter channel features to improve the accuracy of pedestrian detection on multiple datasets. Stefan et al. [3] exploited self-similarity on the color channel to improve the detection performance of still images and video sequences in the dataset, with a 20% performance improvement in pedestrian detection when combined with HOG features. Traditional object detection methods are built on manually designed feature representations and shallow trainable architectures and the algorithms are prone to performance bottlenecks when multiple low-level image features are combined with contextual information from the target detector or scene classifier.

Deep learning-based object detection methods have a large accuracy improvement over traditional methods and are now the mainstream in this field. Deep learning methods are characterized by the introduction of semantic and deep-level features that can be learned, which can compensate for the shortcomings of traditional object detection methods. In recent years, object detection methods based on convolutional neural network have developed rapidly and achieved significant results [4–8]. The release of public datasets such as ImageNet [9], COCO [10], VOC [11] and KITTI [12] has greatly promoted the development of object detection applications. CNN-based object detectors can be divided into two types: (1) one-stage detectors: YOLO9000 [13], YOLOv3 [14], YOLOv4 [15], Scaled-YOLOv4 [16], YOLOv5, YOLOX [17], FCOS [18], DETR [19], etc.; (2) two-stage detectors: Faster R-CNN [4], VFNet [8], CenterNet2 [20], etc. Two-stage detectors require a network to find possible object regions in images and then a network to classify objects. Two-stage detectors such as Faster R-CNN have high detection accuracy in object detection tasks, but their detection speed is slow, and does not meet the real-time requirement of object detection in intelligent transportation systems. The YOLO series [5,13–17] is a typical one-stage detector that demonstrates excellent performance in object detection tasks. The YOLO model takes into account the advantages of speed and precision, and is our first choice for object detection tasks in traffic scenes.

However, the traditional YOLO model is designed for object detection tasks in natural scenes and there are several main problems with using previous models directly to perform object detection on images of traffic scenes, which are intuitively illustrated by some cases in Figure 1. Firstly, traffic scene images are often captured by cameras set up at various intersections and the different camera angles lead to large variations in target size, which can easily lead to missed and false detections. Secondly, due to hardware specifications and lighting conditions, the captured images may have low resolution and blurrier objects. Thirdly, the large coverage area of the camera results in images containing a large number of complex backgrounds, resulting in extremely small sized objects that are difficult to detect. These problems result in the traditional YOLO model performing poorly in traffic scene images and cannot be directly applied to object detection tasks in traffic scenes. Yu et al. [21] used an improved YOLOv3 model to detect traffic lights in traffic scenes and achieved good results, but the model could not be applied to object detection tasks in various complex traffic scenes. Zhu et al. [22] proposed a new multi-sensor and multi-level enhanced convolutional network structural model, MME-YOLO, for object detection in complex traffic scenes, but the detection accuracy of extremely small objects in traffic scenes was not high. Li et al. [23] proposed an Attention-YOLOv4, which introduced the attention mechanism to improve the detection accuracy of small target objects, but the detection ability of objects in low-resolution images was insufficient. Mittal et al. [24] proposed a hybrid model of Faster R-CNN and YOLO and established a rich traffic scene dataset for vehicle object detection and traffic flow detection and achieved good results.

Figure 1. Intuitive examples to explain the three main problems of object detection on traffic scene images. The cases in the first, second and third rows show the problems of large object size variations, blurred images and tiny object sizes that are difficult to detect, respectively. (The Chinese words in the image are the time and place recorded by the surveillance video.)

In this paper, we propose an improved model, multi-scale YOLOv5s based on YOLOv5s to solve the three problems presented above. The overview of the multi-scale YOLOv5s model is shown in Figure 2. We, respectively, use CSPDarknet53 [25] as the backbone and use FPN [26]+PAN [27] as the neck of multi-scale YOLOv5s. In the original YOLOv5 model, three detection heads are included, which are, respectively, used for the detection of small, medium and large objects. In complex traffic scenes, it is easy to miss and misdetect extremely small objects. On this basis, we add a detection head for detecting extremely small objects, which shows a good effect in complex traffic scene object detection. Then, we use a content-aware reassembly of features (CARAFE) module [28], to replace the original upsampling layer. We replace the original convolution module with a new SPD-Conv CNN Module [29], dedicated to low-resolution images and extremely small object detection. Finally, To find the attention region in images with large coverage, we adopt the Normalization-based Attention Module (NAM) [30] to suppress unimportant channels or pixels to improve detection efficiency. Compared to YOLOv5s, our improved multi-scale YOLOv5s can better deal with traffic scene images.

Our contributions are listed as follows:

- We add the fourth detection head for the detection of extremely small objects on the basis of the three detection heads of the original YOLOv5, which improved the problem of wrong detection and missing detection of extremely small objects in complex traffic images.
- A new content-aware reassembly of features (CARAFE) module is used for feature fusion, which enhances the feature fusion capability of the neck part. It is lighter than the traditional upsampling module and requires fewer parameters and less computation.

- A new SPD-Conv CNN Module is used to replace the original convolution module, which improves detection accuracy for low-resolution images and extremely small objects. It uses the space-to-depth and non-strided convolution layers to replace the original pooling and strided convolution layers.
- An effective attention mechanism, Normalization-based Attention Module (NAM), is added to the neck part, which improves the accuracy and robustness of the model. It applies a weight sparsity penalty to the attention modules, making them more computationally efficient while retaining similar performance.

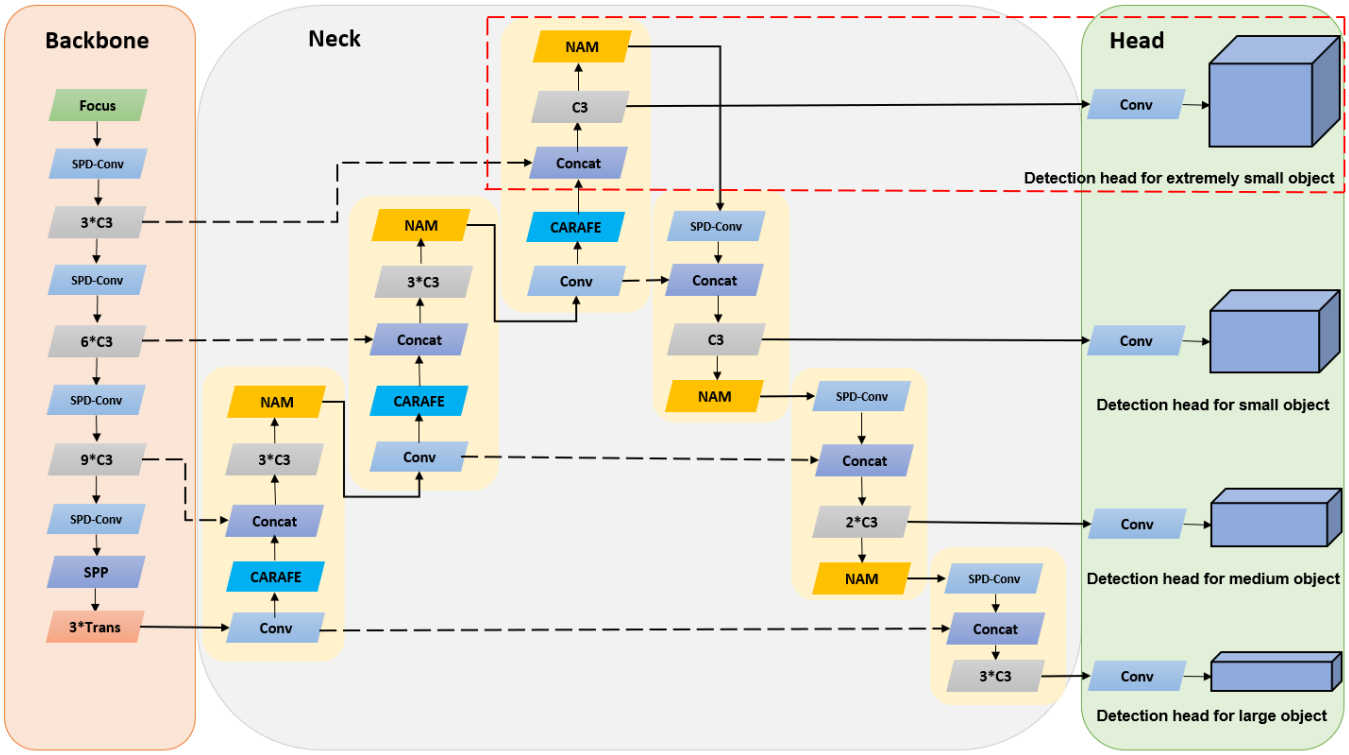

Figure 2. Overview of multi-scale YOLOv5s. It is better suited than the original YOLOv5s for the detection of small objects in complex traffic environments.

2. Related Work

Object detectors usually consist of two parts. One part is the Backbone for feature extraction, which is a convolutional neural network structure that aggregates and forms image features on different fine-grained images. The other part is the detection head used to output prediction results, to predict the image features, generate boundary boxes and predict categories. To enhance the feature extraction effect, some layers are usually added between the backbone and the head, which are called the neck of the detector. We will separately introduce these three structures in detail.

Backbone. The backbone that is often used includes VGG [31], ResNet [32], DenseNet [33], CSPDarknet53 [25], etc. These networks have proven to have strong feature extraction capabilities for problems such as detection and classification and are widely used in the construction of various network models.

Neck. To better enhance feature extraction from the backone, the neck is added between the backbone and the head for feature fusion. The neck is an important link in the detection network. Usually, the neck consists of multiple bottom-up paths and multiple top-down paths. Commonly used path-aggregation blocks in the neck are the following: FPN [26], PANet [27], BiFPN [34], ASFF [35], etc. These modules typically perform feature fusion through operations such as upsampling, downsampling, splicing, dot product, etc.

Head. The head can apply the features extracted by the backbone for target localization and classification. Heads are generally divided into two kinds: one-stage object detector

and two-stage object detector. The YOLO series is a typical one-stage detector, which can predict both the bounding box and the class of the target at the same time, giving a significant speed advantage, but with relatively low detection accuracy.

YOLOv5 generally uses the CSPDarknet53 architecture with SPP layer as backbone, FPN+PANet as neck and YOLO detection head, respectively. YOLOv5 is available in five different models, YOLOv5n, YOLOv5s, YOLOv5m, YOLOv5l and YOLOv5x. Experiments show that the training results of YOLOv5x are much better than those of YOLOv5n, YOLOv5s, YOLOv5m and YOLOv5l. Although the training computational cost of the YOLOv5x model is higher than the other four models, we still choose to use YOLOv5x in order to pursue the best detection performance.

The attention mechanism in deep learning is similar to the human visual attention mechanism, both of which extract more detailed information about the current target information from a large amount of information, which has become a hot topic of academic research in recent years. Squeeze-and-Excitation Networks (SENet) [36] integrates spatial information into the feature response in terms of channels and uses two multilayer perceptron (MLP) layers to compute the corresponding attention. Coordinate Attention (CA) [37] embeds position information into channel attention, capturing long-range dependencies in one spatial direction while retaining accurate position information in the other. Convolutional Block Attention Module (CBAM) [38] provides a solution that embeds the channel and spatial attention submodules sequentially. However, these efforts ignore information from the adjusted weights in training. Therefore, we aim to highlight salient features by using variance measures of the trained model weights.

3. Methods

3.1. Dataset Construction

To improve the generalization ability of the model, it is necessary to make scene-rich and effective datasets. The videos in this paper were remotely collected in the ring highways of Xi'an in November 2021. Opencv was used to read the video and save the video frame by frame and the frame rate was selected as 5 frames per second. The videos of highways in different time periods under various scenes are collected and the sample data are rich, which lays a good foundation for the establishment of diverse datasets. The datasets we built are shown in Figure 3. The datasets contains a wide variety of traffic objects, such as cars, trucks, buses, motorcycles and pedestrians. Unlike most publicly available datasets, ours also includes extremely small objects such as roadblocks, road debris and traffic signs. We also add some low-pixel, blurry images to the datasets to test the accuracy and robustness of our model in complex traffic scenes. We use a ratio of 8:1:1 to classify our training sets, verification sets and testing sets, there are about 8000 images of training sets with tens of thousands of detection objects.

3.2. Data Augmentation

To enhance the robustness of the model in different scenes, we introduce data augmentation techniques. Data augmentation techniques can expand and enrich the datasets at a relatively small cost. Several researchers have proposed unique data enhancement methods that use multiple images together, such as the MixUp [39] and Mosaic [15] methods. The MixUp method randomly selects two samples from the training images for random weighted summation and the labels of the samples correspond to the weighted summation. The Mosaic method stitches together four images, greatly enriching the background of the object being detected, batch normalization is used to calculate the activation statistics for each layer of four different images. In multi-scale YOLOv5s, we combine the MixUp and Mosaic data enhancement methods to expand our datasets.

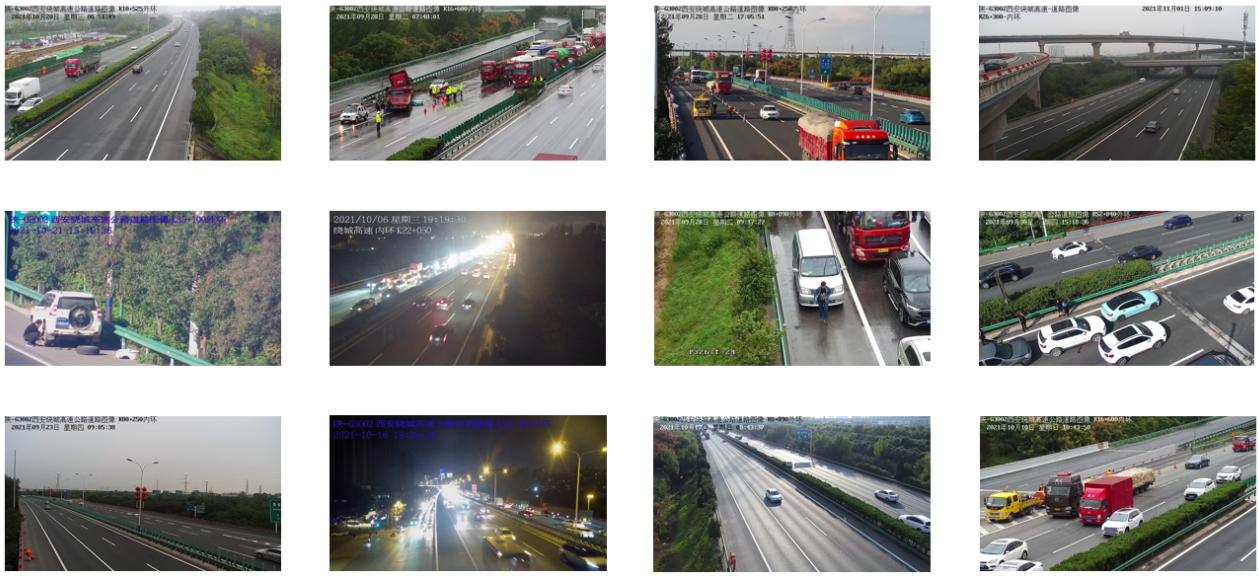

Figure 3. The datasets containing various traffic scenes with a total of approximately 100,000 images. (The Chinese words in the image are the time and place recorded by the surveillance video.)

3.3. Algorithm Optimization

3.3.1. Additional Detection Head

The datasets we built contain some extremely small traffic objects. To improve the detection accuracy of these extremely small objects, we add a prediction head for extremely small object detection. Compared with the original YOLOv5's three-head structure, our four-head structure mitigates the negative effects of drastic object scale changes. With the addition of the detection head, the performance of small object detection becomes larger, although the computation and memory consumption increase.

3.3.2. Content-Aware Reassembly of Feature Module

Feature upsampling operation is an important part of the CNN structure, which is usually used in the feature fusion part for feature enhancement. The upsampling operation enlarges the extracted feature map, so as to display the image with higher resolution. Almost all the upsampling methods use the interpolation method, which inserts new elements between pixels based on the original image pixels by using an appropriate interpolation algorithm, such as the nearest neighbor interpolation, bilinear interpolation and trilinear interpolation. These interpolation methods only consider the sub-pixel neighborhood and the image gray values have obvious discontinuities after resampling and the image quality loss is large, which means it easily causes the loss of semantic information in dense prediction tasks.

We introduce a content-aware reassembly of features (CARAFE) module in FPN structure to replace the original upsampling module. CARAFE consists of two key components, the kernel prediction module and the content-aware reassembly module, as shown in Figure 4. The kernel prediction module contains three sub-modules, namely, channel compressor, content encoder and kernel normalizer. The channel compressor reduces the number of input feature mapping channels, which improves CARAFE's efficiency and significantly reduces the number of parameters and calculations required in subsequent steps. The content encoder generates reassembly kernels based on the content of input features. In order to increase the receptive field of the encoder, a kernel-sized convolution layer is added in the process to make better use of context information within the region. The kernel normalizer normalizes the reassembly kernel spatially with a softmax function, in which the sum of kernel values is enforced to 1.

Compared with the traditional upsampling operator, CARAFE has a large field of view, which can effectively aggregate context information. It can dynamically generate an adaptive kernel, and can be aware of the instance content processing. It also occupies less computing overhead, is more lightweight, can be easily integrated into modern network architecture and has achieved good results in object detection and semantic segmentation tasks.

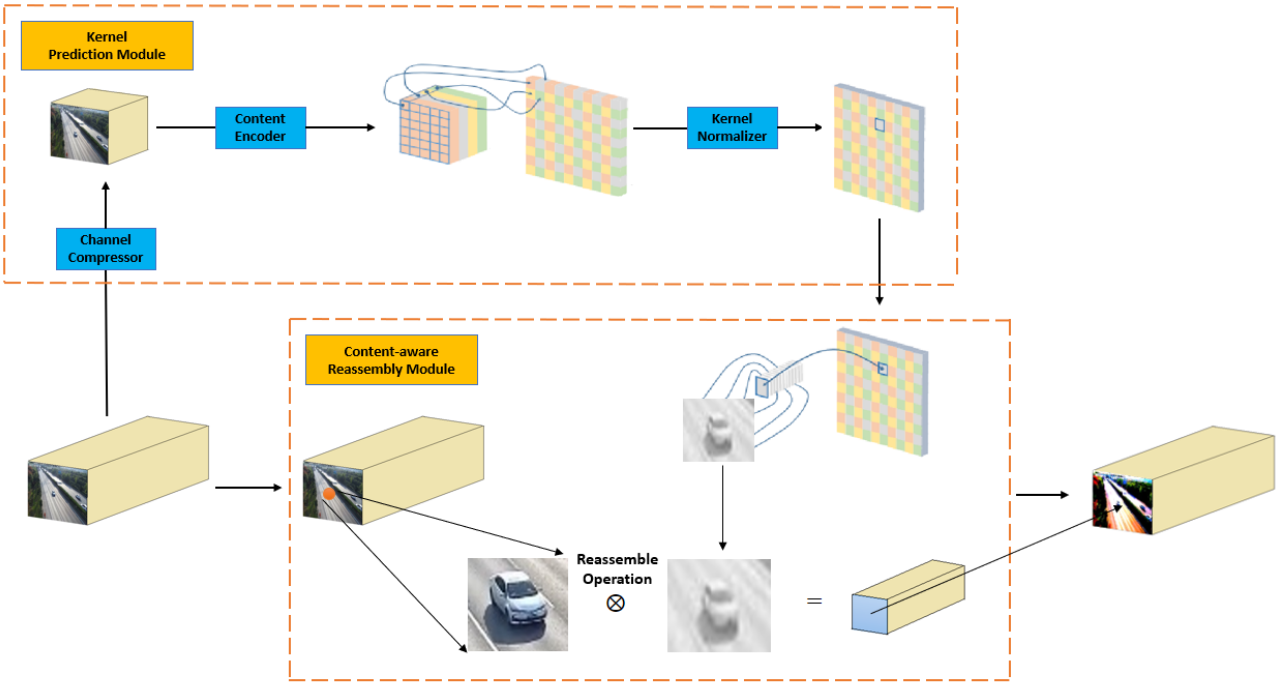

Figure 4. The overall framework of CARAFE.

3.3.3. SPD-Conv CNN Module

Traditional convolutional neural networks will lose a lot of feature information when detecting extremely small objects or objects in low-resolution images, resulting in a sharp decline in their performance. The CNN architecture as originally designed had major drawbacks, which did not manifest themselves because most of the scene images studied early on had good resolution and moderately sized objects for detection. Therefore, there is a large amount of redundant pixel information that strided convolution and pooling can conveniently skip and the model still learns features well. However, in more difficult tasks, when the images are blurred or the objects are extremely small, the currently designed CNN architectures start to lose fine-grained information and features with poor learning capabilities. To this end, we introduced a new convolutional neural network structure, SPD-Conv CNN Module, in the backbone part and the downsampling part of the neck.

The SPD-Conv CNN Module uses a space-to-depth (SPD) layer and a non-strided convolution layer to replace the pooling and strided convolution layers in the traditional CNN module. The SPD layer downsamples the feature map (X) within the entire network, while retaining all the information in the channel dimension without information loss. A non-strided convolution layer is added after each SPD layer, which uses learnable parameters in the increased convolutional layer to reduce the number of channels and reduce the non-discriminatory loss of information. The SPD-Conv CNN Module performs well when targeting low-resolution images and extremely small object detection tasks, greatly reducing information loss. The SPD-Conv CNN Module is shown in Figure 5.

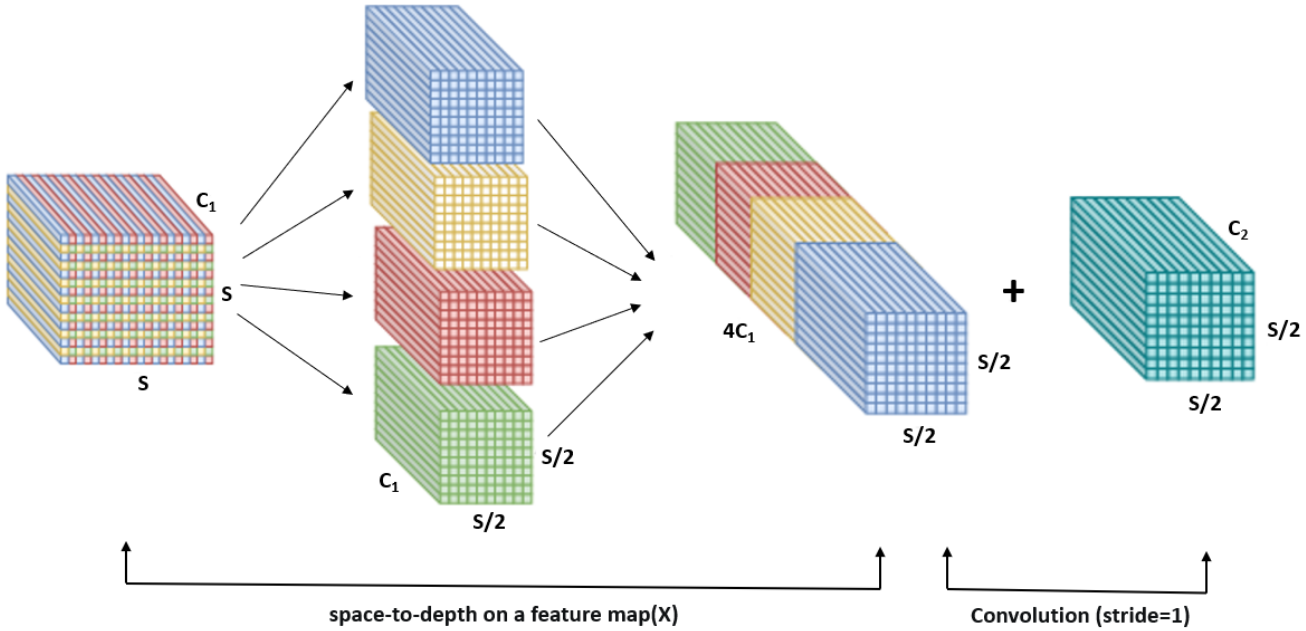

Figure 5. The overall framework of SPD-Conv.

3.3.4. Normalization-Based Attention Module

Traditional attention mechanisms generally obtain significant features from channels and spatial dimensions by means of attention operators and suppress less significant features. These methods successfully discover the mutual information between the different dimensions of the feature. However, the weight contribution factor can further suppress non-significant features and most attention mechanism modules ignore this contribution factor.

We introduce an efficient and lightweight attention mechanism, normalization-based attention module (NAM), into the neck structure to highlight salient features by exploiting variance measures of the training model weights. NAM adopts the modular integration approach of CBAM and redesigns the channel attention submodule and the spatial attention submodule, as shown in Figure 6. In order to avoid adding fully connected and convolution layers like SE and CBAM modules and increase the computing cost of the network model, NAM uses a batch normalization scaling factor to indicate the importance of weights and uses the contributing factors of weights to improve the effect of the attention mechanism. This enables the NAM module to greatly improve the efficiency of the network model detection while remaining light in terms of weight.

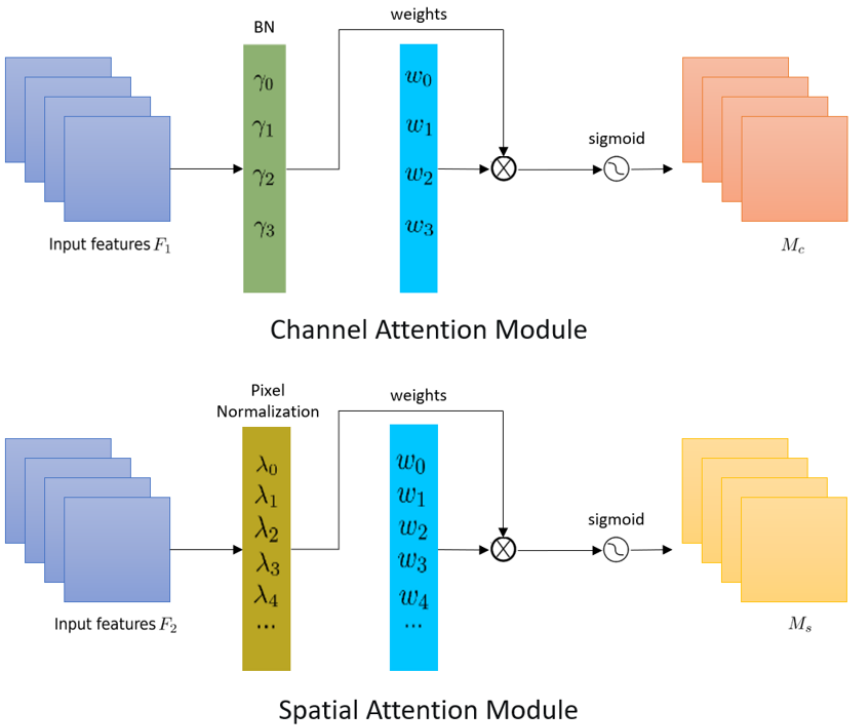

Figure 6. The channel attention submodule and the spatial attention submodule in NAM.

4. Experiments

4.1. Implementation Details

We implement multi-scale YOLOv5s on Pytorch 1.8.1, CUDA 11.3. All of our models use an NVIDIA RTX3080ti GPU for training and testing. In the training phase, we used YOLOv5s as our baseline model and used part of the pre-trained model from YOLOv5s. Since multi-scale YOLOv5s and YOLOv5s share most parts of the backbone and some parts of the head, it is possible to transfer many of the weights from YOLOv5s to multi-scale YOLOv5 and by using these weights a significant amount of training time can be saved.

4.2. Model Algorithm Evaluation Index

In this experiment, parameter quantity, Floating Point Operations (FLOPs), Precision (*P*), Recall (*R*) and mean Average Precision (*mAP*) were used to evaluate the performance of the algorithm, where Precision (*P*), Recall (*R*) and mean Average Precision (*mAP*) are expressed as:

$$P = \frac{TP}{TP + FP} \tag{1}$$

$$R = \frac{TP}{TP + FN} \tag{2}$$

$$mAP = \frac{1}{n}\sum_{n}^{i-1} AP_i \tag{3}$$

where True Positives (*TP*) represents the number of correctly detected objects, False Positives (*FP*) represents the number of incorrectly detected objects, False Negatives (*FN*) represents the number of undetected objects, *n* represents the number of categories that need to be classified. Average Precision (*AP*) represents the average accuracy of a object class.

4.3. Comparison of Multi-Scale YOLOv5s Models' Performances for Each Category

To explore the effect of introducing different improved structures on the detection performance of the model, we refer to the YOLOv5s model with the addition of a fourth detection head as YOLOv5s-F, the YOLOv5s model with the introduction of the CARAFE structure as YOLOv5s-C, the YOLOv5s model with the introduction of the SPD-Conv structure as YOLOv5s-S and the YOLOv5s model with the introduction of the NAM structure as YOLOv5s-N and experimentally explore the mAP of the model on different detection categories of the datasets. The experimental results are shown in Table 1 and the dataset detection categories are shown in Figure 7.

Table 1. Comparison of multi-scale YOLOv5s models' performances for each category, results in mAP@0.5.

Methods	All	Car	Truck	Bus	Person	Fire	Smoke	Cone	Div	Suit	Box	Moto
YOLOv5s	78.3	95.1	93.4	64.9	81.3	98.1	99.5	76.2	62.0	57.9	61.3	72.1
YOLOv5s-F	81.9	94.9	94.1	67.9	85.7	97.9	99.5	81.7	72.3	70.5	65.7	70.8
YOLOv5s-C	79.8	96.3	94.0	68.8	80.2	98.0	99.2	75.1	67.7	60.0	60.3	78.6
YOLOv5s-S	78.6	95.9	93.8	65.1	83.6	98.1	99.5	70.6	68.3	57.1	60.8	72.5
YOLOv5s-N	81.3	98.2	95.6	70.1	80.7	98.7	99.6	78.6	69.1	61.9	61.7	80.7

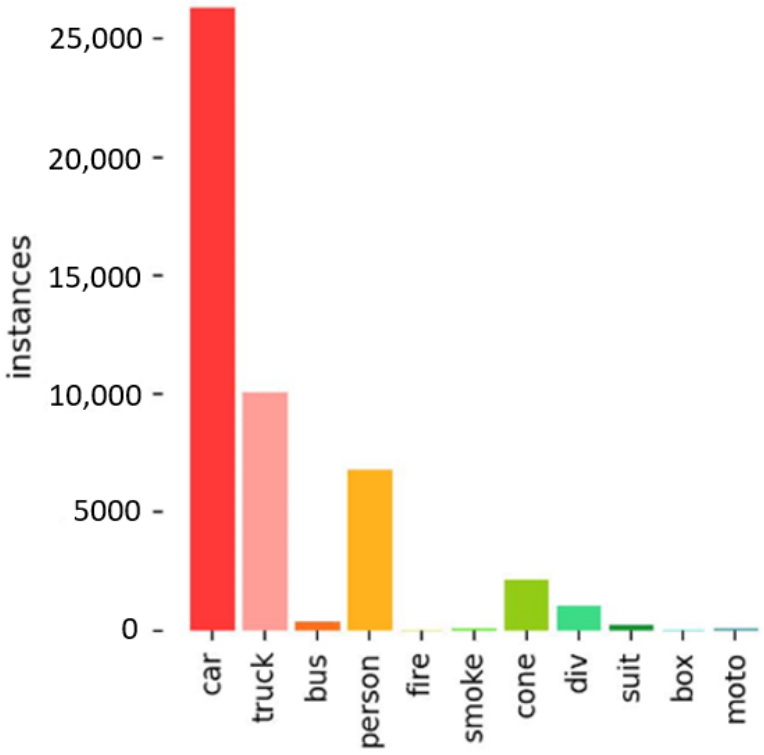

Figure 7. The number of labels of each category.

As can be seen in Table 1, all four improved methods have improved in terms of the detection accuracy of the models. Among them, the improvement of adding detection heads is more obvious in terms of the detection accuracy of extremely small objects, such as pedestrians, roadblocks and signboards. The addition of the NAM module is effective in all detection categories and the improvement in the overall detection accuracy of the model is more obvious.

4.4. Comparison of Multi-Scale YOLOv5s Models' Performances with Different Attention Mechanisms

The above experimental results show that the four improvements have significantly improved the detection accuracy of the algorithm model. In order to make a more comprehensive comparison with other attention mechanism methods, on the basis of the YOLOv5s-FCS model, the NAM and the three commonly used attention mechanisms, SE [36], CA [37] and CBAM [38] are, respectively, embedded into the neck of the algorithm and the other parts are not changed. Experimental comparison is carried out on the established datasets. The experimental results are shown in Table 2 and Figure 8.

Table 2. Comparison of multi-scale YOLOv5 models' performances with different attention mechanisms.

Methods	Params (M)	FLOPs@640 (B)	mAP@0.5 (%)	Precision (%)	Recall (%)
YOLOv5s-FCS	12.5	25.5	83.1	96.5	82.0
YOLOv5s-FCS-NAM	16.1	32.1	85.4	97.2	87.0
YOLOv5s-FCS-SE	15.8	30.7	83.5	96.2	85.3
YOLOv5s-FCS-CA	15.5	30.1	84.9	96.9	85.0
YOLOv5s-FCS-CBAM	16.5	35.7	85.1	97.1	86.9

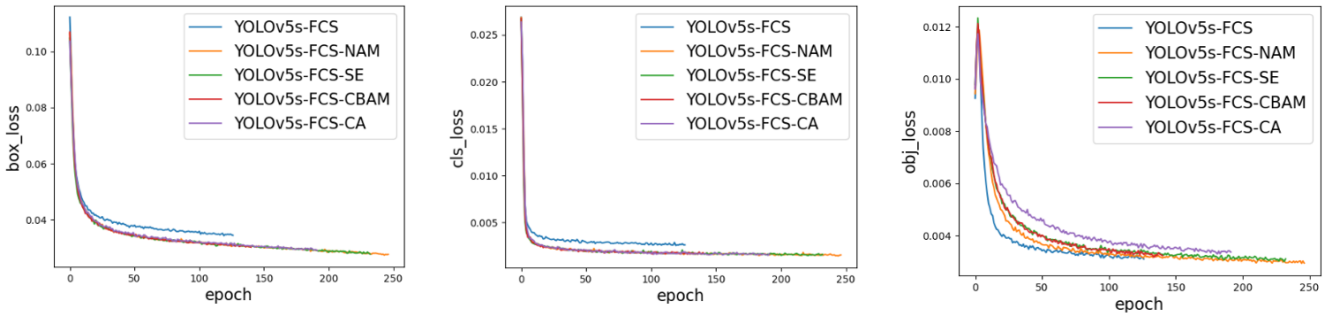

Figure 8. Loss functions for models with different attention mechanisms.

As seen in Table 2 and Figure 8, the addition of the four attention mechanism modules improved the model detection accuracy, with the NAM module showing the most significant improvement, with a 2.3% improvement in detection accuracy. The SE and CA modules were more lightweight compared to the other two modules, with less increase in the number of model parameters and computation, but less improvement in detection accuracy, with 0.4% and 1.8% improvement, respectively. The CBAM module improves the detection accuracy by 2.0%, but increases the number of model parameters and the amount of computation by more. In summary, the NAM module has certain advantages over the current mainstream attention mechanism modules.

In order to test the improvement of the generalization ability of the model by introducing the attention mechanism module, we also conduct comparative experiments on the related public datasets MS COCO and VOC 2007.

As seen in Tables 3 and 4, NAM maintains the highest detection accuracy compared with several other common attention mechanism modules on both public datasets. The experiment shows that NAM improves the detection accuracy and generalization ability of the network model.

Table 3. Comparative experiments of different attention mechanisms on MS COCO datasets.

Methods	Params (M)	FLOPs@640 (B)	mAP@0.5 (%)	Precision (%)	Recall (%)
YOLOv5s-FCS	12.5	25.5	65.4	77.6	67.9
YOLOv5s-FCS-NAM	16.1	32.1	69.7	80.9	69.3
YOLOv5s-FCS-SE	15.8	30.7	66.3	79.2	70.6
YOLOv5s-FCS-CA	15.5	30.1	67.9	80.1	65.4
YOLOv5s-FCS-CBAM	16.5	35.7	68.1	78.6	67.8

Table 4. Comparative experiments of different attention mechanisms on VOC 2007 datasets.

Methods	Params (M)	FLOPs@640 (B)	mAP@0.5 (%)	Precision (%)	Recall (%)
YOLOv5s-FCS	12.5	25.5	79.6	87.6	78.9
YOLOv5s-FCS-NAM	16.1	32.1	82.4	90.7	76.3
YOLOv5s-FCS-SE	15.8	30.7	80.6	88.2	80.9
YOLOv5s-FCS-CA	15.5	30.1	81.7	91.6	81.2
YOLOv5s-FCS-CBAM	16.5	35.7	81.4	90.6	75.5

4.5. Ablation Experiments

In order to verify the effectiveness of the four different improvement methods, this paper designed ablation experiments from the following two directions: (1) based on the original YOLOv5s algorithm, only one improvement method was added to verify the improvement effect of each improvement method on the original algorithm; (2) based on the final YOLOv5s-FCSN algorithm, only one improvement method was eliminated to verify the impact of each improvement method on the final algorithm.

As can be seen from Table 5, compared with the original YOLOv5s algorithm, the introduction of the fourth detection head has the most obvious improvement in detection accuracy, which is increased by 3.6%. Compared with the final YOLOv5s-FCSN algorithm, the elimination of NAM has the greatest impact on the detection accuracy, which is reduced by 2.3%. At the same time, compared with the original YOLOv5s, the detection accuracy of the proposed YOLOv5s-FCSN algorithm on the applied datasets is increased by 7.1%, which can cause the algorithm to have high detection accuracy while maintaining good real-time performance. The confusion matrix of YOLOv5s-FCSN is shown in the Figure 9

Table 5. Ablation experiments after the introduction of different improved methods. "+" represents the introduction of this method.

Methods	F	C	S	N	mAP@0.5 (%)	Precision (%)	Recall (%)
YOLOv5s					78.3	96.0	81.0
YOLOv5s-F	+				81.9	96.5	82.0
YOLOv5s-C		+			79.8	96.0	81.0
YOLOv5s-S			+		78.6	96.2	81.2
YOLOv5s-N				+	81.3	96.9	82.1
YOLOv5s-CSN		+	+	+	84.1	96.9	82.3
YOLOv5s-FSN	+		+	+	84.2	97.0	82.2
YOLOv5s-FCN	+	+		+	83.9	96.9	82.0
YOLOv5s-FCS	+	+	+		83.1	96.5	82.0
YOLOv5s-FCSN	+	+	+	+	85.4	97.2	87.0

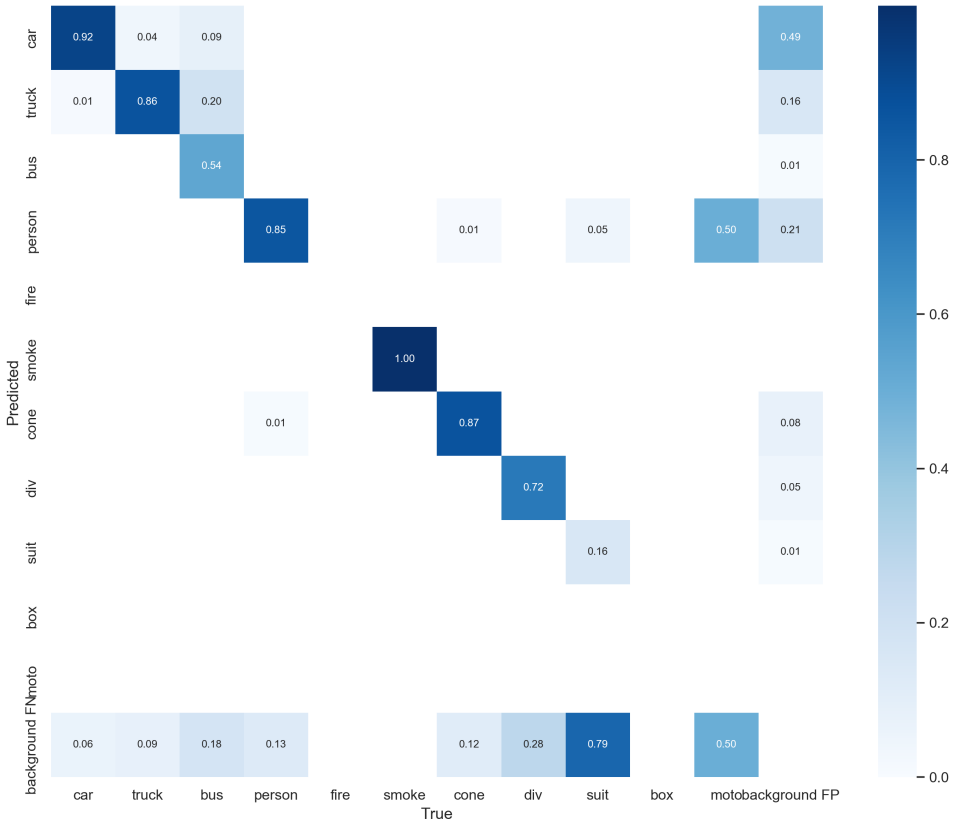

Figure 9. Confusion Matrix.

4.6. Methods' Comparative Experiment

In order to further confirm the effectiveness and superiority of the proposed algorithm, the proposed algorithm model is compared with the current mainstream algorithm model in the same scene and the performance of the algorithm is compared. The algorithm in this paper is compared with YOLOv5n, YOLOv5s, YOLOv5m, YOLOv5l, YOLOv5x, Faster R-CNN, SSD, YOLOv4-tiny and YOLOv4 and the results are shown in Table 6.

Table 6. Comparative experiment with the current mainstream methods.

Methods	Params (M)	FLOPs@640 (B)	mAP@0.5 (%)	Precision (%)	Recall (%)
YOLOv4	62.1	128.4	74.0	89.6	80.1
YOLOv4-tiny	6.1	3.4	75.9	90.7	80.5
SSD	50.4	114.2	70.1	85.4	73.2
Faster R-CNN	67.9	147.2	73.6	88.9	77.8
YOLOv5x	86.7	205.7	85.3	97.1	85.0
YOLOv5l	46.5	109.1	82.9	96.2	81.1
YOLOv5m	21.2	49.0	80.6	96.2	82.4
YOLOv5s	7.2	16.5	78.3	96.0	81.0
YOLOv5n	1.9	4.5	69.7	87.6	78.9
Multi-scale YOLOv5s	16.1	32.1	85.4	97.2	87.0

Compared with the experimental results in Table 6, it can be seen that the algorithm proposed in this paper has the highest detection accuracy compared with other mainstream detection models while taking up a small number of parameters and computations, keeping the model lightweight. The improved YOLOv5s model improves the detection accuracy by 7.1% compared to the original YOLOv5s model, which is comparable to YOLOv5x and the number of parameters and amount of computation are much smaller than the YOLOv5x model. The improved multi-scale YOLOv5s algorithm also has a substantial improvement

in detection accuracy compared to the mainstream algorithms YOLOv4, YOLOv4-tiny, SSD and Faster R-CNN. To sum up, the multi-scale YOLOv5s algorithm proposed in this paper has the highest detection accuracy while maintaining good real-time performance, which proves the feasibility and superiority of the algorithm in this paper. The detection results of the multi-scale YOLOv5s algorithm on the datasets are shown in Figure 10.

Figure 10. Detection results of the multi-scale YOLOv5s algorithm. (The Chinese words in the image are the time and place recorded by the surveillance video.)

5. Conclusions

In order to improve the detection accuracy of traffic objects in complex road scenes, we add a detection head for extremely small objects to the original YOLOv5s model, which significantly improves the detection accuracy of extremely small traffic objects. A content-aware reassembly of features (CARAFE) module is introduced in the feature fusion part to enhance the feature fusion. A new SPD-Conv CNN Module is introduced instead of the original convolutional structure to enhance the overall computational efficiency of the model. Finally, the normalization-based attention module (NAM) is introduced, allowing the model to focus on more useful information during training and significantly improving detection accuracy.

The experimental results show that compared with the original YOLOv5s algorithm, the detection accuracy of the multi-scale YOLOv5s model proposed in this paper is improved by 7.1% on the constructed diverse traffic scene datasets, which is comparable to YOLOv5x, maintaining the lightness of the model with respect to its weight while having a high detection accuracy. Compared with the current mainstream object detection algorithms, the multi-scale YOLOv5s model has the highest detection accuracy and is superior to the current mainstream object detection algorithms in the detection of traffic objects in complex road scenes.

Author Contributions: Methodology, A.L.; Validation, A.L.; Writing—original draft, A.L.; Writing–review & editing, S.S., Z.Z., M.F., C.W. and W.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China grant number 62072053, 62006026 and the Central Universities Basic Research Special Funds grant number 300102241304, 300102241202. The APC was funded by the National Natural Science Foundation of China grant number 62072053, 62006026 and the Central Universities Basic Research Special Funds grant number 300102241304, 300102241202.

Data Availability Statement: Not applicable.

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

References

1. Matthews, N.; An, P.; Charnley, D.; Harris, C. Vehicle Detection and Recognition in Greyscale Imagery. *IFAC Proc. Vol.* **1995**, *4*, 473–479.
2. You, M.; Zhang, Y.; Shen, C.; Zhang, X. An Extended Filtered Channel Framework for Pedestrian Detection. *IEEE Trans. Intell. Transp. Syst.* **2018**, *19*, 1640–1651. [CrossRef]
3. Walk, S.; Majer, N.; Schindler, K.; Schiele, B. New features and insights for pedestrian detection. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010.
4. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*. [CrossRef] [PubMed]
5. Joseph, R.; Santosh, D.; Ross, G.; Ali, F. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
6. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In Proceedings of the Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016.
7. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
8. Zhang, H.; Wang, Y.; Dayoub, F.; Sünderhauf, N. VarifocalNet: An IoU-aware Dense Object Detector. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020.
9. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009.
10. 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.
11. Everingham, M.; Gool, L.V.; Williams, C.; Winn, J.; Zisserman, A. The Pascal Visual Object Classes (VOC) Challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–308. [CrossRef]
12. Geiger, A.; Lenz, P.; Stiller, C.; Urtasun, R. Vision meets robotics: The KITTI dataset. *Int. J. Robot. Res.* **2013**, *32*, 1231–1237.
13. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
14. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
15. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
16. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. Scaled-YOLOv4: Scaling Cross Stage Partial Network. In Proceedings of the IEEE/cvf Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020.
17. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. YOLOX: Exceeding YOLO Series in 2021. *arXiv* **2021**, arXiv:2107.08430.
18. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: Fully Convolutional One-Stage Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019.
19. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable DETR: Deformable Transformers for End-to-End Object Detection. *arXiv* **2020**, arXiv:2010.04159.
20. Zhou, X.; Koltun, V.; Krähenbühl, P. Probabilistic two-stage detection. *arXiv* **2021**, arXiv:2103.07461.
21. Yu, F.; Zhong, M.; Tang, S.; Zheng, Z. Improved traffic signal light recognition algorithm based on YOLO v3. In Proceedings of the International Conference on Optics and Machine Vision (ICOMV 2022), Guangzhou, China, 14–16 January 2022.
22. Zhu, J.; Li, X.; Jin, P.; Xu, Q.; Sun, Z.; Song, X. MME-YOLO: Multi-Sensor Multi-Level Enhanced YOLO for Robust Vehicle Detection in Traffic Surveillance. *Sensors* **2020**, *21*, 27. [CrossRef] [PubMed]
23. Li, Y.; Li, J.; Meng, P. Attention-YOLOV4: A real-time and high-accurate traffic sign detection algorithm. *Multimed. Tools Appl.* **2022**, *82*, 7567–7582. [CrossRef]
24. Mittal, U.; Chawla, P.; Tiwari, R. EnsembleNet: A hybrid approach for vehicle detection and estimation of traffic density based on faster R-CNN and YOLO models. *Neural Comput. Appl.* **2022**, *35*, 4755–4774.
25. Wang, C.Y.; Liao, H.Y.M.; Yeh, I.H.; Wu, Y.H.; Chen, P.Y.; Hsieh, J.W. 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, Long Beach, CA, USA, 16–17 June 2019.
26. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016.

27. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
28. Wang, J.; Chen, K.; Xu, R.; Liu, Z.; Loy, C.C.; Lin, D. CARAFE: Content-Aware ReAssembly of FEatures. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27–28 October 2019.
29. Sunkara, R.; Luo, T. No More Strided Convolutions or Pooling: A New CNN Building Block for Low-Resolution Images and Small Objects. *arXiv* **2022**, arXiv:2208.03641.
30. Liu, Y.; Shao, Z.; Teng, Y.; Hoffmann, N. NAM: Normalization-based Attention Module. *arXiv* **2022**, arXiv:2111.12419.
31. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2015**, arXiv:1409.1556.
32. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
33. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
34. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and Efficient Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
35. Liu, S.; Huang, D.; Wang, Y. Learning Spatial Fusion for Single-Shot Object Detection. *arXiv* **2019**, arXiv:1911.09516.
36. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018.
37. 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, Virtual, 19–25 June 2021.
38. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
39. Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. mixup: Beyond Empirical Risk Minimization. *arXiv* **2017**, arXiv:1710.09412.