drones

MDPI

*Article*

# Research on Improved YOLOv5 Vehicle Target Detection Algorithm in Aerial Images

**Xue Yang [1,2], Jihong Xiu [1,*] and Xiaojia Liu [1,2]**

[1] Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130000, China
[2] University of Chinese Academy of Sciences, Beijing 100000, China
* Correspondence: xiujihong@ciomp.ac.cn

**Abstract:** Aerial photoelectric imaging payloads have become an important means of reconnaissance and surveillance in recent years. However, aerial images are easily affected by external conditions and have unclear edges, which greatly reduces the accuracy of imaging target recognition. This paper proposes the M-YOLOv5 model, which uses a shallow feature layer. The RFBs module is introduced to improve the receptive field and detection effect of small targets. In the neck network part, the BiFPN structure is used to reuse the underlying features to integrate more features, and a CBAM attention mechanism is added to improve detection accuracy. The experimental results show that the detection effect of this method on the DroneVehicle dataset is better than that of the original network, with the precision rate increased by 2.8%, the recall rate increased by 16%, and the average precision increased by 2.3%. Considering the real-time problem of target detection, based on the improved model, the Clight-YOLOv5 model is proposed, by lightweighting the network structure and using the depth-separable convolution optimization module. After lightweighting, the number of model parameters is decreased by 71.3%, which provides a new idea for lightweight target detection and proves the model's effectiveness in aviation scenarios.

**Keywords:** object detection; aerial images; YOLOv5; feature fusion; lightweight network



**Citation:** Yang, X.; Xiu, J.; Liu, X. Research on Improved YOLOv5 Vehicle Target Detection Algorithm in Aerial Images. *Drones* **2024**, *8*, 202. https://doi.org/10.3390/drones8050202

Academic Editor: Anastasios Dimou

Received: 2 April 2024
Revised: 12 May 2024
Accepted: 13 May 2024
Published: 16 May 2024



## 1. Introduction

Aerial imaging is a technology that utilizes aircraft such as drones or aerial cameras to obtain ground information [1]. It allows for the multi-angle, all-round collection of ground targets and has been widely applied as an indispensable optical remote sensing method. Due to the long photographic distance of aerial payload imaging, small targets such as cars have fewer effective pixels in the image, often getting lost in complex background clutter, thereby further increasing the difficulty of detecting vehicle targets. The rapid and effective detection and identification of small targets such as cars in aerial images under complex background conditions are of great importance [2,3].

This paper improves the current target detection algorithm based on images captured by aerial cameras, with aerial vehicle images as the foundation [4,5]. It adopts the YOLOv5 algorithm to analyze and study existing images and improve the original algorithm for identification. Firstly, according to the characteristics of aerial images, research is conducted on small vehicle targets with low pixel proportions in the images. Attention is focused on how to enhance target detection accuracy under complex and blurred backgrounds. The specific research content is as follows.

The edge features of aerial images are not distinct, and the targets occupy a relatively small area in the original image. Moreover, vehicle features are often characterized by single features in small targets. Therefore, the convolutional model structure is modified to enhance the feature information of small targets and increase the receptive field to improve the detection accuracy of small targets. Various datasets covering different objects and

scenes are utilized. Images obtained at different heights contain objects of various scales, images obtained at different angles display objects of various shapes, and images obtained under different lighting conditions contain more shadow occlusions and brightness changes. Additionally, experimental environments are configured.

In response to the need for real-time detection in aerial images with complex backgrounds and blurred edges [6,7], all performance metrics of the original algorithm need improvement. This paper begins by analyzing the YOLOv5 algorithm and its backbone network and improving it based on the characteristics of the dataset. To address the characteristic of small targets, improvements are made to shallow features and feature pyramid structures, enhancing low-level features to make the model focus more on crucial features, thereby effectively improving the ability to extract features. Finally, the algorithms before and after improvement are compared to validate the effectiveness of the improved model.

Addressing the requirement for real-time detection in aerial images, this paper proposes a lightweight vehicle target detection algorithm based on deep learning [8,9]. Researching common lightweight methods, this algorithm focuses on detecting small targets in aerial images and conducts lightweight processing while optimizing the model. Initially, the network structure is simplified by reducing redundant down-sampling sections and optimizing Bottleneck in the C3 module through separable convolution operations. Experimental data and evaluation metrics demonstrate the effectiveness of the proposed lightweight approach.

## 2. Model and Method

### *2.1. Introduction to the YOLOv5 Algorithm*

This paper adopts YOLOv5s, whose structure is shown in Figure 1, as the base model, saving some computational costs while maintaining a fast detection speed and relatively high detection accuracy [10,11]. Firstly, images are resized to 640 × 640 before being fed into the network. Building upon the Mosaic augmentation method, the mixup data augmentation method is introduced. In this method, two input images undergo certain flips and scaling, and are then mixed together according to a certain ratio to form a new image that is passed into the network. By linearly interpolating between samples and labels, new training samples are generated, which can enhance the model's robustness and generalization, reduce noise in the samples, and alleviate overfitting. Assuming samples are random and λ ranges from 0 to 1, the core formulas are shown in Equation (1).

$$\begin{aligned} \widetilde{x} &= \lambda x_i + (1-\lambda)x_j \\ \widetilde{y} &= \lambda y_i + (1-\lambda)y_j \end{aligned} \tag{1}$$

Using the SPPF module, three 5 × 5 pooling operations are employed instead of a single 13 × 13 pooling operation, and two 5 × 5 pooling operations are used instead of a 9 × 9 pooling operation, resulting in higher computational efficiency. In the multi-scale object detection algorithm of YOLOv5, five down-sampling operations are performed, producing feature layers P1 to P5. Each Pi-1 layer is half the size of the Pi layer. The feature layers P3 to P5 are then passed into the Neck for feature fusion. However, as the feature scale decreases, the resolution of objects, after multiple down-sampling operations, also gradually decreases, further increasing the difficulty of detecting aerial vehicle targets.

The output consists mainly of the Neck feature fusion structure and the detection part. Combining the feature pyramid with the path aggregation structure [12,13], strong semantic information from the top layers of the FPN is passed down to the lower layers. Then, through the bottom-up PAN structure, strong localization features are conveyed, achieving parameter aggregation for different detection layers. This approach helps enhance the detection effectiveness of vehicles of various sizes in aerial images.

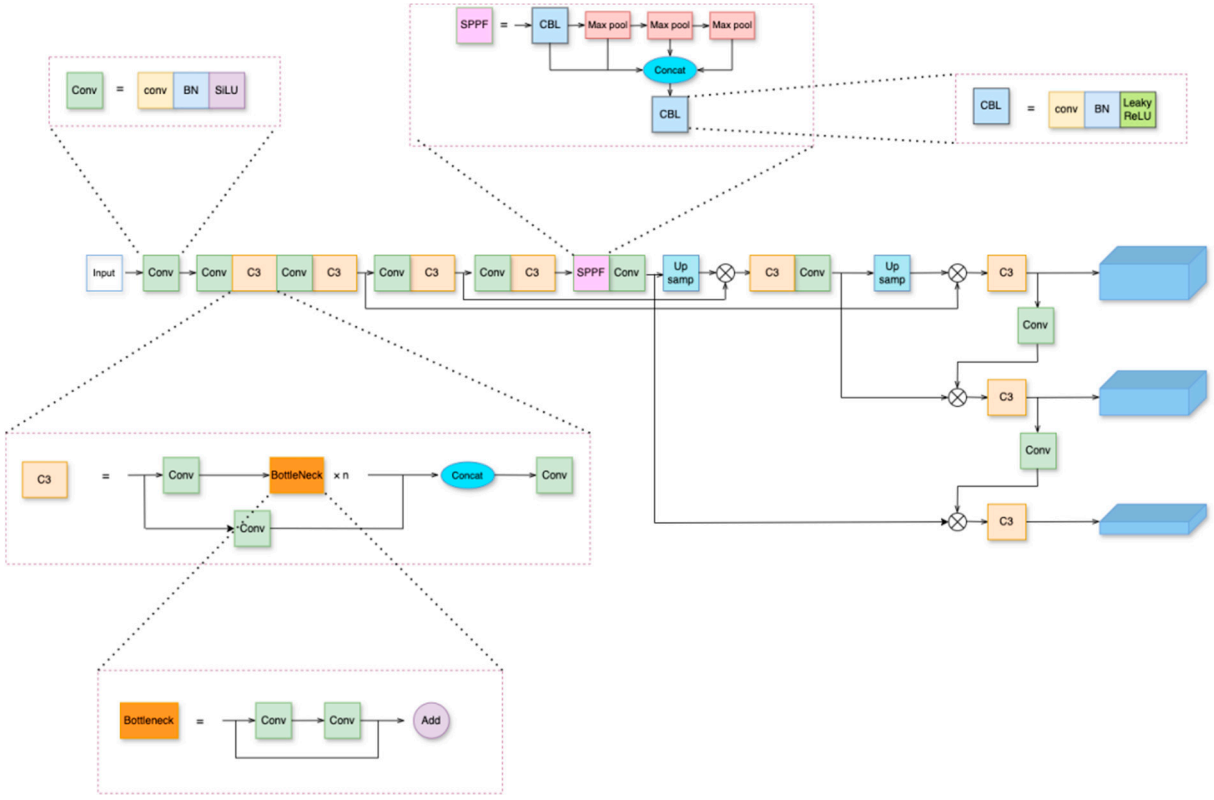



**Figure 1.** YOLOv5 network structure diagram.

### *2.2. Model Improvement Based on Receptive Fields*

In aerial image scenarios, the key to detection lies in effectively distinguishing between the edges of targets and the background. However, targets captured by aerial cameras are often small and can be submerged in the background, leading to potential false positives [14,15]. To address the challenges posed by factors such as the limited number of pixels occupied by targets, complex backgrounds, and low contrast, a vehicle small target detection algorithm based on the combination of attention mechanism and receptive field module is proposed. This algorithm is validated through experiments on publicly available datasets to verify the effectiveness of the improved model.

The RFBs module simulates the characteristics of human visual neurons and draws inspiration from the parallel thinking of Inception. It utilizes dilated convolutions with dilation rates of 1, 3, and 5 to adjust the sizes of convolutional kernels, enabling lightweight tasks and enhancing the network's feature extraction capabilities. Compared to the original RFB module, the RFBs module reduces parameter volume while also capturing smaller receptive fields, effectively maintaining the lightweight nature of the YOLOv5 network's speed while improving detection accuracy. The structure of the RFBs module, as shown in Figure 2, employs different convolution combinations of $3 \times 3$, $1 \times 3$, and $1 \times 1$, which are equivalent to convolutions of $3 \times 3$, $5 \times 5$, and $7 \times 7$, respectively. Furthermore, dilated convolutions with rates of 1, 3, and 5 are introduced to increase the receptive field. Feature fusion is achieved through concatenation, and the detection speed of the model is enhanced using $1 \times 1$ convolutions.

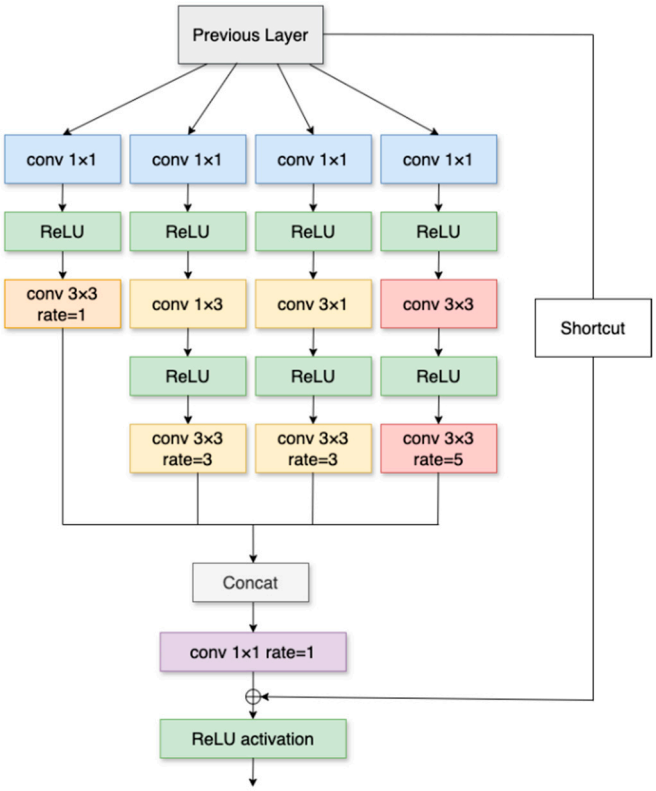



**Figure 2.** RFBs structure diagram.

This paper applies the RFBs module to the shallow layers of the YOLOv5 network structure, with the following advantages: (1) it can focus on relatively localized information; (2) it significantly reduces computational complexity, enhancing the network's detection speed; and (3) it can be easily integrated into the network. In conclusion, based on the theoretical analysis above, utilizing the receptive field module can improve object detection accuracy while enhancing network performance.

### *2.3. Improvements Based on Attention Mechanism*

Due to the blurred edges and small pixel occupancy of objects in aerial images, although the receptive field module can help the model focus more on local information and improve detection accuracy, the characteristics of small, dense, and easily mixed targets in the background warrant the consideration of introducing an attention mechanism. This allows the network to pay more attention to important information such as vehicle targets, reducing the impact of irrelevant information on detection results and improving model detection performance.

CBAM is a lightweight attention module that scores both channel and spatial attention [16]. Channel attention module and spatial attention module are sequentially applied to achieve adaptive feature refinement. The structure of the CBAM attention module, which is shown in Figure 3, enhances the ability to extract target features, allowing the network to effectively focus on the extraction region, thereby improving detection performance. Moreover, it can be widely applied in fields such as natural language processing.

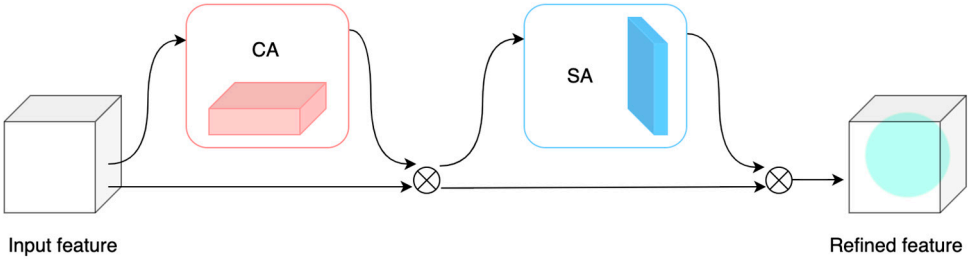



**Figure 3.** CBAM module structure diagram.

Incorporating the CBAM module into the Neck section, the RC-YOLOv5 model offers the following advantages.

Due to the characteristics of aerial datasets including blurred edges and small and dense targets prone to be mixed in the background, introducing an attention mechanism allows the network to focus on important information, reducing the influence of irrelevant information on detection results. This enables the network to pay more attention to small vehicle targets, thereby improving detection accuracy.

1. CBAM enables the network to simultaneously focus on important information in both channel and spatial dimensions, precisely locating vehicle targets. Unlike standalone attention modules that may lose information, CBAM combines the advantages of channel and spatial attention mechanisms, resulting in better detection performance.
2. The CBAM module incurs low computational costs, can be integrated into any network, and is easy to operate with plug-and-play capability, allowing end-to-end training.
3. By serializing information in two dimensions and adaptively refining features through the multiplication of the two types of information, CBAM effectively generates output feature maps.

In summary, introducing attention modules into the YOLOv5 network model is highly necessary. It not only allows the network to focus more on important information but also effectively saves resources and enhances network performance. By placing the CBAM module before the fusion of multi-scale features in the original network, the enhanced feature maps are fused, allowing for the discarding of redundant information and focusing on key information. This facilitates the fusion of weighted important information, resulting in feature maps containing more and more critical valid information. For aerial images, this not only improves the detection accuracy of vehicle targets in the input images but also enhances the model's localization speed. Adding the CBAM module enables the network to continuously learn more effective features as it deepens, thereby enhancing network performance.

### *2.4. Cross-Connected Vehicle Detection Based on Feature Pyramid*

To achieve a more accurate detection of vehicle targets captured by aerial photography, reduce interference from backgrounds similar to object targets, improve detection accuracy, and effectively address the issue of information loss, enabling the network to focus more on the position and edge information of vehicle targets in the image, particularly on the areas where small vehicle targets are located, this subsection considers leveraging a feature pyramid structure. It involves cross-connecting low-level features into high-level features, combining bottom-up and top-down pathways, thereby enhancing the network's robustness. High-level features can capture detailed features from low-level ones, while low-level features can also incorporate semantic information from high-level ones, thus improving the network's feature extraction and detection capabilities. Therefore, this chapter primarily focuses on improving the model's detection accuracy and performance by optimizing and enhancing the RC-YOLOv5 based on the feature pyramid structure.

By adding skip connections, features extracted by the backbone are transmitted to the fused features, enabling more features to be integrated without additional costs and

retaining more features. A weighted strategy is proposed to assign weights to features, allowing the adjustment of their importance and directing the network's attention to crucial information. The repeated stacking of both top-down and bottom-up pathways allows for the fusion of more advanced features. Removing nodes with only one input effectively reduces computational costs.

In response to the challenges posed by aerial payload images with small pixels and indistinct edge features, and to enhance the detection effectiveness of targets with low contrast in aerial vehicle images, three improvements are made to the original YOLOv5 model. The M-YOLOv5 model structure is proposed, which involves the following aspects: introducing the RFBs receptive field module in shallow feature layers to enhance the detection accuracy of small-sized targets, implementing higher level feature fusion using the BiFPN structure, and effectively improving detection accuracy by adding the CBAM attention module. These enhancements make the model more suitable for detecting small vehicle targets with unclear edges. The network components and architecture are illustrated in Figure 4.

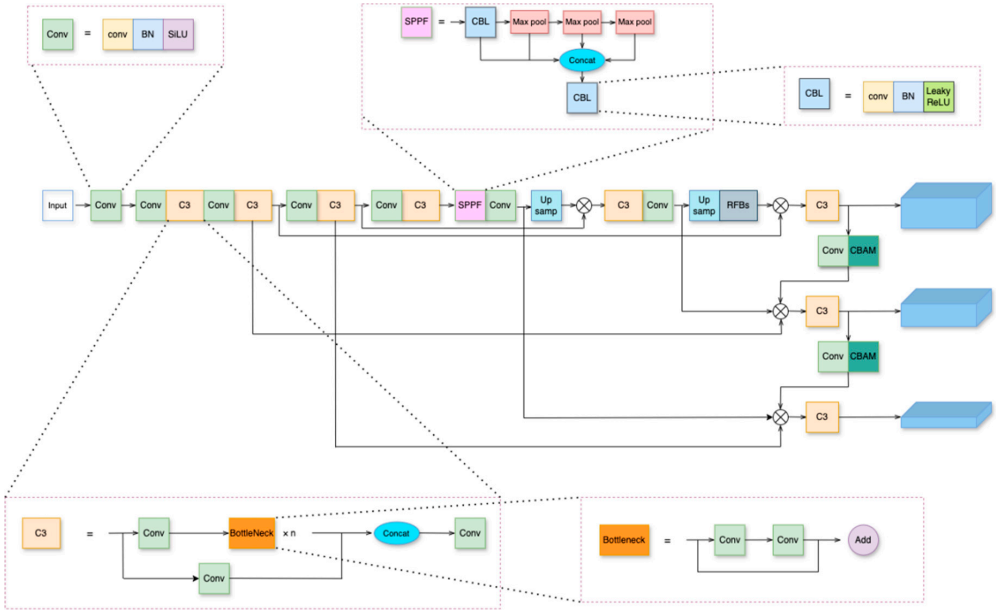



**Figure 4.** M-YOLOv5.

The improved model, as demonstrated by the experimental data, shows enhancements in all evaluation metrics. Furthermore, through multiple sets of comparative experiments, it is proven to effectively reduce both the missed detection rate and false detection rate in the images. It achieves a balance between speed and accuracy, thereby enhancing the overall detection performance.

### *2.5. Lightweight Aerial Vehicle Object Detection Algorithm*

Lightweight object detection plays a crucial role in the field of computer vision, allowing for the avoidance of the computational resource-intensive drawbacks associated with traditional object detection algorithms. It is commonly employed in environments with limited computational resources such as mobile devices. By designing lightweight models and algorithm structures, it is possible to reduce model parameters, save time and resources, reduce costs, and improve real-time responsiveness.

In the M-YOLOv5 network, the backbone undergoes a total of 4 repeated down-sampling operations. To effectively reduce the number of parameters, the network structure is modified by eliminating one down-sampling operation and outputting two detection heads. With these improvements, the network complexity of the target detection network can be reduced, and the detection speed of the model can be effectively increased, thereby

achieving a lightweight detection network. The modified network structure is termed light-YOLOv5, and its simplified structure is illustrated in Figure 5.

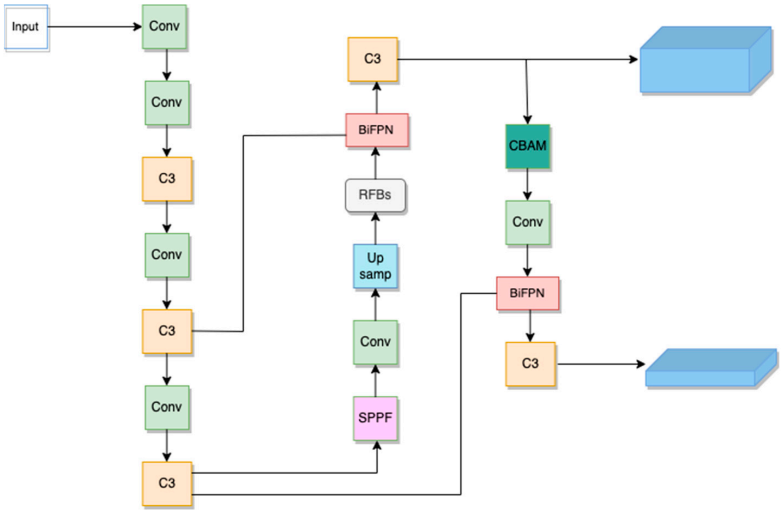



**Figure 5.** light-YOLOv5 structure graph.

Depthwise separable convolution decomposes conventional convolution into depthwise (DW) convolution and pointwise (PW) convolution [17]. In DW convolution, each channel corresponds to a separate convolutional kernel, which, for a 3-channel input image, produces 3 feature maps. In this paper, we utilize depthwise separable convolution to modify the Bottleneck structure, thereby forming modifications to the C3 module. The principles of the two modules are illustrated in Figures 6 and 7. In the original Bottleneck, the input x undergoes two conventional convolutions. In our modified M_Bottleneck, the input x undergoes two depthwise separable convolutions. Utilizing a shortcut connection, the modified M_C3 module in this paper is divided into two branches: one branch stacks n Bottleneck modules using convolutions, and the other branch consists of basic convolution modules.

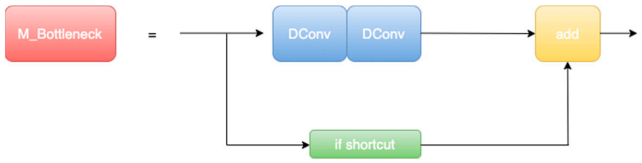



**Figure 6.** M_Bottleneck.

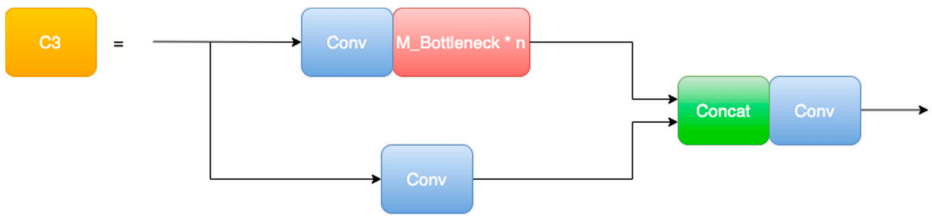



**Figure 7.** M_C3.

## 3. Experimental Results and Analysis

### *3.1. Experimental Environment and Parameter Configuration*

This experiment utilizes the DroneVehicle dataset [18], consisting of a total of 28,439 samples. Among these, 17,990 images are allocated for training, 8980 images for testing, and 1469 images for validation. The experimental configuration for this chapter is presented in Table 1.

**Table 1.** Experiment-related configuration.

| Configuration Information | Title 2 |
|---|---|
| Operating System | Ubuntu 18.04 |
| CPU | Intel Core i7 |
| Memory | 32G |
| GPU | NVIDA GeForce RTX-4070Ti |
| Framework | Pytorch3.8 |

### *3.2. Analysis of Vehicle Detection Results of Small Targets in Aerial Images*

When discussing the strengths and weaknesses of a model, it is necessary to conduct ablation experiments for comparative analysis. This involves gradually adding or removing components from the network to study the effects of different modules on the network. Below, we introduce the relevant metrics used in the experiments.

Precision is the proportion of correctly identified positive samples among all recognized targets, while recall is the proportion of correctly identified targets among all actual targets [19–21]. By plotting the precision–recall (P-R) curve using P and R, we can reflect the relationship between precision and recall. The formulas are as follows:

$$\begin{array}{c} P = \frac{TP}{TP+FP} \times 100\%, \\ R = \frac{TP}{TP+FN} \times 100\% \end{array} \tag{2}$$

The F1 score is the harmonic mean of precision and recall, ranging from 0 to 1, which considers the importance of both precision and recall. The formula is as follows:

$$F_1 = \frac{2TP}{2TP + FP + FN} \tag{3}$$

AP stands for the area under the precision–recall (P-R) curve. mAP, which stands for mean average precision, is the average of AP values for multiple categories, which can validate the model's performance. The formula is as follows:

$$\begin{array}{c} AP = \int_0^1 P(R)dR, \\ mAP = \frac{\sum P_A}{N_c} \end{array} \tag{4}$$

GFlops represents the unit of computational performance of a model, standing for giga floating-point operations per second, which means billions of floating-point operations per second. It is one of the standards for evaluating computational capabilities.

To verify the detection performance of different improvement methods, a comparative analysis was conducted through ablation experiments. Based on the feature pyramid structure, the improved object detection method not only increased the recall rate by 16% compared to the original model, and raised mAP@0.5 by 2.3%, but also addressed the problem of the inability to improve precision using methods based on receptive fields and attention mechanisms. This improvement not only enhances the feature extraction capability of the network but also enables a more precise localization of aerial small objects, thereby improving the performance and accuracy of detection. From Table 2, it can be observed that BiFPN increased precision by 2.8%, while the receptive field and attention mechanism played a crucial role in the recall rate, increasing it by 11% and 3%, respectively. This led to an increase in average precision by 1% and 0.7%, respectively. Overall, the improved algorithm showed improvements in evaluation metrics compared to the original algorithm, validating the effectiveness of the improvement methods.

**Table 2.** M-YOLOv5 ablation experiment results.

| Model | P | R | mAP@0.5 |
|---|---|---|---|
| YOLOv5 | 0.824 | 0.82 | 0.818 |
| YOLOv5 + RFBs | 0.824 | 0.93 | 0.828 |
| YOLOv5 + RFBs + CBAM | 0.824 | 0.96 | 0.835 |
| YOLOv5 + RFBs + CBAM + BiFPN | 0.852 | 0.98 | 0.841 |

In summary, the improved M-YOLOv5 effectively enhances the detection accuracy of various objects and reduces the missed detection issues of small vehicle targets in aerial images captured from long distances, thereby significantly improving the model's performance. The mAP of the model increased by 0.023 after the improvement, and its confusion matrix is shown in Figures 8 and 9.

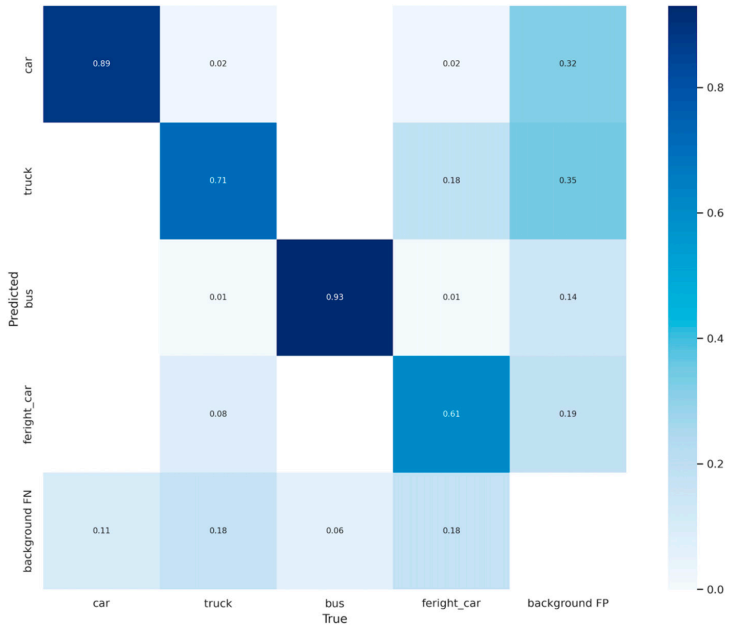



**Figure 8.** The original model's confusion matrix.

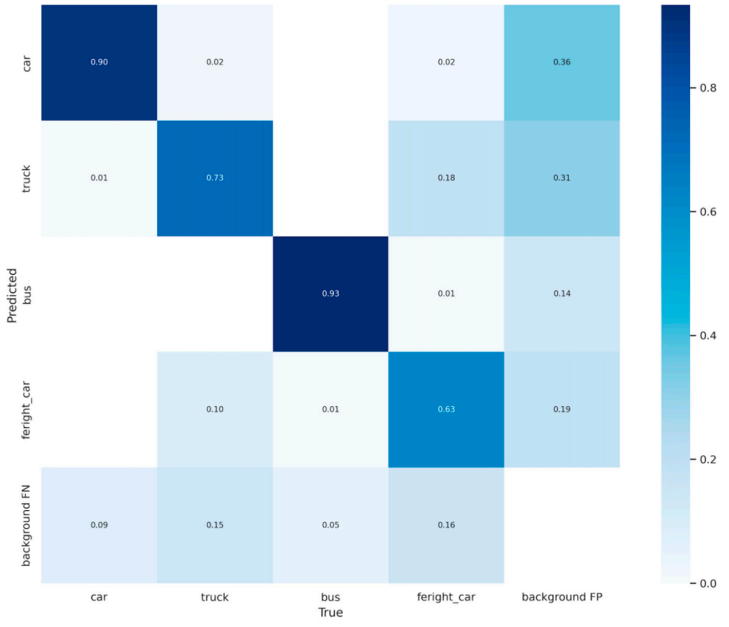



**Figure 9.** M-YOLOv5 model's confusion matrix.

The evaluation metrics of the M-YOLOv5 model used in this study are superior to those of the original model. Through validation on the DroneVehicle dataset, the improved model effectively detects objects of different classes in aerial images. Even for partially occluded targets, the M-YOLOv5 model can detect them well. It strikes a balance between speed and accuracy. After training for 150 epochs, validation was conducted, and some results are shown in Figure 10. In the figure, the bounding boxes of four different colors represent four categories in the dataset: red for cars, yellow for freight cars, pink for trucks, and orange for buses.

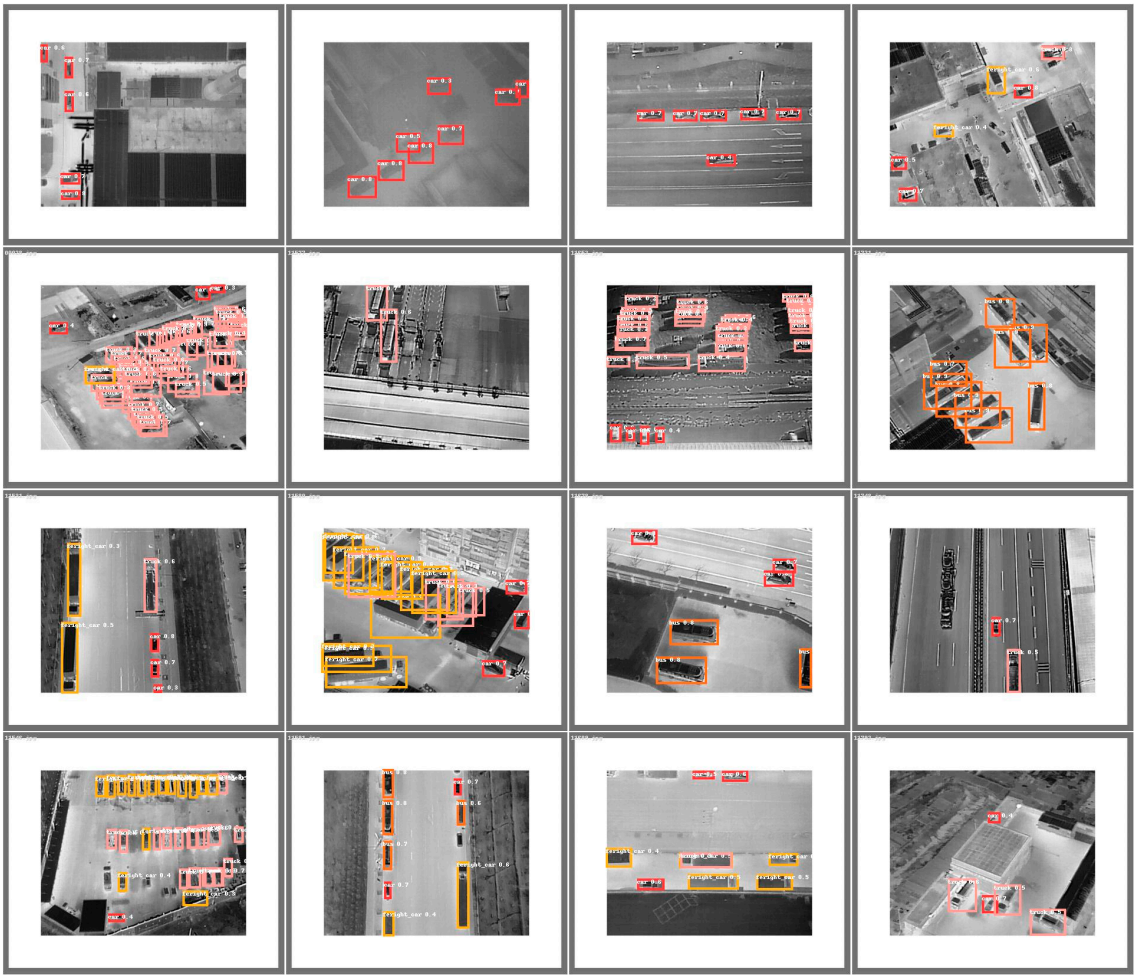


**Figure 10.** Test result.

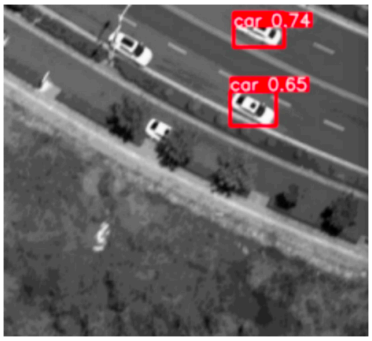
Even for vehicle targets with very few pixels, M-YOLOv5 can still recognize them with high confidence. For example, in the image in the first row and first column of Figure 10, the two vehicle targets on the left are small and easily submerged in the background, making them difficult to identify with the naked eye. However, the method used in this study predicts the "car" category with a confidence of over 50%. Similarly, in the second row and first column of Figure 10, the vehicle at the bottom is a partially occluded incomplete target. The method used in this study accurately predicts its category, further validating the usability and effectiveness of the model.

To verify the effectiveness of the model, this study applied the proposed model to aerial images taken from actual scenarios using an aerial camera. In contrast to the various car images in the DroneVehicle dataset, the target features in the real images are not as clear, which further increases the difficulty of detection for the model. This process effectively validates the generalization and robustness of the improved model.

This paper validates the improved model using aerial images captured by the aerial camera. Due to the inclusion of significant redundant information, large capture range, and complex backgrounds in the images taken by the camera, they differ from the vehicle targets in the DroneVehicle dataset. Targets in aerial real-world images exhibit differences in clarity, contrast, and background blur, with backgrounds predominantly consisting of fields and trees. The shooting locations are random, leading to variations in image clarity. Moreover, factors such as wind can cause camera instability, resulting in poor image quality, shaking, and blurriness, thereby increasing the difficulty of detection.

To annotate the targets in the aerial images, the MakeSense online annotation tool was utilized. This tool enables the annotation of vehicles and other targets in the images and allows for exportation in Annotation and VOC formats. As the DroneVehicle dataset adopts the VOC annotation method, a unified approach was employed during the validation process.

Figure 11 depicts the detection results of applying the M-YOLOv5 model to actual aerial images captured by a certain type of aerial camera. The left side shows the detection results of YOLOv5, while the right side displays the detection results of the M-YOLOv5 model. As shown in (a1,a2) and (b1,b2), for occluded and missed vehicles, the original YOLOv5 fails to detect the targets, whereas the improved model can detect them, reducing the miss detection rate. The detection accuracies are 0.65 and 0.66, respectively. In (c1,c2), the original YOLOv5 incorrectly identifies the zebra crossing as a target, whereas the improved model accurately identifies the true target, reducing the false detection rate. In (d1,d2), (e1,e2), and (f1,f2), the M-YOLOv5 model effectively improves the detection accuracy by 1% to 9%, further validating the detection performance of M-YOLOv5. Under challenging conditions such as occlusion, changing angles/heights, and complex backgrounds, the detection accuracy of the improved model is further improved, and the false detection rate is reduced, verifying that the model detection performance has been further improved.




(**a1**)


(**a2**)

(**b1**)


(**b2**)

**Figure 11.** *Cont.*

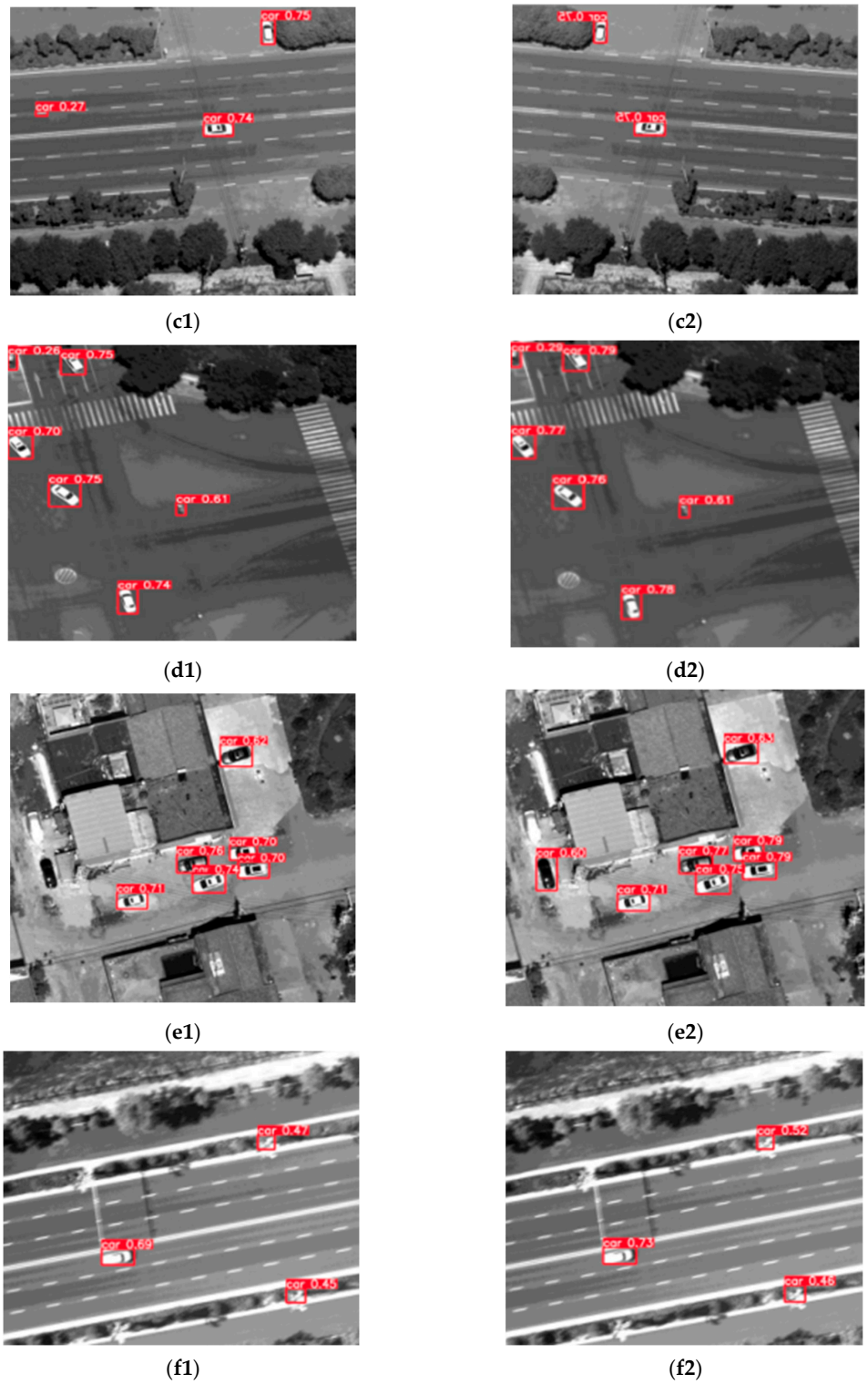



**Figure 11.** Industrial image inspection results. (**a1**–**f1**) represents the original model detection results, (**a2**–**f2**) represents the improved model detection results.

### *3.3. Analysis of Object Detection Results for Lightweight Aerial Vehicles*

The model is based on separable convolution and lightweight model structure to form the Clight-YOLOv5 architecture. The M-YOLOv5 model is defined as Model 1, the lightweight model structure light-YOLOv5 is defined as Model 2, and the application of depth separable convolution, Clight-YOLOv5, is defined as Model 3, which is the optimized

model in this paper. The dataset is inputted into the respective models, and validation is conducted based on evaluation metrics such as mAP@0.5, time, parameter count (Params), and GFlops. The specific experimental results are presented in Table 3.

**Table 3.** Ablation experiment results.

| Model | mAP@0.5 | Speed | Params | GFlops |
|---|---|---|---|---|
| 1 | 0.841 | 4.1 | 9,560,126 | 34.8 |
| 2 | 0.841 | 2.7 | 3,371,068 | 26.9 |
| 3 | 0.765 | 2.6 | 2,743,228 | 23.6 |

There is no decrease in mAP for Model 2 compared to Model 1. The processing time of each image is reduced from 4.1 ms to 2.7 ms, in which inference is reduced from 3.5 ms to 2.1 ms, the number of parameters is reduced from 9,560,126 to 3,371,068 (a reduction of approximately 64.7%), and GFlops are reduced from 34.8 to 26.9. Compared with Model 2, the number of parameters in Model 3 is reduced from 3,371,068 to 2,743,228, a reduction of approximately 18.6%. Compared with Model 1, the mAP of Model 3 dropped by 7.6%, the inference dropped to 1.5 ms, the number of parameters was reduced from 9,560,126 to 2,743,228 (a reduction of about 71.3%), and the GFlops were reduced from 34.8 to 23.6. The speed is increased by about 36.5%. Although the lightweight model loses some accuracy, it leads to a significant reduction in the number of parameters, achieving the expected results. The confusion matrix is shown in Figures 12 and 13.


**Figure 12.** light-YOLOv5 model's confusion matrix.

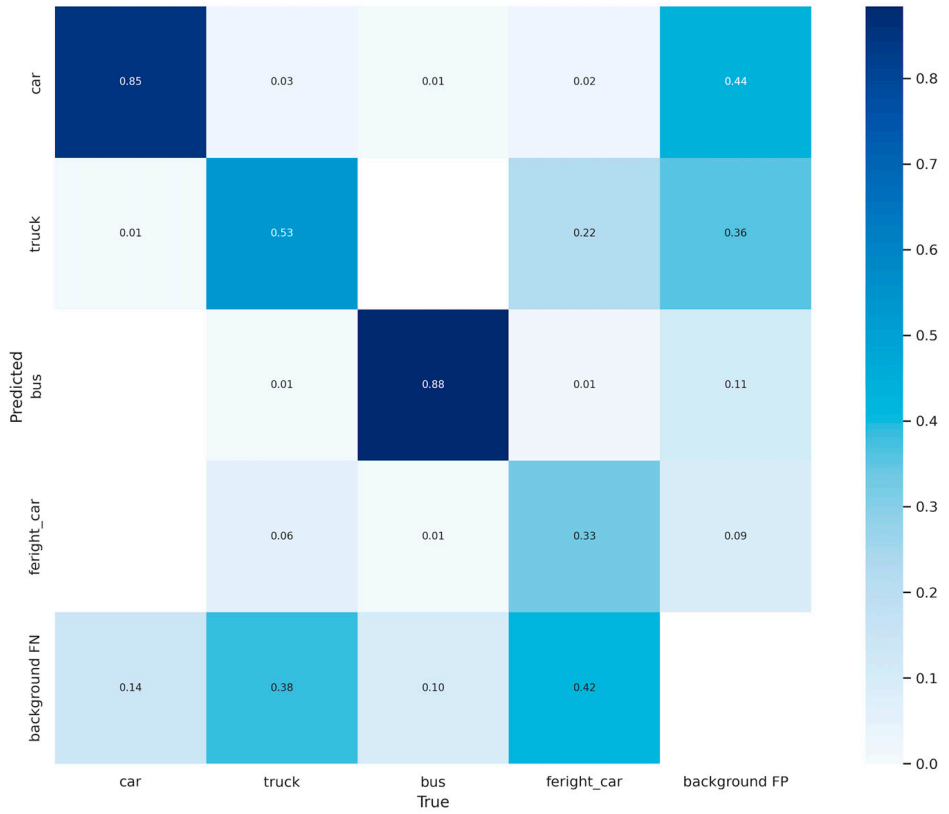



**Figure 13.** Clight-YOLOv5 model's confusion matrix.

In conclusion, under the condition of ensuring minimal loss in accuracy and retaining essential functional modules, experimental results demonstrate that light-YOLOv5, compared to the original model, achieves minimal loss in accuracy while significantly reducing the parameter count through optimized network structures. On the other hand, Clight-YOLOv5, by simplifying the model network structure and modifying module architectures, albeit experiencing a slight decline in detection performance, effectively reduces energy consumption and parameter count, thus enhancing detection speed. Both algorithms can be considered and chosen based on practical applications.

## 4. Discussion

To verify the scientific added value of the research conducted using the improved model, we further verified the original model and the improved model on the VisDrone dataset for a more comprehensive evaluation. The experimental data showed that the accuracy of the original model was 0.7, and the model was unstable. The accuracy of the improved Clight_YOLOv5 model is 0.81, and the model tends to be stable. The recall rate increases from 0.1 to 0.37, an increase of 36%. The detection results are shown in Figure 14. However, the value of the recall rate is low. The reason is that certain types of data in the VisDrone dataset are too small and the angles are too tilted. Although the model has improved the evaluation index, there is still some room for improvement, which we will conduct as a future research direction.

**Figure 14.** VisDrone dataset detection results.

This paper has conducted in-depth research on detecting small vehicle targets in aerial images and has achieved some research results. However, there are still some issues that need to be addressed for further improvement of the algorithm. In order to enhance the algorithm's performance, several future work directions are proposed with the aim of making contributions to the field of object detection. The following will be the focus of future work.

1. The weighted bidirectional fusion method used in this paper increases the size of the model. Future work will focus on reducing the model size while improving its accuracy. Additionally, due to class imbalance, fixed-size anchor boxes may limit the universality of detection. Future work will explore improvements using anchor-free methods.
2. In natural scenes, images often suffer from occlusion. Moreover, images captured from the air may exhibit variations in angle and height, leading to deformations of vehicles. Irregular arrangements of objects may also result in a significant overlap of anchor boxes. Future work will consider using rotated bounding boxes for vehicle detection to reduce overlap and improve detection accuracy, thereby further enhancing the performance of the improved model.

## 5. Conclusions

Addressing the detection challenges of vehicle targets in aerial images involves several difficulties. On one hand, due to the long shooting distance of aerial cameras, targets appear small in the original images with small pixel sizes, and may exhibit angular tilting or jitter, which increases the difficulty of detection. On the other hand, targets in the images are often similar to the background, making many targets prone to being submerged within the background, leading to frequent false positives and false negatives. This paper proposes an improved model, M-YOLOv5, based on YOLOv5, with the following key innovations:

1. The original model's large receptive field in the feature maps makes it prone to losing small targets and information due to fixed receptive fields. To address this, a receptive field module is introduced in the shallow feature layers, utilizing dilated convolutions

to adjust the eccentricity of the convolutional kernel, enabling sampling on the feature map based on different ranges without losing information, thereby enhancing the detection of small targets. The experimental results show that on the DroneVehicle dataset, the mAP@0.5 of the original model is 0.818, while the improved model achieves an mAP@0.5 of 0.828, with improvements in both recall and precision.

2. Since targets in aerial images are small in proportion, it is easy to predict positive samples as background or other class samples. Therefore, a CBAM module is added before feature fusion to enhance the model's focus on blurry small targets, reducing irrelevant information interference and improving the feature extraction capability. The experimental results on the DroneVehicle dataset show that M-YOLOv5 achieves a 1.7% increase in mAP@0.5 compared to the original model, enhancing target localization.
3. The bidirectional feature pyramid structure based on weighted connections strengthens the bottom-level features, enabling the full cross-fusion of bottom-level features with top-level features, enhancing feature transmission across different scales. The experimental results demonstrate that on the DroneVehicle dataset, M-YOLOv5 achieves a 2.3% increase in mAP@0.5 compared to the original model, effectively improving the model's target detection performance.
4. Considering the real-time issue of target detection, a lightweight network structure and optimization module using depth-wise separable convolutions are applied to the improved M-YOLOv5 model, sacrificing some accuracy to reduce the model's parameter count by 71.3%, providing a new direction for lightweight target detection.

Multiple comparative experiments conducted on the DroneVehicle dataset and real aerial images demonstrate the effectiveness of the proposed method, with improvements in precision by 2.8%, recall by 16%, and average precision by 2.3%, reducing false negatives and false positives. Applying the proposed method to aerial images captured by aerial cameras and comparing it with the original model, multiple image comparisons show that the improved M-YOLOv5 enhances detection performance, reducing false negatives and false positives in real image detection, and further validating the excellence of the proposed algorithm.

The proposed lightweight Clight-YOLOv5 model achieves the lightweight processing of the optimized network, significantly reducing the model's parameter count, effectively reducing detection time, and improving detection speed. Through ablation experiments, the effectiveness of both light-YOLOv5 and Clight-YOLOv5 in lightweight processing is confirmed.

**Author Contributions:** Conceptualization, X.Y.; methodology, X.Y.; software, X.Y.; validation, X.Y.; formal analysis, J.X.; data curation, X.Y.; writing—original draft preparation, X.Y.; writing—review and editing, X.L.; project administration, J.X. All authors have read and agreed to the published version of the manuscript.

**Funding:** Major Project of High Resolution Earth Observation System (GFZX0403260312); Major Project of High Resolution Earth Observation System (GFZX04030201).

**Data Availability Statement:** The data underlying the results presented in this paper are not publicly available at this time but maybe obtained from the authors upon reasonable request.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Hamet, P.; Tremblay, J. Artificial intelligence in medicine. *Metabolism* **2017**, *69*, S36–S40. [CrossRef] [PubMed]
2. Zaidi, S.S.A.; Ansari, M.S.; Aslam, A.; Kanwal, N.; Asghar, M.; Lee, B. A survey of modern deep learning based object detection models. *Digit. Signal Process.* **2022**, *126*, 103514. [CrossRef]
3. Sukanya, C.M.; Gokul, R.; Paul, V. A survey on object recognition methods. *Int. J. Sci. Eng. Comput. Technol.* **2016**, *6*, 48.
4. Nguyen, T.T.; Grabner, H.; Bischof, H.; Gruber, B. On-Line Boosting for Car Detection from Aerial Images. In Proceedings of the 2007 IEEE International Conference on Research, Innovation and Vision for the Future, Hanoi, Vietnam, 5–9 March 2007; pp. 87–95.

5. Cao, X.; Wu, C.; Lan, J.; Yan, P.; Li, X. Vehicle detection and motion analysis in low-altitude airborne video under urban environment. *IEEE Trans. Circuits Syst. Video Technol.* **2011**, *21*, 1522–1533. [CrossRef]
6. Yu, C.; Jiang, X.; Wu, F.; Fu, Y.; Zhang, Y.; Li, X.; Fu, T.; Pei, J. Research on Vehicle Detection in Infrared Aerial Images in Complex Urban and Road Backgrounds. *Electronics* **2024**, *13*, 319. [CrossRef]
7. Kuma, D.; Sinha, B. Vehicle Detection in Aerial Images: A Survey. In *Data Science and Communication*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 145–158.
8. Ali, S.; Jalal, A. Vehicle Detection and Tracking from Aerial Imagery via YOLO and Centroid Tracking. In Proceedings of the Conference ICACS'23, Larissa, Greece, 19–21 October 2023.
9. Makrigiorgis, R.; Kyrkou, C.; Kolios, P. How High Can You Detect? Improved Accuracy and Efficiency at Varying Altitudes for Aerial Vehicle Detection. In Proceedings of the 2023 International Conference on Unmanned Aircraft Systems (ICUAS), Warsaw, Poland, 6–9 June 2023.
10. Wu, T.H.; Wang, T.W.; Liu, Y.Q. Real-Time Vehicle and Distance Detection Based on Improved Yolo v5 Network. In Proceedings of the 2021 3rd World Symposium on Artificial Intelligence (WSAI), Guangzhou, China, 18–20 June 2021; pp. 24–28.
11. Tang, H.; Liang, S.; Yao, D.; Qiao, Y. A visual defect detection for optics lens based on the YOLOv5-C3CA-SPPF network model. *Opt. Express* **2023**, *31*, 2628–2643. [CrossRef] [PubMed]
12. Zhang, T.; Zhang, X.; Ke, X. Quad-FPN: A novel quad feature pyramid network for SAR ship detection. *Remote Sens.* **2021**, *13*, 2771. [CrossRef]
13. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768. [CrossRef]
14. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
15. Liu, S.; Huang, D. Receptive Field Block Net for Accurate and Fast Object Detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 385–400.
16. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the 15th European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 3–19. [CrossRef]
17. Tan, M.X.; Pang, R.M.; Le, Q.V. EfficientDet: Scalable and Efficient Object Detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
18. Sun, Y.; Cao, B.; Zhu, P.; Hu, Q. Drone-Based RGB-Infrared Cross-Modality Vehicle Detection via Uncertainty-Aware Learning. *Trans. Circuits Syst. Video Technol.* **2022**, *32*, 6700–6713. [CrossRef]
19. Glas, A.S.; Lijmer, J.G.; Prins, M.H.; Bonsel, G.J.; Bossuyt, P.M. The diagnostic odds ratio: A single indicator of test performance. *J. Clin. Epidemiol.* **2003**, *56*, 1129–1135. [CrossRef] [PubMed]
20. Buckland, M.; Gey, F. The relationship between recall and precision. *J. Am. Soc. Inf. Sci.* **1994**, *45*, 12–19. [CrossRef]
21. Yacouby, R.; Axman, D. Probabilistic Extension of Precision, Recall, and f1 Score for More thorough Evaluation of Classification Models. In Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems, Online, 20 November 2020; pp. 79–91.