electronics

MDPI

Article

Improved Traffic Small Object Detection via Cross-Layer Feature Fusion and Channel Attention

Qinliang Chuai, Xiaowei He * and Yi Li

School of Computer Science and Technology, Zhejiang Normal University, Jinhua 321004, China; qinliang@zjnu.edu.cn (Q.C.); leeyee@zjnu.edu.cn (Y.L.)
* Correspondence: jhhxw@zjnu.edu.cn

Abstract: Small object detection has long been one of the most formidable challenges in computer vision due to the poor visual features and high noise of surroundings behind them. However, small targets in traffic scenes encompass a multitude of complex visual interfering factors, bearing crucial information such as traffic signs, traffic lights, and pedestrians. Given the inherent difficulties faced by generic models in addressing these issues, we conduct a comprehensive investigation on small target detection in this application scenario. In this work, we present a Cross-Layer Feature Fusion and Channel Attention algorithm based on a lightweight YOLOv5s design for traffic small target detection, named CFA-YOLO. To enhance the sensitivity of the model toward vital features, we embed the channel-guided Squeeze-and-Excitation (SE) block in the deep layer of the backbone. Moreover, the most excellent innovation of our work belongs to the effective cross-layer feature fusion method, which maintains robust feature fusion and information interaction capabilities; in addition, it simplifies redundant parameters compared with the baseline model. To align with the output features of the neck network, we adjusted the detection heads from three to two. Furthermore, we also applied the decoupled detection head for classification and bounding box regression tasks, respectively. This approach not only achieves real-time detection standards, but also improves the overall training results in parameter-friendly manner. The CFA-YOLO model significantly pays a lot of attention to the detail features of small targets, thereby it also has a great advantage in addressing the issue of poor performance in traffic small target detection results. Vast experiments have validated the efficiency and effectiveness of our proposed method in traffic small object detection. Compared with the latest lightweight detectors, such as YOLOv7-Tiny and YOLOv8s, our method consistently achieves superior performance both in terms of the model's accuracy and complexity.

Keywords: computer vision; feature fusion; small object detection; intelligence transportation

check for updates

Citation: Chuai, Q.; He, X.; Li, Y. Improved Traffic Small Object Detection via Cross-Layer Feature Fusion and Channel Attention. *Electronics* **2023**, *12*, 3421. https://doi.org/10.3390/electronics12163421

Academic Editor: Silvia Liberata Ullo

Received: 12 July 2023
Revised: 7 August 2023
Accepted: 9 August 2023
Published: 12 August 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The rapid advancement of science and technology represents the culmination of human intellect, showcasing the extension of our capabilities. In recent years, intelligent automobile manufacturing has garnered significant attention. The concept of autonomous driving revolves around vehicles possessing the ability to operate on roads with minimal manual intervention, integrating various subjects such as car-making engineering, mathematics, computer vision science, and more. The development of autonomous driving is driven by the necessity to minimize traffic accidents. Autonomous vehicles must understand the surrounding environment and make judgments during operation, necessitating an object detection system to identify targets captured by vehicle cameras on the road, including pedestrians, traffic signs, traffic lights, and nearby vehicles. Object detection is a crucial component of the autonomous driving system, allowing it to perceive the external world and provide digital image information for vehicle decision-making. To accomplish this, the system often requires a large amount of surrounding vision information to accurately distinguish objects. Firstly, multiple cameras are installed to capture visual

data from different directions, providing comprehensive knowledge about the vehicle's surroundings. Subsequently, detection algorithms analyze these images, recognize targets, and facilitate subsequent control decisions based on the obtained results. The development of deep learning [1] in computer vision has significantly propelled the progress of object detection research over the years, providing valuable technical support for addressing specific challenges in traffic scenarios. However, existing detection algorithms fall short of meeting the requirements of this application scenario, highlighting the increasing urgency for visual detection studies in traffic scenarios.

Object detection serves as a fundamental aspect of computer vision, forming a solid foundation for other visual tasks, such as instance segmentation [2] and object tracking [3]. Deep learning methods have exhibited superior capabilities in complex representations compared to traditional object detection algorithms, significantly enhancing algorithm performance. In the realm of traffic scenes, numerous researchers have proposed deep-learning-based object detection algorithms. For example, Gu et al. [4] proposed a lightweight model based on YOLOv4 [5], effectively reducing computational costs and achieving a balanced trade-off between performance and deployment complexity, making it suitable for deployment on resource-constrained edge devices like mobile and embedded devices. Liu et al. [6] presented the M-YOLO algorithm, which improves detection efficiency while ensuring a certain level of generalization and robustness. Additionally, He et al. [7] focused on multi-category small object detection in traffic scenes and proposed the YOLO-MXANet model, which improves detection accuracy while reducing model complexity.

Small target detection has been a difficult problem in the field of target detection because small targets contain less information and are greatly affected by environmental noise. Although there are many research works on small target detection [8–10], these works mainly address the issues of dense small target detection and how to improve small target feature extraction capability. However, the detection of traffic small targets, such as traffic signs and traffic signals, is still a major challenge in traffic scenarios. The challenges of traffic small target detection are mainly twofold: first, the need to accurately discriminate the category information contained in the small targets; and second, how to accurately recognize the targets in the complex environmental noise. Existing detection algorithms, despite their progress in general target detection, are inadequate for addressing the specific requirements of this application scenario. Consequently, their applicability is limited, and their performance falls short of expectations.

In this study, we begin by considering the characteristics of small objects and aim to build a model architecture that enables more effective feature fusion and precise localization. Based on this objective, we propose the CFA-YOLO algorithm, which achieves highly accurate detection of small objects. Our algorithm aims to significantly improve feature expression by redesigning the network structure, reducing information redundancy, and enhancing the interaction of information flow between different levels. Thereby, these designs boost overall the model's performance to recognize small targets. Firstly, we integrate the SE attention block into the feature extraction framework to focus on detailed information about small objects. Additionally, we introduce a novel neck network architecture called the Cross-layer Alternating Pyramid Aggregation Network (CAPAN), which targets the enhanced propagation of features. This architecture ensures a comprehensive combination of semantic and spatial information necessary for recognizing small objects, thereby providing richer data for detection. Moreover, we adjust the coupling detection heads to decoupling detection heads and conduct separate regression training for the classification and localization tasks to achieve more accurate recognition. Furthermore, a series of experiments are performed to validate the effectiveness of the two decoupling heads paradigm. By optimizing the model structure, we achieve more concise and efficient results. Extensive experiments on the TT100K [11] and BDD100K [12] datasets highlight the excellent performance of our model design.

The main contributions of our work can be summarized as follows:

- We introduced an attention mechanism module into the feature extraction process to enhance the model's focus on important information and reduce feature redundancy. This ensures that the model allocates more attention to crucial features, improving its overall performance;
- We designed a novel CAPAN neck network that effectively enriches the feature information of small objects. This network facilitates the communication and fusion of spatial and semantic information across different levels, resulting in enhanced feature representation for accurate detection;
- To optimize model training, we replaced the original three-coupling detection heads with two decoupling counterparts. This decoupling style allows for independent regression of the classification and localization tasks, leading to a more useful and effective training period with fewer model parameters;
- Extensive experiments are conducted on the TT100K [11] and BDD100K [12] datasets to validate the superiority of our designed model. CFA-YOLO noticeably outperforms the current latest lightweight methods, showing its outstanding performance and advancement in small object detection in traffic scenes.

2. Related Work

General object detection algorithms based on deep learning methods can be broadly classified into two categories. The first type is two-stage detectors, which typically owe higher accuracy but with slower inference speed, while the other belongs to one-stage types, possessing high running speed but with lower accuracy of overall performance. To date, a large number of works with the help of two types of detectors have been applied to traffic object detection applications.

2.1. Two-Stage Method in Traffic Scenes

Two-stage detectors, such as R-CNN series [13–15], have the following steps: Firstly, they extract proposed regions by Selective Search [16] or a region proposal network (RPN). Then, these regions are fed forward through subsequent neural networks for classification and regression. Numerous studies have applied two-stage detectors to traffic scenarios. Qian et al. [17] designed a traffic sign detection system with a Fast R-CNN detector; it has also been successfully applied to pedestrian detection by Li et al. [18], who proposed a framework incorporating two sub-networks capable of detecting pedestrians at different spatial scales. Fan et al. [19] demonstrated the superior car detection results achieved by improved Faster R-CNN design. Zhao et al. [20] adapted Faster R-CNN for pedestrian detection by adjusting and training a pedestrian-specific network, achieving faster inference speed and an acceptable miss rate compared to Li et al. [18], based on Fast R-CNN.

2.2. One-Stage Method in Traffic Scenes

Although two-stage detectors outperform inaccurate results, the one-stage detectors startle in terms of real-time performance. This kind of design involves outstanding advantages like simplifying the model's complexity and improving inference speed, which meets the real-time detection standard. The classic one-stage detection algorithms include YOLO [21] and SSD [22]. Due to their superior network design, extensive works based on this approach are applied in traffic scenes. Shortly after the introduction of SSD, Kim et al. [23] exploited the framework for road detection and achieved pleasant results. Xie et al. [24] proposed a Multi-directional YOLO (MD-YOLO) for detecting car license plates; it also retains great potential for detecting multiple objects simultaneously. Jensen et al. [25] proposed an improved YOLOv2 for traffic light detection scenes. Yang et al. [26] employed a larger 14×14 grid instead of the standard 7×7 grid to extract larger feature maps and raise recognition accuracy.

2.3. Attention-Based Method

The attention mechanism method mimics the human visual ability to capture specific information about objects, focusing more details on obvious target features. There are many efforts to incorporate attention modules to improve detection performance. Guo et al. [27] used an attention mechanism to improve the defects of the feature pyramid networks (FPN) and enhanced the ability of feature fusion. Zuo et al. [28] applied an attention network to the Faster R-CNN framework for traffic sign detection. Wang et al. [29] introduced the BANet, which integrates multi-channel attention and multi-attention fusion modules to enhance the network's basic information.

3. Methodology

In this section, we provide an overview of the CFA-YOLO structure, which is illustrated in Figure 1. The specific design of our model is detailed in Section 3.1. Subsequently, Section 3.2 explains the construction of the backbone network. The Cross-layer Alternating Pyramid Aggregation Network (CAPAN) is introduced in Section 3.3. Furthermore, Section 3.4 presents the improved detection head structure. Lastly, the utilized loss function is described in Section 3.5.

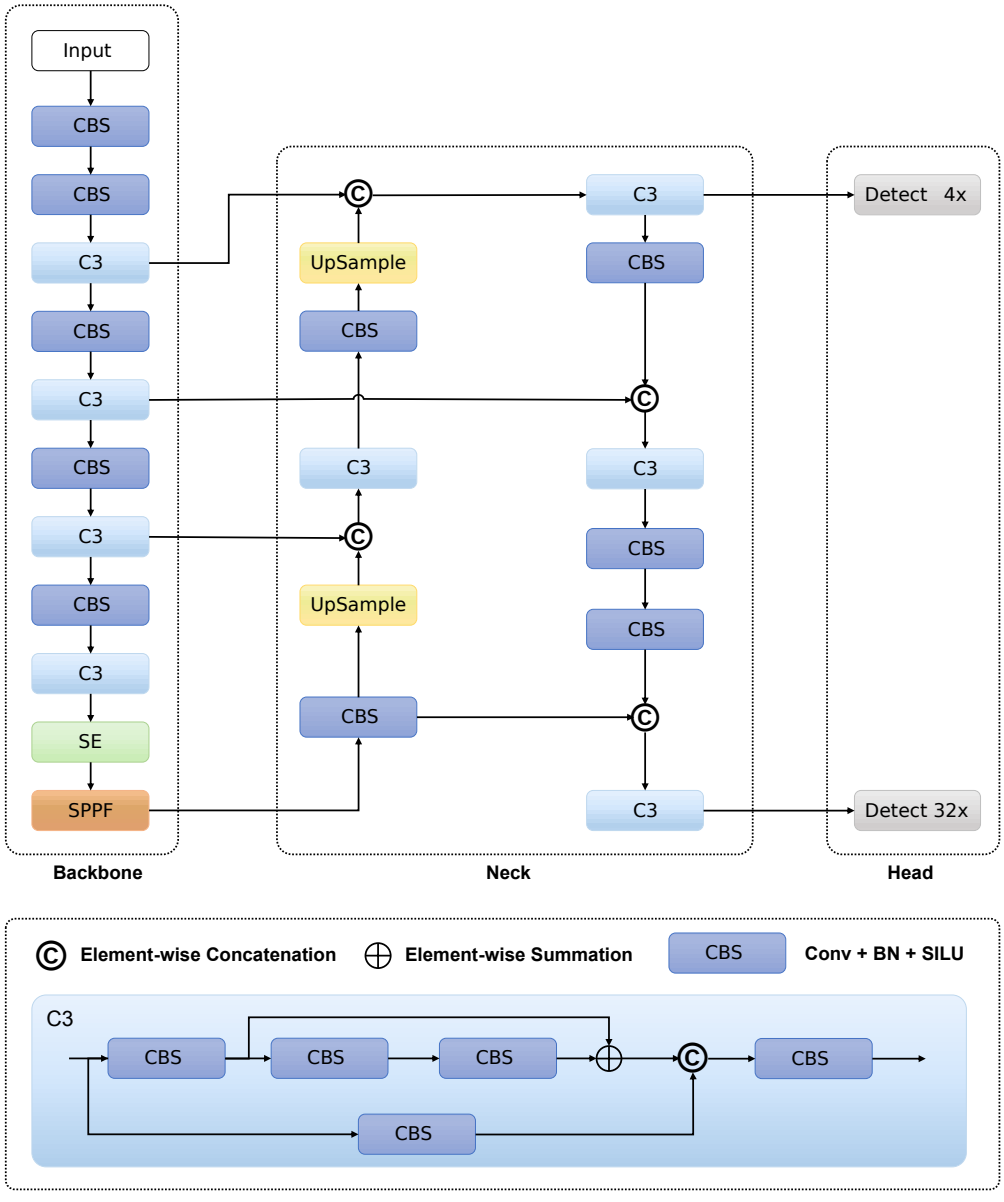


Figure 1. The structure of CFA-YOLO.

3.1. CFA-YOLO Architecture

In this study, we selected YOLOv5s as our baseline algorithm, which is an improvement over YOLOv3 [30] and YOLOv4 [5]; while YOLOv5s is more suitable for mobile devices, its accuracy in detecting small targets still has room for improvement.

To address the issue of small target detection in traffic scenes, we propose a model called CFA-YOLO, which incorporates three enhancements to the baseline model. Firstly, we integrate the original CSP-Darknet 53 backbone network with a Squeeze-and-Excitation (SE) block and replace the Spatial Pyramid Pool (SPP) with a more efficient Spatial Pyramid Pool Fast (SPPF) module. This allows the model to pay more attention to important information during feature extraction and handle images of arbitrary scales. Additionally, we introduce a novel neck network structure called the Cross-layer Alternating Pyramid Aggregation Network (CAPAN), which facilitates effective cross-layer feature interaction, ensuring information fusion while minimizing redundancy. For the prediction layers, after analyzing the output features, we use the two neck network layers that contain the most informative features for output prediction. Consequently, we employ only two prediction heads and replace the original coupled head with a decoupled design that is responsible for classification and bounding box regression tasks, respectively.

3.2. Improved Feature Extraction Network

The overview of the CFA-YOLO framework is illustrated in Figure 1. The feature extraction network typically consists of deep convolutional neural networks, where increasing depth allows the network to capture more target-related feature information for subsequent recognition. However, deeper layers of the network often exhibit a higher number of parameters and complexity, which can introduce information redundancy and bias during the model's training time. To address this issue, we introduce the Squeeze-and-Excitation (SE) block to explore more detailed features of small objects. The attention block emulates human vision by selectively focusing on more informative features, and also enables the model to prioritize channels with higher weights and effectively learn important and precise knowledge. The structure of the SE block is depicted in Figure 2.

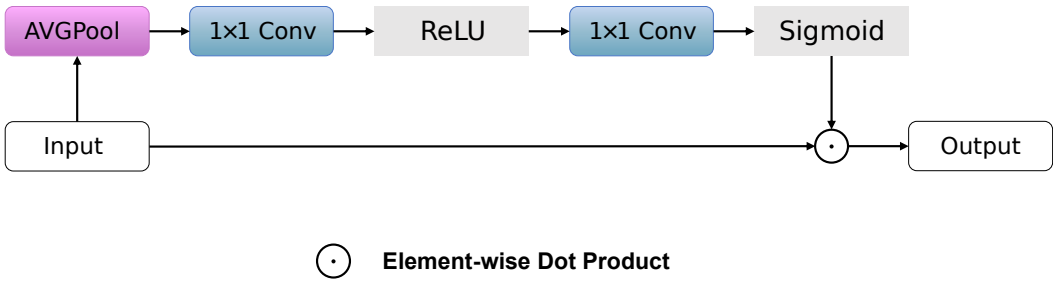


Figure 2. Squeeze-and-Excitation block.

Many deep learning neural networks require fixed-size input images, which limits their compatibility with images of arbitrary scales. To address this limitation, He et al. proposed SPPNet [31]. The SPP module serves the purpose of handling non-uniform input image sizes, ensuring consistent-sized outputs regardless of the input size. Additionally, SPP enables the fusion of features from different spatial scales, which is particularly useful when dealing with images containing objects of varying sizes, especially in complex multi-target scenarios. However, the structure of the SPP module can be further optimized. In our approach, we introduce the Spatial Pyramid Pool Fast (SPPF) module as a replacement for the SPP in our designed network. The SPPF reduces the model's parameters while preserving the rich features of objects. Figure 3a illustrates the structure of the SPP module, while Figure 3b demonstrates the structure of the SPPF module.

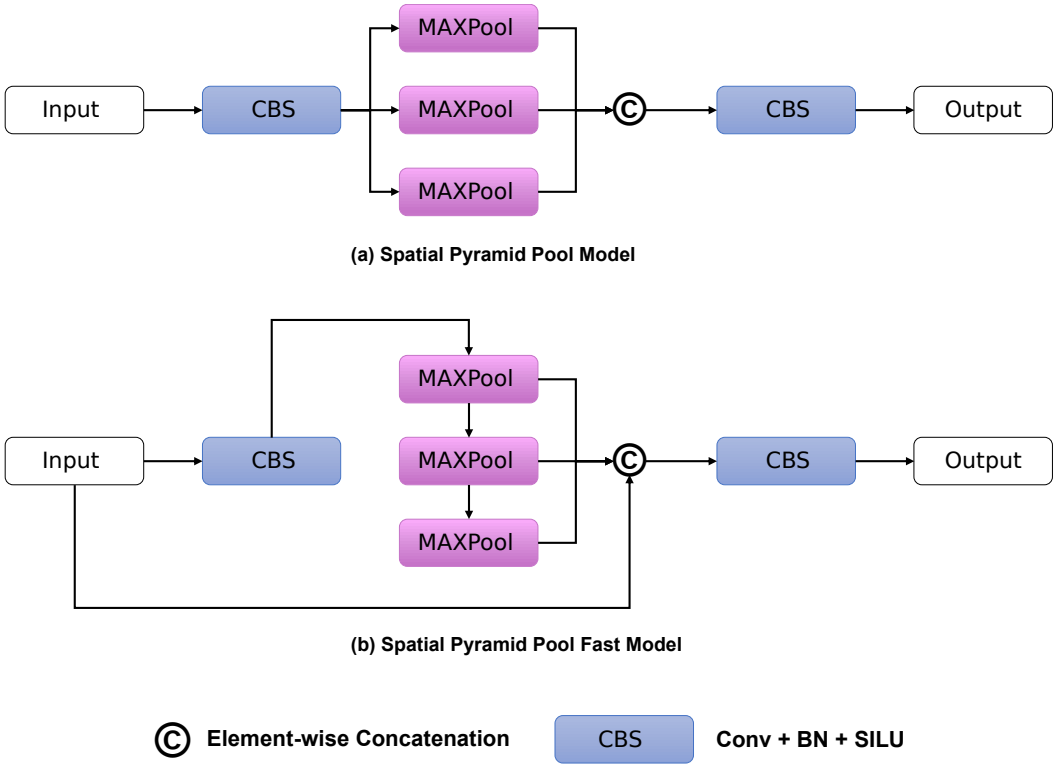


Figure 3. (**a**) The structure of the Spatial Pyramid Pool module. (**b**) The structure of the Spatial Pyramid Pool Fast module.

3.3. Cross-layer Alternating Pyramid Aggregation Network

With layers in the backbone going deeper, the relationship between local and global information gradually weakens. Shallow features possess higher resolution, capturing more detailed and positional information, but they lack semantic meaning and may contain more noise due to weak extraction abilities, while deep features contain richer semantic information but with lower resolution and may lack fine-grained details. Therefore, fusing these two types of information can significantly enhance detection performance, especially for small targets.

Traditional feature pyramid networks (FPN) [32] consist of a single top-down pathway, which limits the multi-scale interaction of information flow. To maximize the utilization of features, YOLOv5s introduced a bottom-up branch in addition to the top-down process, combined with the Path Aggregation Network (PANet) [33]. This design facilitates the transmission of deep semantic information to the earlier layers, and enables the integration of detailed information from the shallow layers into the deeper layers, enhancing feature representation ability. However, this structure suffers from redundancy in many aspects, and the fusion of information is not targeted but rather a simple superposition. To address these issues, we further improve the PANet architecture and propose the Cross-layer Alternating Pyramid Aggregation Network (CAPAN).

For the top-down branching in CAPAN, we specifically select features from the two deepest layers to fuse with the four-fold downsampled feature maps. This choice is motivated by the fact that the shallow layers contain crucial details necessary for detecting small targets and localizing objects, while the deeper layers offer rich semantic information that aids in instance discrimination. Regarding the bottom-up branch, we consider the lack of spatial information in the deeper features to be the reason for poor performance in small target detection. Thus, we make the 8-fold downsampled features with the former-fused 4-fold downsampled features fused, then they are going fused with the 32-fold downsampled features. Finally, we feed the deepest and shallowest features of the neck network into the prediction layer for result analysis. The structure of CAPAN is illustrated in Figure 4d.

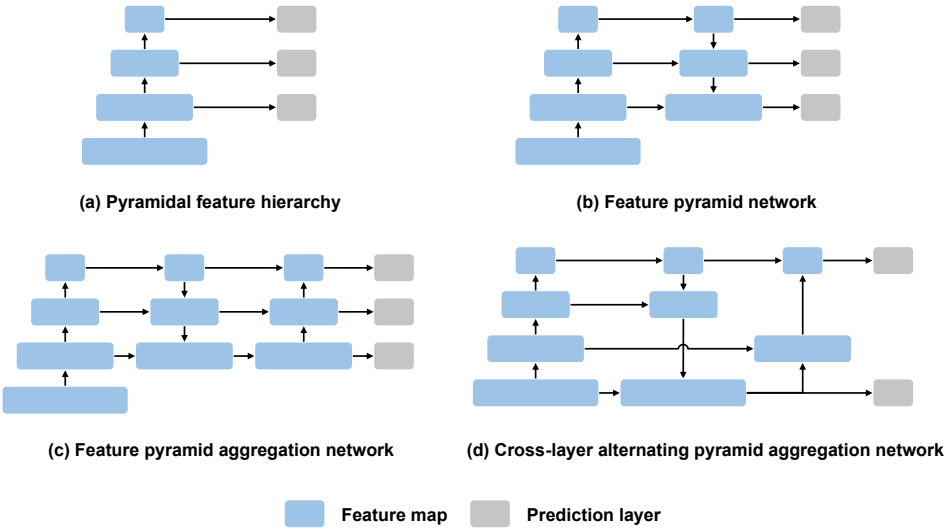


Figure 4. (**a**) Reusing different hierarchical features of the backbone network for multi-scale prediction. (**b**) Top-down feature fusion of features at different levels of the backbone. (**c**) The Path Aggregation Network (PANet) contains two feature fusion branches, top-down and bottom-up. (**d**) We design the Cross-layer Alternating Pyramid Aggregation Network (CAPAN).

Furthermore, we replace the original nearest-neighbor interpolation upsampling method with a bilinear interpolation approach, which results in a more natural transition of the feature maps. The calculation process of bilinear interpolation is illustrated by the following formula:

$$\begin{cases} f(R_1) = f(A_{11})\frac{x_2-x}{x_2-x_1} + f(A_{21})\frac{x-x_1}{x_2-x_1} & R_1 = (x, y_1) \\ f(R_2) = f(A_{12})\frac{x_2-x}{x_2-x_1} + f(A_{22})\frac{x-x_1}{x_2-x_1} & R_2 = (x, y_2) \end{cases} \tag{1}$$

$$f(P) = f(R_1)\frac{y_2-y}{y_2-y_1} + f(R_2)\frac{y-y_1}{y_2-y_1} \tag{2}$$

In the given formula, we have the coordinates of points $A_{11}(x_1, y_1)$, $A_{12}(x_1, y_2)$, $A_{21}(x_2, y_1)$, and $A_{22}(x_2, y_2)$, and we want to calculate the value of point $P(x, y)$. Bilinear interpolation can be decomposed into two linear interpolations. First, we determine the function for the abscissa direction using Equation (1). Then, we determine the function for the ordinate direction using Equation (2). Finally, we calculate the value of $P(x, y)$ using the obtained functions.

3.4. Improved Detect Head Structure

The baseline model adopts three coupled detection heads, which directly perform convolution processing on the input features to generate the outputs for object classification and regression. While this design is structurally simple, it will bring some negative results.

In practice, the tasks of target detection, namely classification and localization, suffer from spatial misalignment [34,35]. This means that each task focuses on different regions of interest. Specifically, the classification task is more concerned with determining which extracted features related the most to the object, while the localization task emphasizes correcting the coordinates of the ground truth and proposed boxes. Therefore, treating both tasks jointly in one step does not yield optimal results.

To address the aforementioned issues, we replace the coupled detection head in the model with a decoupled head style. The structure is illustrated in Figure 5b. The decoupled head employs two separate network branches to learn the classification and localization features of the target. By utilizing this approach, we not only enhance the model's generalization performance, but also improve its convergence speed during the training process. Additionally, through comparative analysis and experimental demonstration, we validate

that the top and bottom layers of the neck part contain the richest information levels. To reduce the feature redundancy of the detection head, we decrease the three detection heads to two.

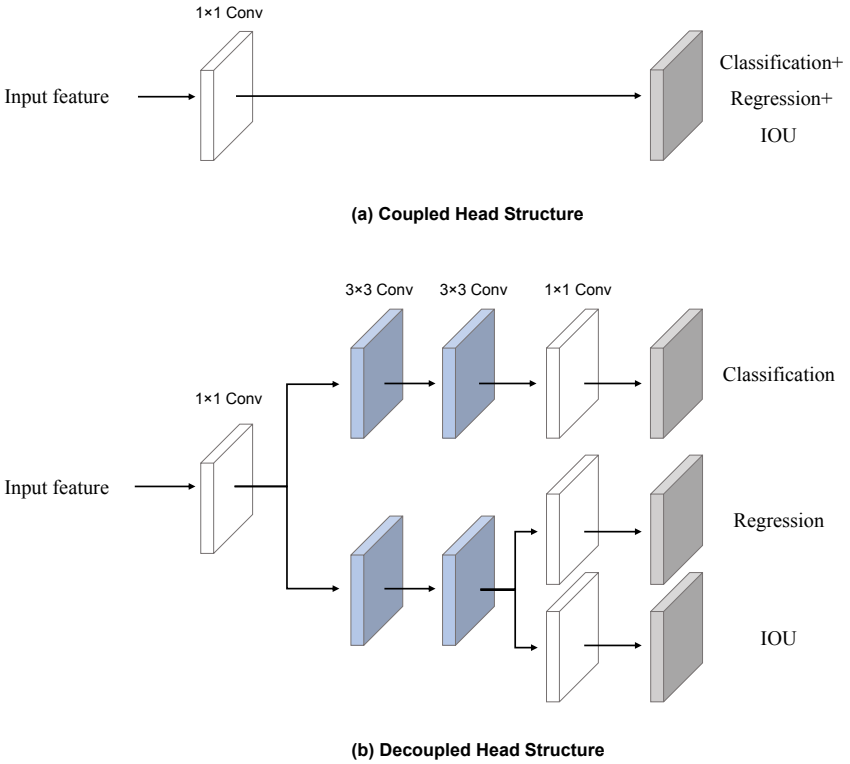


Figure 5. (**a**) The structure of coupled detection head. (**b**) The structure of decoupled detection head.

3.5. Loss Function

The loss function of CFA-YOLO includes confidence loss, classification loss, and bounding box regression loss. The total loss function of our model is presented in Equation (3).

$$Loss = \sum_{k=0}^{K} \left(\alpha_{obj} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \prod_{k_{ij}}^{obj} L_{obj} + \alpha_{cls} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \prod_{k_{ij}}^{obj} L_{class} + \alpha_{box} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \prod_{k_{ij}}^{obj} L_{CIoU} \right) \tag{3}$$

$$L_{obj}(p_{obj}, p_{iou}) = BCE_{obj}^{sig}(p_{obj}, p_{iou}; w_{obj}) \tag{4}$$

$$L_{cls}(c_{pre}, c_{gt}) = BCE_{cls}^{sig}(c_{pre}, c_{gt}; w_{cls}) \tag{5}$$

K, S^2, and B represent the number of output features, grid cells, and anchor boxes on each grid cell, respectively. α_* denotes the weight of the corresponding component. $\prod_{k_{ij}}^{obj}$ indicates whether the j anchor box of the i grid in the k feature map is a positive sample.

The bounding box loss function has received extensive research attention due to its impact on the localization accuracy of detection results. The conventional localization loss regresses the four coordinate points independently, which has several drawbacks. The IoU loss [36] addresses these limitations by treating the four-point coordinates of the bounding box as a whole for regression calculation. Currently, the most widely adopted approach is CIoU loss [37], which improves upon the IoU method and enhances regression accuracy based on DIoU [37]. The CIoU loss is defined as shown in Equation (9).

$$\alpha = \frac{v}{1 - IoU + v} \tag{6}$$

$$v = \frac{4}{\pi^2}(arctan\frac{w^{gt}}{h^{gt}} - arctan\frac{w}{h})^2 \tag{7}$$

$$CIoU = IoU - \frac{\rho^2(b, b^{gt})}{c^2} - \alpha v \tag{8}$$

$$L_{CIoU}(b, b_{gt}) = 1 - CIoU \tag{9}$$

Among these, $\rho^2(b, b^{gt})$ represents the real distance between the center point of the bounding box and ground truth, while c represents the diagonal distance of the smallest enclosing regions.

4. Experiments and Results

In this section, we conduct a comprehensive evaluation of the CFA-YOLO through a series of experiments. We begin by providing an introduction of two datasets in the experiments in Section 4.1. Next, we introduce the evaluation metrics commonly employed for target detection in Section 4.2. The experimental configuration is described in Section 4.3. Furthermore, in Section 4.4, we compare the performance of CFA-YOLO with several commonly used methods. To validate the effectiveness of each proposed scheme, we conduct ablation experiments in Section 4.5. Lastly, we present some visualization results in Section 4.6.

4.1. Traffic Scenario Dataset

The TT100K dataset [11] was compiled and published by a joint lab of Tsinghua University and Tencent. It has 9180 data samples, which contain a relatively comprehensive collection of traffic signs and consist of images captured from Tencent Street View panoramas using six wide-angle SLR cameras with high pixel counts.

The BDD100K dataset [12] was released in 2018 by the Berkeley Artificial Intelligence Laboratory (BAIR) at the University of California, Berkeley. It includes 80,000 images depicting real-world driving scenarios across different categories, encompassing diverse weather conditions, environments, and times of day.

For the TT100K dataset, we analyzed different types of traffic signs and selected 45 categories with more than 100 samples to address the sample discrepancy caused by varying sign categories. Furthermore, we focus on the two small object categories in the BDD100K dataset. Table 1 provides a concise description of the two datasets.

Table 1. Introduction table of datasets.

Dataset	Training and Validation Images	Test Images	Images Resolution	Categories
TT100K [11]	8487	970	2048 × 2048	45
BDD100K [12]	72,000	8000	1280 × 720	6

4.2. Evaluation Metric

The mean Average Precision (mAP) is a widely used evaluation metric in object detection algorithms. Average Precision (AP) is a measure of the detection capability of a trained model for a specific category of interest. It is calculated as the area enclosed by the Precision–Recall (P–R) curve and the coordinate axes, as defined in Equation (10).

$$Average\ Precision(AP) = \sum_{i=1}^{n-1}(R_{i+1} - R_i)\max_{R' \geqslant R} P(R') \tag{10}$$

In the formula, the precision P and the recall R_i are calculated as shown in Equation (11). The term $\max_{R' \geqslant R} P(R')$ represents the maximum precision value for recall R' after applying curve smoothing techniques.

$$Precision(P) = \frac{TP}{TP + FP} \quad Recall(R) = \frac{TP}{TP + FN} \tag{11}$$

In detection tasks involving multiple categories, each category is assigned an Average Precision (AP) value. The mean Average Precision (mAP) is then calculated as the average of all AP values, representing the overall detection performance across all classes. The calculation formula for mAP is shown in Equation (12), where K represents the number of categories.

$$mAP = \frac{\sum_{i=1}^{K} AP_i}{K} \tag{12}$$

4.3. Experimental Configuration

For the experiments, we trained the model weights from scratch to minimize the influence of external factors and ensure fairness. We utilized the SGD optimizer [38] for training on both the TT100K dataset and BDD100K dataset. The dataset will be split into three parts including training, validation, and test sets in a ratio of 8:1:1. The hyperparameter configuration used in the experiments was as follows: a batch size of 32, initial learning rate of 0.01 with a final learning rate of 0.01, momentum of 0.937, and weight decay of 0.0005. The model was trained for 200 epochs on the TT100K dataset and 100 epochs on the BDD100K dataset. In the training process, the first three epochs employed a warm-up strategy, while the remaining epochs followed a cosine annealing strategy. The warm-up initial learning rate and momentum were set to 0.1 and 0.8, respectively. The detailed information about the experimental platform in this work is shown in Table 2.

Table 2. Configuration of the experimental platform.

Experimental Setting	Configuration
CPU	Intel(R) Core(TM) i7-11700 CPU @2.50 GHz
GPU	NVIDIA GeForce RTX 3090
OS	Ubuntu 20.04
Compiling Tool	PyTorch 1.11.0
Language	Python 3.8

4.4. Comparative Experiments

We conducted experiments on two well-established datasets in the field of traffic analysis to compare the performance of our proposed model with the state-of-the-art methods. The comparative models used in the experiments were implemented using widely accepted codes.

Table 3 presents the comparative experimental results on the TT100K dataset. The data in the table clearly demonstrate that our CFA-YOLO model achieves a significant improvement over the baseline model YOLOv5s, with a 21.4% higher mAP@0.5 accuracy and a resulting mAP@0.5 value of 84.2%. This improvement indicates that we effectively address the limitations of the baseline model in small object detection. Furthermore, when compared to other state-of-the-art object detectors such as YOLOv7-Tiny and YOLOv8s, our algorithm consistently outperforms them, achieving the best results. Remarkably, despite only introducing an additional 4.8M parameters to the baseline model, our method still maintains real-time detection capabilities. Among lightweight models of similar scale, our approach exhibits a distinct advantage in traffic small object detection.

Table 3. Comparison with other mainstream algorithms on the TT100K dataset.

Method	Backbone	mAP/%	Param	Speed (ms)
Faster R-CNN [15]	ResNet-50	56.8	41.4M	26.5
Cascade R-CNN [39]	ResNet-50	66.0	69.1M	31.2
Sparse R-CNN [40]	ResNet-50	64.7	106.1M	32.8
RetinaNet [41]	ResNet-50	44.8	37.1M	25.6
YOLOv5s	CSP-Darknet53-C3	62.8	7.1M	9.3
YOLOv5m	CSP-Darknet53-C3	76.0	21.1M	12.6
YOLOv7 [42]	ELAN-Net	60.7	36.7M	10.4
YOLOv3-Tiny	Darknet53	61.9	8.8M	3.1
YOLOv4-Tiny	CSP-Darknet53	56.4	3.2M	4.2
YOLOv7-Tiny	ELAN-Net	49.2	6.2M	7.5
YOLO-MAXNet [7]	SA-MobileNeXt	74.5	14.1M	15.6
YOLOv8s	CSP-Darknet53-C2f	82.9	11.2M	8.2
CFA-YOLO (Ours)	CSP-Darknet53-C3	**84.2**	11.9M	10.4

To ensure the validity of our experiments, we also conducted experiments on the BDD100K dataset, and the experimental results are presented in Table 4. Since the BDD100K dataset consists of targets of various scales, with our focus being on small-scale objects, we not only provide the mAP@0.5 values for all categories but also the AP values specifically for two types of small targets: traffic signs and traffic lights. It can be observed from the AP values of these two categories that our model outperforms others in terms of small object detection, achieving AP values of 71.2% and 67.6% for traffic signs and traffic lights, respectively. Although the mAP@0.5 value may not reach state-of-the-art results, the overall accuracy remains highly competitive. Additionally, our method exhibits a low parameter count and is capable of real-time detection.

Table 4. Comparison with other mainstream algorithms on the BDD100K dataset.

Method	Backbone	Traffic Sign/%	Traffic Light/%	mAP/%	Param	Speed (ms)
Faster R-CNN [15]	ResNet-50	66.3	52.4	65.3	41.2M	34.4
Cascade R-CNN [39]	ResNet-50	66.1	52.1	65.2	68.9M	40.1
Sparse R-CNN [40]	ResNet-50	69.7	64.3	66.4	106.0M	36.9
RetinaNet [41]	ResNet-50	64.8	52.0	63.1	36.2M	34.7
YOLOv5s	CSP-Darknet53-C3	61.9	56.4	61.1	7.1M	6.9
YOLOv5m	CSP-Darknet53-C3	67.3	62.1	65.9	20.9M	9.2
YOLOv3 [30]	Darknet53	70.9	65.8	69.0	61.5M	9.1
YOLOv4 [5]	CSP-Darknet53	71.1	66.0	**69.4**	60.4M	14.0
YOLOv3-Tiny	Darknet53	34.8	27.2	39.8	8.7M	2.8
YOLOv4-Tiny	CSP-Darknet53	40.1	29.6	41.5	3.1M	3.9
YOLOv7-Tiny	ELAN-Net	57.1	52.8	59.0	6.1M	7.0
YOLO-MAXNet [7]	SA-MobileNeXt	68.6	63.4	66.4	13.9M	15.0
YOLOv8s	CSP-Darknet53-C2f	63.9	57.9	64.8	11.1M	6.8
CFA-YOLO (Ours)	CSP-Darknet53-C3	**71.2**	**67.6**	67.6	11.8M	9.7

The extensive experimental data mentioned above demonstrate that the model we designed is well-suited for detecting small targets in the traffic field. Moreover, the superiority of performance across different datasets further confirms the versatility of CFA-YOLO in addressing these problems.

4.5. Ablation Experiments

We conducted numerous ablation experiments on the TT100K dataset to validate the effectiveness of our proposed approaches. Table 5 presents the various improvement schemes and their corresponding experimental results.

The first row of the table represents the baseline model. Scheme (a) incorporates the SE and SPPF modules into the baseline model, resulting in a slight improvement in accuracy. This indicates that the SE module enhances the attention to crucial information within deep networks. Scheme (b) showcases the performance of our designed CAPAN neck network, which serves as the core component of CFA-YOLO. By comparing the data in the table, we can observe that using CAPAN increases the baseline mAP@0.5 from 62.8% to 78.2%, indicating CAPAN's strong feature fusion capability and significant enhancement in small target detection performance. Next, scheme (c) and scheme (d) both employ decoupled detection heads based on CAPAN, differing only in the number of detection heads. Through the experimental data, we discovered that the two detection heads can cover nearly all targets, resulting in a more concise model design. Scheme (e) represents the final results of our proposed designs, which achieve the best results. Our model achieves a 21.4% improvement in mAP@0.5 compared to the baseline model. Fundamentally, the experimental results validate the efficiency and effectiveness of our proposed CFA-YOLO along with enhanced performance.

Table 5. Ablation studies of the CFA-YOLO with different improvement schemes.

Method	SE	SPPF	CAPAN	Decoupled Head	mAP/%	Param
YOLOv5s					62.8	7.1 M
(a)	✓	✓			64.1	7.1 M
(b)			✓		78.2	7.1 M
(c)			✓	3 Heads	83.9	14.4 M
(d)			✓	2 Heads	83.7	11.9 M
(e)	✓	✓	✓	2 Heads	84.2	11.9 M

We utilize Grad-CAM [43] to generate heatmaps for both the CFA-YOLO model and the baseline model on the TT100K dataset, as depicted in Figure 6. The baseline model is highly affected by the ambient noise and does not detect the target in the first image, while our model still recognizes it. It can be observed in the second image that our model focuses on the target more accurately. By examining the images, it becomes evident that our model exhibits greater sensitivity towards small targets and can precisely focus on the intended targets.

4.6. Visualization of Experimental Results

To provide a more intuitive comparison between the CFA-YOLO algorithm and the baseline model, we present the detection results of several images from the test set in Figure 7. By examining the visualization results, we can observe that the baseline model has more instances of missed detections for traffic signs because targets are highly integrated with the background environment, whereas our method successfully identifies a higher number of traffic signs. Our model also exhibits superior performance in detecting more distant traffic signs and handling dense traffic signage scenarios. The visualization of the detection results serves as further evidence of the effectiveness of our approach.

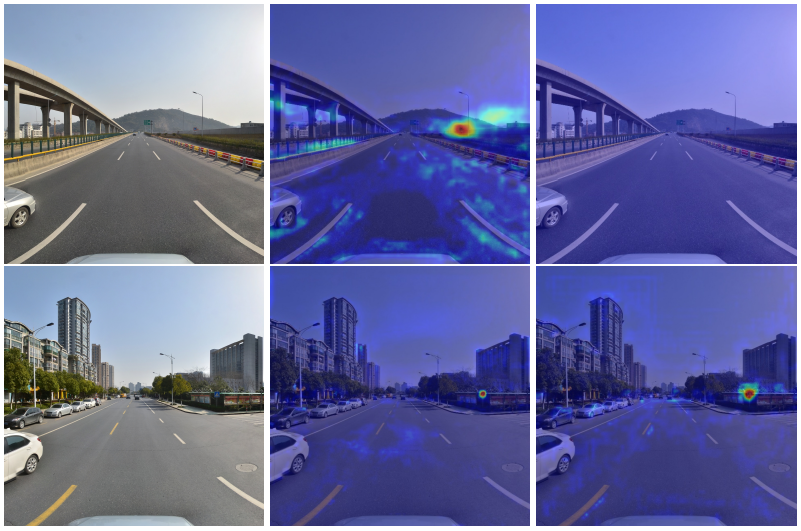


Figure 6. Heatmap visualization of detection results from CFA-YOLO model and baseline model.

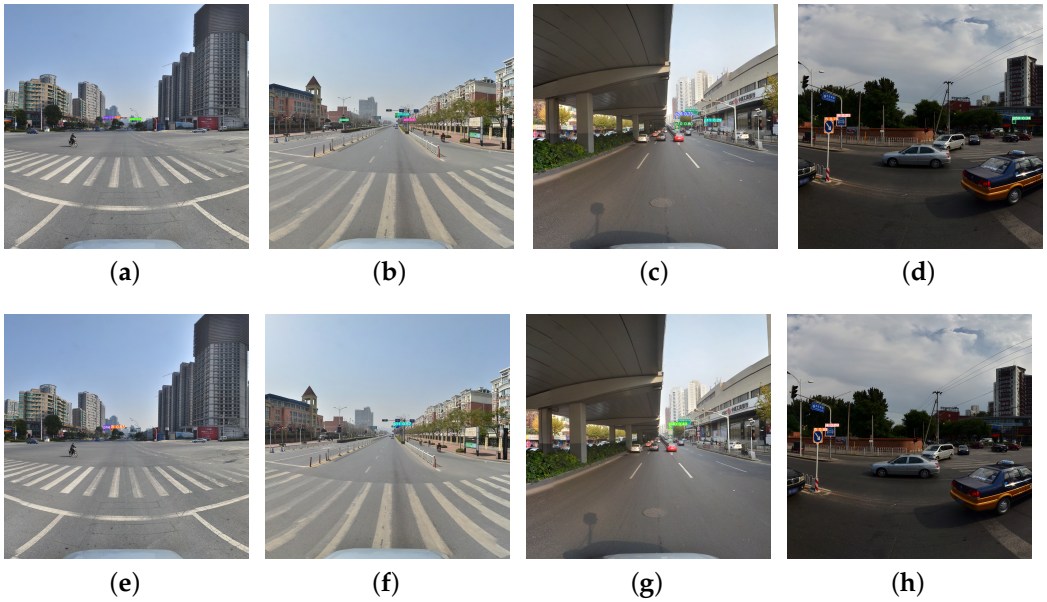


Figure 7. Partial detection results visualization of YOLOv5s and CFA-YOLO on the TT100K test dataset. (**a**–**d**) the detection results of CFA-YOLO. (**e**–**h**) the detection results of YOLOv5s.

5. Conclusions

In order to address the challenge of detecting small traffic objects, we propose an object detector called CFA-YOLO in this work. Our method achieves remarkable detection performance for small targets while maintaining real-time detection standards. Firstly, our model incorporates the Squeeze-and-Excitation (SE) block and the Spatial Pyramid Pooling Fast (SPPF) module in the backbone to select more crucial feature information for various objects. Additionally, we introduce a novel neck network named the Cross-layer Alternating Pyramid Aggregation Network (CAPAN) to facilitate efficient feature fusion. CAPAN reduces feature redundancy while maintaining strong feature interaction capabilities, making it a key contribution of the CFA-YOLO. Furthermore, through analysis of the features in the neck network, we adjust the number of detection heads to two, which shortens the time-consuming training. We also replace the original one-step coupled

detection heads with decoupled heads to further boost the network's training effectiveness. Extensive experimental results demonstrate that our algorithm achieves higher accuracy compared to SOTA lightweight detectors with a similar magnitude of parameters. However, there is still room for improvement in the performance of multi-scale object detection. In future work, we will focus on addressing targets at different scales to tackle more challenging problems in traffic scenarios. In addition, we are going to test and optimize the algorithm on a more portable platform to meet the demands of the computing-limited edge devices in real application scenarios.

Author Contributions: Conceptualization, Q.C. and X.H.; methodology, Q.C. and X.H.; software, Q.C. and X.H.; validation, Q.C., X.H. and Y.L.; formal analysis, Q.C. and X.H.; investigation, Q.C. and X.H.; resources, Q.C. and X.H.; data curation, Q.C. and X.H.; writing—original draft preparation, Q.C. and X.H.; writing—review and editing, Q.C., X.H. and Y.L.; visualization, Q.C. and X.H.; supervision, Q.C., X.H. and Y.L.; project administration, Q.C. and X.H.; funding acquisition, X.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by National Natural Science Foundation of China (NSFC) (61572023, 62272419); Natural Science Foundation of Zhejiang Province (LZ22F020010).

Data Availability Statement: Some or all data, models, or code generated or used during the study are available from the corresponding author by request.

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

References

1. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
2. Dai, J.; He, K.; Sun, J. Instance-aware semantic segmentation via multi-task network cascades. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 3150–3158.
3. Kang, K.; Li, H.; Yan, J.; Zeng, X.; Yang, B.; Xiao, T.; Zhang, C.; Wang, Z.; Wang, R.; Wang, X.; et al. T-cnn: Tubelets with convolutional neural networks for object detection from videos. *IEEE Trans. Circuits Syst. Video Technol.* **2017**, *28*, 2896–2907. [CrossRef]
4. Gu, Y.; Si, B. A novel lightweight real-time traffic sign detection integration framework based on YOLOv4. *Entropy* **2022**, *24*, 487. [CrossRef] [PubMed]
5. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
6. Liu, Y.; Shi, G.; Li, Y.; Zhao, Z. M-YOLO: Traffic sign detection algorithm applicable to complex scenarios. *Symmetry* **2022**, *14*, 952. [CrossRef]
7. He, X.; Cheng, R.; Zheng, Z.; Wang, Z. Small object detection in traffic scenes based on YOLO-MXANet. *Sensors* **2021**, *21*, 7422. [CrossRef] [PubMed]
8. Gong, H.; Mu, T.; Li, Q.; Dai, H.; Li, C.; He, Z.; Wang, W.; Han, F.; Tuniyazi, A.; Li, H.; et al. Swin-transformer-enabled YOLOv5 with attention mechanism for small object detection on satellite images. *Remote Sens.* **2022**, *14*, 2861. [CrossRef]
9. Liu, H.; Sun, F.; Gu, J.; Deng, L. Sf-yolov5: A lightweight small object detection algorithm based on improved feature fusion mode. *Sensors* **2022**, *22*, 5817. [CrossRef]
10. Qi, G.; Zhang, Y.; Wang, K.; Mazur, N.; Liu, Y.; Malaviya, D. Small object detection method based on adaptive spatial parallel convolution and fast multi-scale fusion. *Remote Sens.* **2022**, *14*, 420. [CrossRef]
11. Zhu, Z.; Liang, D.; Zhang, S.; Huang, X.; Li, B.; Hu, S. Traffic-Sign Detection and Classification in the Wild. In Proceedings of the IEEE 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2110–2118.
12. Yu, F.; Xian, W.; Chen, Y.; Liu, F.; Liao, M.; Madhavan, V.; Darrell, T. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv* **2018**, arXiv:1805.04687
13. 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.
14. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
15. 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]
16. Uijlings, J.R.; Van De Sande, K.E.; Gevers, T.; Smeulders, A.W. Selective search for object recognition. *Int. J. Comput. Vis.* **2013**, *104*, 154–171. [CrossRef]
17. Qian, R.; Liu, Q.; Yue, Y.; Coenen, F.; Zhang, B. Road surface traffic sign detection with hybrid region proposal and fast R-CNN. In Proceedings of the 2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), Changsha, China, 13–15 August 2016; pp. 555–559.

18. Li, J.; Liang, X.; Shen, S.; Xu, T.; Feng, J.; Yan, S. Scale-aware fast R-CNN for pedestrian detection. *IEEE Trans. Multimed.* **2017**, *20*, 985–996. [CrossRef]
19. Fan, Q.; Brown, L.; Smith, J. A closer look at Faster R-CNN for vehicle detection. In Proceedings of the 2016 IEEE Intelligent Vehicles Symposium (IV), Gothenburg, Sweden, 19–22 June 2016; pp. 124–129.
20. Zhao, X.; Li, W.; Zhang, Y.; Gulliver, T.A.; Chang, S.; Feng, Z. A faster RCNN-based pedestrian detection system. In Proceedings of the 2016 IEEE 84th Vehicular Technology Conference (VTC-Fall), Montreal, QC, Canada, 18–21 September 2016; pp. 1–5.
21. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 779–788.
22. 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 European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
23. Kim, H.; Lee, Y.; Yim, B.; Park, E.; Kim, H. On-road object detection using deep neural network. In Proceedings of the 2016 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia), Seoul, Republic of Korea, 26–28 October 2016; pp. 1–4.
24. Xie, L.; Ahmad, T.; Jin, L.; Liu, Y.; Zhang, S. A new CNN-based method for multi-directional car license plate detection. *IEEE Trans. Intell. Transp. Syst.* **2018**, *19*, 507–517. [CrossRef]
25. Jensen, M.B.; Nasrollahi, K.; Moeslund, T.B. Evaluating state-of-the-art object detector on challenging traffic light data. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 9–15.
26. Yang, W.; Zhang, J.; Wang, H.; Zhang, Z. A vehicle real-time detection algorithm based on YOLOv2 framework. In Proceedings of the Real-Time Image and Video Processing 2018, Orlando, FL, USA, 15–19 April 2018; Volume 10670, pp. 182–189.
27. Guo, C.; Fan, B.; Zhang, Q.; Xiang, S.; Pan, C. Augfpn: Improving multi-scale feature learning for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 12595–12604.
28. Zuo, Z.; Yu, K.; Zhou, Q.; Wang, X.; Li, T. Traffic signs detection based on faster r-cnn. In Proceedings of the 2017 IEEE 37th International Conference on Distributed Computing Systems Workshops (ICDCSW), Atlanta, GA, USA, 5–8 June 2017; pp. 286–288.
29. Wang, S.y.; Qu, Z.; Li, C.j.; Gao, L.y. BANet: Small and multi-object detection with a bidirectional attention network for traffic scenes. *Eng. Appl. Artif. Intell.* **2023**, *117*, 105504. [CrossRef]
30. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
31. 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]
32. 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, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
33. 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; pp. 8759–8768.
34. Song, G.; Liu, Y.; Wang, X. Revisiting the sibling head in object detector. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11563–11572.
35. Wu, Y.; Chen, Y.; Yuan, L.; Liu, Z.; Wang, L.; Li, H.; Fu, Y. Rethinking classification and localization for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10186–10195.
36. Jiang, B.; Luo, R.; Mao, J.; Xiao, T.; Jiang, Y. Acquisition of localization confidence for accurate object detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 784–799.
37. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU loss: Faster and better learning for bounding box regression. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12993–13000.
38. Ruder, S. An overview of gradient descent optimization algorithms. *arXiv* **2016**, arXiv:1609.04747.
39. Cai, Z.; Vasconcelos, N. Cascade r-cnn: Delving into high quality object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6154–6162.
40. Sun, P.; Zhang, R.; Jiang, Y.; Kong, T.; Xu, C.; Zhan, W.; Tomizuka, M.; Li, L.; Yuan, Z.; Wang, C.; et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 14454–14463.
41. 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, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
42. 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. *arXiv* **2022**, arXiv:2207.02696.
43. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.