

Article

Small Target Detection Algorithm Based on Improved YOLOv5

Ruiyun Chen ¹, Zhonghua Liu ^{1,*} , Weihua Ou ² and Kaibing Zhang ³

¹ School of Information Engineering, Zhejiang Ocean University, Zhoushan 316022, China; 18135360845@163.com

² School of Big Data and Computer Science, Guizhou Normal University, Guiyang 550025, China; ouweihua@gznu.edu.cn

³ College of Electronics and Information, Xi'an Polytechnic University, Xi'an 710048, China; kbzhang0505@gmail.com

* Correspondence: lzhua_217@163.com

Abstract: Small targets exist in large numbers in various fields. They are broadly used in aerospace, video monitoring, and industrial detection. However, because of its tiny dimensions and modest resolution, the precision of small-target detection is low, and the erroneous detection rate is high. Therefore, based on YOLOv5, an improved small-target detection model is proposed. First, in order to improve the number of tiny targets detected while enhancing small-target detection performance, an additional detection head is added. Second, involution is used between the backbone and neck to increase the channel information of feature mapping. Third, the model introduces the BiFormer, wherein both the global and local feature information are captured simultaneously by means of its double-layer routing attention mechanism. Finally, a context augmentation module (CAM) is inserted into the neck in order to maximize the structure of feature fusion. In addition, in order to consider among the required real frame as well as the prediction frame simultaneously, YOLOv5's original loss function is exchanged. The experimental results using the public dataset VisDrone2019 show that the proposed model has P increased by 13.43%, R increased by 11.28%, and mAP@.5 and mAP@[.5:.95] increased by 13.88% and 9.01%, respectively.

Keywords: small-target detection; YOLOv5; BiFormer; CAM



Citation: Chen, R.; Liu, Z.; Ou, W.; Zhang, K. Small Target Detection Algorithm Based on Improved YOLOv5. *Electronics* **2024**, *13*, 4158. <https://doi.org/10.3390/electronics13214158>

Received: 22 September 2024

Revised: 20 October 2024

Accepted: 21 October 2024

Published: 23 October 2024



Copyright: © 2024 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

Target detection primarily identifies the target of interest in a picture and determines its category and location. It is an important field of computer vision research that serves as the basis for increasingly difficult vision tasks. In recent years, with the increasing diversification and enrichment of demands in fields such as drone reconnaissance, intelligent video surveillance, industrial inspection, and medical diagnosis, the research on target detection has become particularly important. The widespread application of deep-learning technology has led to the rapid development of target detection technology, and the current detection model has achieved good results for large- and medium-sized targets. However, the size of a small target in the image is less than 32×32 pixels, or even smaller. These targets are challenging to detect because they occupy limited space in the image or video; have little detailed information; and are susceptible to resolution, noise, and background interference. Small-target detection has consequently emerged as one of computer vision's most difficult problems.

A lot of research and improvement has been performed to improve the performance of small-target detection in order to address these issues. An enhanced small-target detection technique is presented by Yang et al. [1]. The K-means++ clustering technique is first applied in the suggested strategy to increase the precision of the anchor dimension of small targets. Then, to focus more on the small targets' feature information, the scSE (spatial and channel compression and excitation) attention module is included. Ultimately,

the effectiveness of small-target detection is enhanced by the addition of a small-target detection layer. TCA-YOLOv5m is a new tiny target detection model that Huang et al. [2] propose. First, to allow the model to gain more image feature information, the YOLOv5 backbone network integrates the Transformer algorithm in the end. Second, the Path Aggregation Network (PANet) and the Transformer method are coupled in the neck layer of the model to increase the feature pyramid's expression capability and the accuracy of detecting occluded small targets with high densities. Furthermore, in high-density settings, the CA algorithm is used to improve target location accuracy. Last, the capacity to gather small targets is enhanced by the addition of a detecting layer. In order to maintain large-scale feature information, Cao et al. [3] introduce a multiple-scale feature fusion module and suggest a new connection approach that combines a multi-scale module with an attention module. Nevertheless, there is still much room for improvement in the small-object detection performance, because the above methods still have problems such as low detection accuracy and high computational complexity.

This paper proposes an improved YOLOv5 small-target detection model. First, in order to catch more small targets and enhance the effectiveness of small-target detection, an extra detection head is added. Second, to enhance the channel information of the feature mapping, involution is inserted between the neck and the trunk. Furthermore, the neck is where the BiFormer attention mechanism is used, enabling the simultaneous acquisition of global and local feature information via its double-layer routing attention method. In order to maximize the neck reticulum's feature-fusion structure, CAM is finally inserted into the neck.

The following is a summary of this paper's primary contributions:

- (1) A fourth detection head is added to YOLOv5 to extract information from more fine-grained features. Involution is added to adaptively learn the weights in the spatial dimension, and the features of small targets in space and channels can be adaptively enhanced.
- (2) The BiFormer attention module and the context augmentation module (CAM) are introduced in the neck of YOLOv5. BiFormer enables the fusion of multi-scale features through bidirectional information flow and captures the information of targets at different scales at the same time, which is very beneficial for situations that need to detect large and small targets at the same time.
- (3) On the basis of the traditional IoU, the SIoU loss function introduces the center-point distance penalty term. This makes the model consider the distance between the center point of the detection frame and the real frame when calculating the loss, which is helpful for the model to optimize the positioning of the detection frame. SIoU can still provide gradient information without overlap, so the model can obtain effective feedback faster during initial training, and the helper box can quickly converge to the target.

The structure of the remaining sections of this paper is as follows: Section 2 discusses the work as well as the YOLOv5 network architecture overview, Section 3 details the modules covered in this paper, and Section 4 covers our experimental setup and analysis of the results. Finally, Section 5 concludes the paper.

2. Related Works

Determining the category and location information of individual targets in an image is the fundamental problem of target detection, which forms the foundation for advanced visual tasks like target tracking, real-time segmentation, and image description. Like the majority of models in machine vision, object detection models have evolved from the conventional method of using shallow classifiers and artificially created features for identification to using convolutional neural networks to learn features deeply.

Sliding window technology and manually created features are used in traditional target-detection algorithms for feature extraction and candidate region construction [4]. The Histogram of Oriented Gradient (HOG) [5], Scale-Invariant Feature Transform (SIFT) [6],

Haar-like features (Haar) [7], and Local Binary Pattern (LBP) [8] are examples of traditional feature-extraction techniques. Support Vector Machine (SVM) [9], AdaBoost [10], and deformation part model DPM [9,10] are examples of traditional classifiers. The disadvantages of traditional target-detection techniques are first, that there is no selection method for sliding windows; second, that there are redundant windows and high time complexity; and finally, that the manually designed features lack robustness. Therefore, the detection performance of traditional target-detection technology is not optimal for small-target detection tasks in complex scenes.

AlexNet, which was created in 2012 by Hinton et al. [11] using a convolutional neural network (CNN), outperformed conventional target-detection models in terms of detection performance and went on to win the ImageNet Large-Scale Visual Recognition Challenge. A new era of deep learning in target detection has been brought about by the success of AlexNet, which has elevated CNNs to the status of the most crucial tool in deep-learning models [12].

One-stage and two-stage target detection algorithms are the two main groups into which deep-learning-based algorithms fall. The former primarily consists of region-based convolutional neural networks, which are represented by the R-CNN series [13–15], while the latter primarily consists of regression-based techniques, which are represented by the YOLO series [16–18].

The following is a description of how CNN finds targets. First, it creates a number of candidate regions with potential targets. Next, it extracts the properties of each candidate region separately, utilizing a convolutional neural network. So as to ascertain the target's category, it then forwards the information to the SVM classifier. Last, the regressor is used to change the location of the candidate box. Despite the significant increase in accuracy over conventional object-detection algorithms, R-CNN still has the following two shortcomings: One is that the created candidate zones will overlap, and the convolution process will involve a lot of repetitive calculations. This can lead to a lengthy training period, high space costs, and sluggish detection speed. The other is that, in order to extract the features, the candidate region must first be scaled to the same size, which could quickly result in the loss of picture data.

In 2016, Joseph Redmon et al. [18] came up with the YOLO (You Only Look Once) target detection technique, which was a continuation of the R-CNN series. The main idea behind that is to make use of a single neural network to locate and categorize targets concurrently, turning the target-detection challenge into an issue of regression and achieving efficient and real-time target detection. Liu et al. introduced a one-stage target detection technique for the Single Shot Multibox Detector (SSD) [19] in 2016, in addition to the YOLO series algorithms. YOLOv3 was introduced by Joseph Redmon et al. in 2018 [20]. To extract picture features, YOLOv3 employs a deeper DarkNet-53. The efficacy of small-target detection is improved by the model's capacity to identify targets at various feature map scales. The year 2020 saw the simultaneous appearance of YOLOv4 [21] and YOLOv5. YOLOv4 uses CSPDarkNet53 to replace the backbone network. The PAN (Path Aggregation Network) [22] module is added to the FPN (Feature Pyramid Networks) module of YOLOv3 [23] to better integrate the feature map data at various scales, and SPP (Spatial Pyramid Pooling) [24] is imported in the neck section to increase the receptive field. Furthermore, YOLOv4 incorporates a variety of optimization techniques, maintaining precision of detection while significantly enhancing the speed of detection. An improved version of YOLOv4 is called YOLOv5, which has a tiny memory footprint, is fast, and is highly suited for real-time applications. YOLOv5-6.0 is the experimental algorithm chosen for this paper. Figure 1 displays the YOLOv5-6.0 structure.

The three essential components of YOLOv5 are the head, neck, and backbone. The backbone's main job is to gather features from incoming images. Multi-scale feature fusion on the feature map and feature transfer to the prediction layer are handled by the neck. The final regression prediction is made by the head.

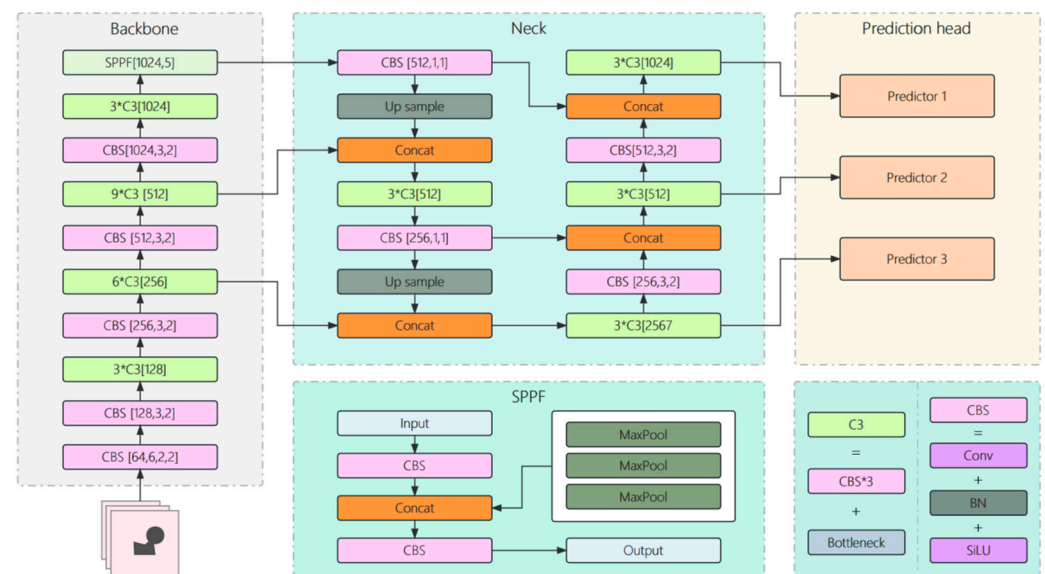


Figure 1. Structure of YOLOv5-6.0.

The backbone network's primary job is feature extraction and ongoing feature-map reduction. The Conv, C3, and SPPF modules are the three primary modules that make up the backbone. An activation function, a BatchNorm2d, and a Conv2d make up the Conv module. The Conv module is capable of extracting features; organizing feature maps; and finishing feature-map nonlinearity, normalizing, downsampling, and dimension enhancement and reduction. Figure 2 depicts the Conv module's structure. Three Conv modules and one bottleneck module make up the C3 module. The most important module in the backbone for feature extraction is C3. Figure 3 depicts the C3 module's structure. Yolov5 improves SPPF based on spatial pyramid pooling (SPP), which produces the same results more quickly. The feature map with an uneven scale can be transformed into a single scale using SPP. Yolov5, on the other hand, uses padding to cushion SPP and SPPF. The size of the feature maps in the front and back are actually equal because they are pooled together. Thus, in Yolov5, the main purpose of SPPF is to aggregate features from various scales. The same feature map's feature representations at various scales are concatenated together. Figure 4a depicts the SPP module's structure, while Figure 4b shows the SPPF module's structure.

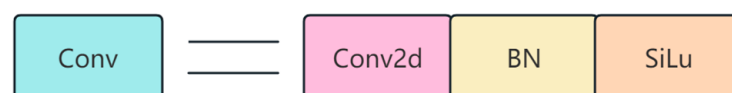


Figure 2. Conv structure.

In reality, the neck constructs a feature pyramid (FPN) by concatenating deeper semantic features with relatively shallow features from the backbone. By sampling from the left upsampling, the feature map is interpolated, allowing for a wider feature map scale. As a result, the feature map continues to increase and is easy to fuse upward toward the backbone, and it keeps growing. Downsampling is still being performed on the right side of the neck, with the aim of obtaining feature maps of various sizes and improving the integration of the shallow graphic features and the deep semantic features, rather than merely basic connections. In order to gain more complete features, the neck layer combines the deep semantic and the shallow graphic features.

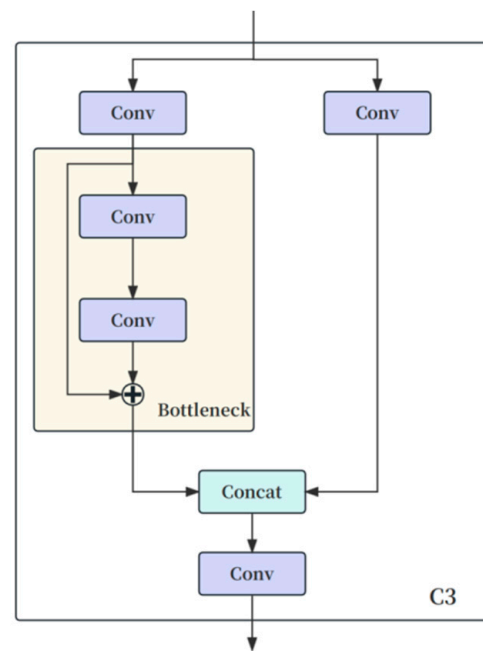


Figure 3. The C3 structure.

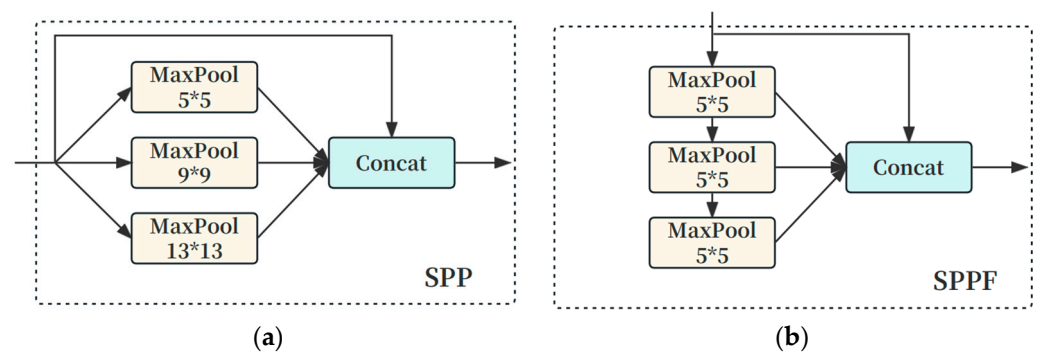


Figure 4. The SPP (a) and SPPF (b) structures.

The three detectors that make up the head's main body use grid-based anchoring to identify objects at various scales on feature maps. Three 1×1 convolutions, or three detection feature layers of various scales, make up the detection module's extremely basic network topology. Ultimately, three distinct feature maps with varying scales of 80×80 , 40×40 , and 20×20 are produced and can be utilized for forecasting big, medium, and small targets, respectively.

3. Methodology

YOLO has consistently been a breakthrough algorithm in the domain of target detection. Since the advent of the YOLO algorithm, it has evolved into many versions, among which is YOLOv5, which uses the Pytorch framework. YOLOv5 is the preferred option for real-time applications due to its several advantages, including higher detection accuracy, faster detection speed, a more customizable model structure, and ease of training and deployment. These features enable it to better fulfill the demands of practical applications. In this paper, YOLOv5-6.0 was chosen as the fundamental framework, and Figure 5 shows the structure of the proposed model.

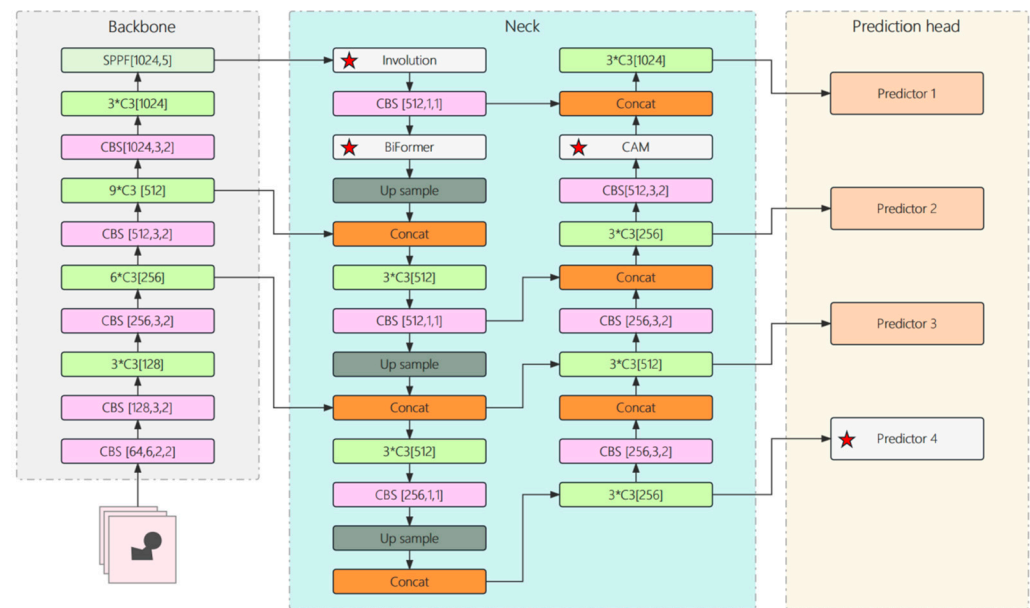


Figure 5. The framework structure of the proposed model. The red star indicates the improvements made on YOLOv5 in this paper.

3.1. BiFormer

A visual Transformer model called BiFormer combines BHlevel Routing Attention, whose core idea is to introduce a two-layer routing attention mechanism. In BiFormer, each image block is associated with a location router, which can assign image blocks to upper and lower routers based on specific rules. The lower-layer router is in charge of gathering local area specifics, and the upper-layer router is in charge of gathering global contextual data. Specifically, the upper-layer router interacts with all image blocks through a global self-attention mechanism and generates a global image representation. The lower-level router uses a local self-attention mechanism to interact with each image tile and its neighbors and generates a local image representation. BiFormer can gather both global and local feature information by using the attention-based two-layer routing approach, which enhances the model's performance in vision tasks. Inspired by [25], we implemented the BiFormer attention mechanism in YOLOv5's neck region, which captures both global and local feature information through its double-layer routing attention mechanism and captures more small targets.

3.2. Context Augmentation Module (CAM)

The target in target detection is often obscured by other targets or obstructions, which hinders the model's ability to identify and precisely locate the target. It is challenging to detect them, particularly in complicated settings captured by photos from airborne monitoring, due to small size, low resolution, and the target's intricate context. Many target detectors attempt to superimpose more and more pooling and downsampling procedures in attempts to obtain trustworthy semantic information. However, this gradually reduces the detection performance of small targets by erasing the features of targets with fewer pixels in forward propagation. A feature pyramid network with a composite structure that includes a feature refinement module and a context enhancement module is proposed by [26] et al. To strengthen the context data of small targets, the context enhancement module integrates extended convolution to collect context data from various receiving domains and incorporates it into the FPN. Highlighting significant traits and stifling competing data from the channel and spatial dimensions, the feature refinement module combines channel adaptive fusion and spatial adaptive fusion. Furthermore, our model incorporates the Copy-Reduce-Paste data augmentation strategy to eliminate the training imbalance. Figure 6 depicts the three CAM structures.

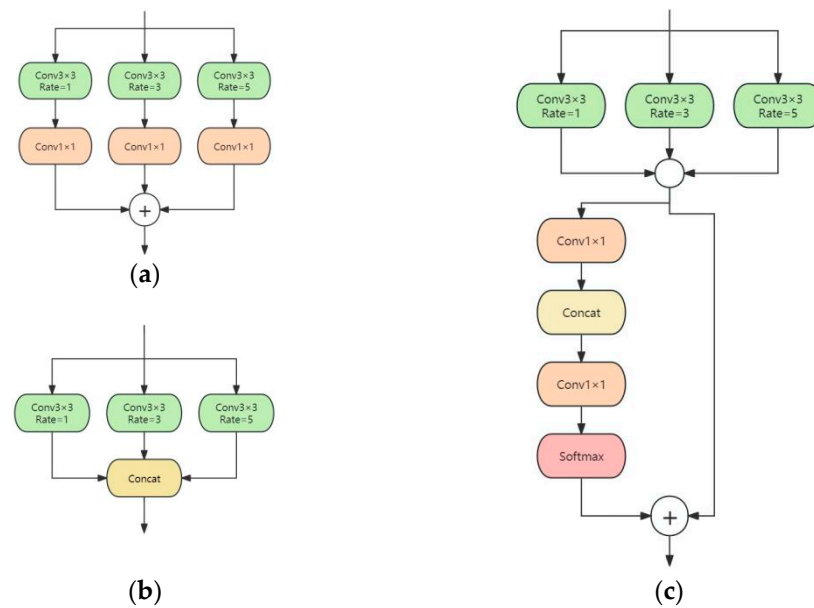


Figure 6. The different feature-fusion methods: (a) Weighted Fusion, (b) Concatenation Fusion, and (c) Adaptive Fusion.

Simultaneously, to minimize the excessive number of parameters during training, one-quarter of the feature map's channels were used as convolution kernels. This allowed the number of channels to be expanded back to its original amount using (1×1) convolution, resulting in three feature maps of equal size but distinct receptive fields. The difference is in the way features are blended. Figure 6a shows the CAM-W module using weighted feature fusion, Figure 6b shows the CAM-C module using cascade operation fusion, and Figure 6c shows the CAM-A module using adaptive feature fusion, which is similar to the implementation principle of the SE attention mechanism. Specifically, the CAM module obtains the adaptive weights of the input channels through convolution operation and Softmax operation, and the three channels correspond to three outputs, and the contextual information is aggregated into the output feature map through calculation weighting.

3.3. Channel Feature Fusion with Involution (CFFI)

YOLOv5 reduces the number of channels at the neck's beginning by using a 1×1 convolution, which significantly improves computational efficiency and reduces channel information. Inspired by this [27], we added an involution block between the backbone and neck to reduce the loss of information by sharing channel information. In Figure 7 [28], the involutory structure is displayed. The information contained in the channel dimension of a single pixel is implicitly dispersed in the vicinity of its space, which helps to obtain rich receptive-field information.

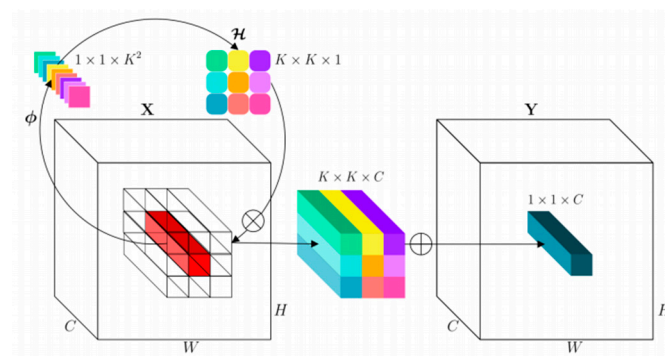


Figure 7. Structure of the involution block.

3.4. Prediction Head

YOLOv5-6.0 has three detection layers. The neck network executed 8-, 16-, and 32-folds of downsampling when the 640×640 picture size was used as the input size. The resulting prediction-layer feature map sizes were 80×80 , 40×40 , and 20×20 and were employed to identify small, medium, and large objects in turn. To enhance the precision of small-target detection, we supplemented the initial YOLOv5-6.0 network with a 160×160 prediction layer.

3.5. Loss Function

The loss function's job is to quantify the difference between the neural network's expected and forecasted data. The loss function's value diminishes as the predicted and expected information become closer to each other. The loss function of YOLO mainly consists of three parts: bounding box loss, confidence loss, and class loss. YOLO optimizes the difference between each prediction box and the real box as a loss. This loss is divided into position loss (mainly including the coordinates of the center point) and size loss (width and height). The confidence loss indicates whether there are targets in each prediction box and the degree of overlap (IoU) between the prediction box and the real box. The confidence loss indicates whether there are targets in each prediction box and the degree of overlap (IoU) between the prediction box and the real box. The total loss function is expressed as follows:

$$\text{Loss}_{\text{total}} = a * \text{Loss}_{\text{bbox}} + \text{Loss}_{\text{conf}} + \text{Loss}_{\text{class}} \quad (1)$$

where a is a weight used to balance the loss of the bounding box. $\text{Loss}_{\text{bbox}}$ represents bounding box loss, $\text{Loss}_{\text{conf}}$ represents the confidence loss, and $\text{Loss}_{\text{class}}$ represents the category loss.

By optimizing the loss function, the YOLO model can optimize the positioning, target confidence, and the accuracy of category classification at the same time during the detection process so as to achieve efficient target detection.

Inspired by [29], we modified the loss function to the SIOU loss function to help the model predict the target's position and size more accurately. This is because previous approaches, such as IoU [30], GIoU [31], DIoU [32], CIoU [33], Focal EIou [34], et al., did not account for the mismatch between the prediction box and the necessary real box, resulting in a slow rate of convergence and low efficiency. Conversely, SIOU redefines the penalty measures while accounting for the vector angles among the necessary regressions.

4. Experiments

YOLOv5 has four network models with different depth_multiple and width_multiple: Yolov5s, Yolov5m, Yolov5l, and Yolov5x. Yolov5s is the network with the least feature map width and depth. Based on this, the last three are all expanding and deepening. This research experimented with small-target detection performance using YOLOv5m as the foundational model.

This article's experiments were conducted using an RTX 3090 (24 GB) GPU and an Intel(R) Xeon(R) Platinum 8255C CPU running at 2.50 GHz. Python 3.8 was the programming language used in the experiment, together with PyCharm, Anaconda 3, PyTorch 1.11.0, and Cuda 11.3 deep-learning frameworks.

4.1. Dataset

In an experiment, the VisDrone2019 dataset was used to validate different target detection models, including the proposed framework and the comparison methods. The AISKEYEYE team at Tianjin University's Machine Learning and Data Mining Laboratory gathered the dataset. A total of 288 video clips with 261,908 frames and 10,209 still photos taken with different drone cameras were included in the benchmark dataset. These clips covered a wide range of topics, such as location, environment, targets, and density. Five

primary targets comprised the VisDrone2019 dataset's challenges: target detection in photos, target detection in videos, single-target tracking, multi-object tracking, and crowd counting. This paper used target detection in photos, which consisted of 10,209 still images taken at various altitudes and locations by the UAV platform, divided into 10 categories: pedestrian, people, bicycle, car, van, truck, tricycle, awning-tricycle, bus, and motor.

4.2. Evaluation Criterion

The quality of a target-detection model can be evaluated from the following aspects:

- (1) Accuracy: speaks of the percentage of targets that the model detects accurately.

$$\text{acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{FN}} \quad (2)$$

When the actual and forecast are both positive, as indicated by TP; when the actual and forecast are both negative, as indicated by FP; when the actual and forecast are both positive, as indicated by FP; FN denotes a negative case for the actual and a negative case for the predicted.

- (2) Precision: speaks of the percentage of samples that are targets in reality that the model predicts to be targets.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (3)$$

- (3) Recall: the percentage of the sample of items that the model has detected as targets.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (4)$$

- (4) mAP (Mean Mean Accuracy): The area of the P-R curve for a particular class is called AP, and the average value of APs for each class of APs is in mAP. The mAP's size falls between 0 and 1. The accuracy of the model increases with the size of the mAP.

$$\text{mAP} = \frac{\sum_{i=1}^C \text{AP}_i}{C} \quad (5)$$

- (5) IoU: This is commonly used to gauge how well targets are localized in target detection algorithms. Its definition is the ratio of the detection frame's union area to the intersection area of the real frame and the detection frame. The accuracy of the detection result increases with the IoU value.

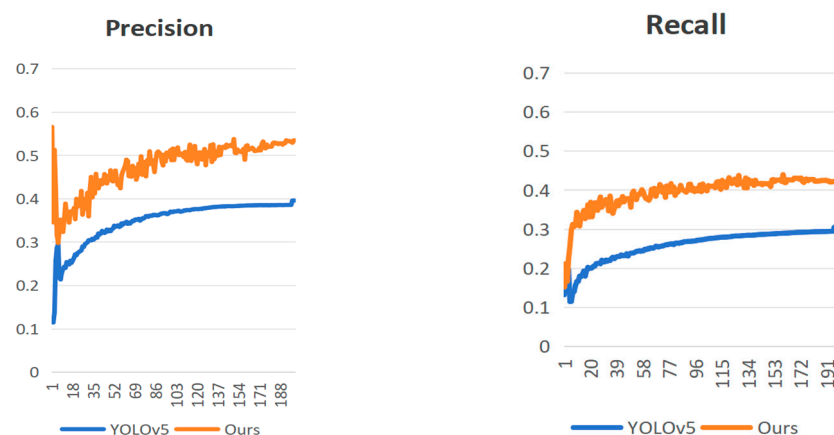
4.3. Experimental Results

We tested the improved model using the VisDrone2019 dataset to confirm its impact on the small-target detection performance. Table 1 shows the improved model's results in comparison with previous research and YOLOv5. On the VisDrone2019 dataset, we discovered that the improved model greatly enhanced the performance of small-target detection, with P rising by 13.43%, R by 11.28%, mAP@.5 by 13.88%, and mAP@[.5:.95] by 9.01%.

As shown in Figure 8, we selected one image each from daytime and nighttime in the VisDrone-2019 validation set to verify the improved model's ability to detect small targets. Figure 8a shows the visual effect of detection using YOLOv5, and Figure 8b shows the visual effect of detection using the improved model. From the figure, we can see that the improved model detected targets that YOLOv5 misdetected and missed, as well as more small targets. The experiment results show that the improved model's ability to detect small targets is more evident than YOLOv5's. Figure 9 compares the precision and recall of the newly improved model in this study to that of YOLOv5, and it is clear that the new model has significantly higher precision and recall than YOLOv5.

Table 1. Comparison of different algorithms.

Method	Precision	Recall	mAP@.5	mAP@[.5:.95]
YOLOv5	39.89%	30.87%	29.97%	17.13%
Tang et al. [27]	47.88%	38.49%	36.95%	20.85%
Yang et al. [1]	48.33%	40.13%	39.28%	23.57%
Cao et al. [3]	46.92%	37.44%	35.73%	20.35%
Huang et al. [2]	45.21%	35.62%	33.18%	18.98%
Ours	53.32%	42.15%	43.85%	26.14%

**Figure 8.** Comparison of the different models' detection effects: (a) YOLOv5 and (b) the proposed model.**Figure 9.** Comparison of the P and R of different models.

As seen above in Figure 10, Figure 10a displays the data for the 10 categories of the training set in the experimental dataset VisDrone2019. A confusion matrix diagram is shown in Figure 10c, where a class's actual instances are represented by each row, and class instance predictions are represented by each column. The region enclosed under the PR_curve is shown in Figure 10b. The average of all forms of AP is mAP, and the

PR curve is AP. It is evident that the R decreases as the P increases. The wider the area contained by the PR curve, the greater the AP value, the more accurate this category is, and the closer the curve is to the (1,1) point in which it is placed. Figure 10b illustrates that the AWNING-CYCLE category has the lowest accuracy, while the car category has the best accuracy. Combined with Figure 10c, we can see that a large part of the awning-tricycle is misjudged as car and tricycle, resulting in the low accuracy of awning-tricycle.

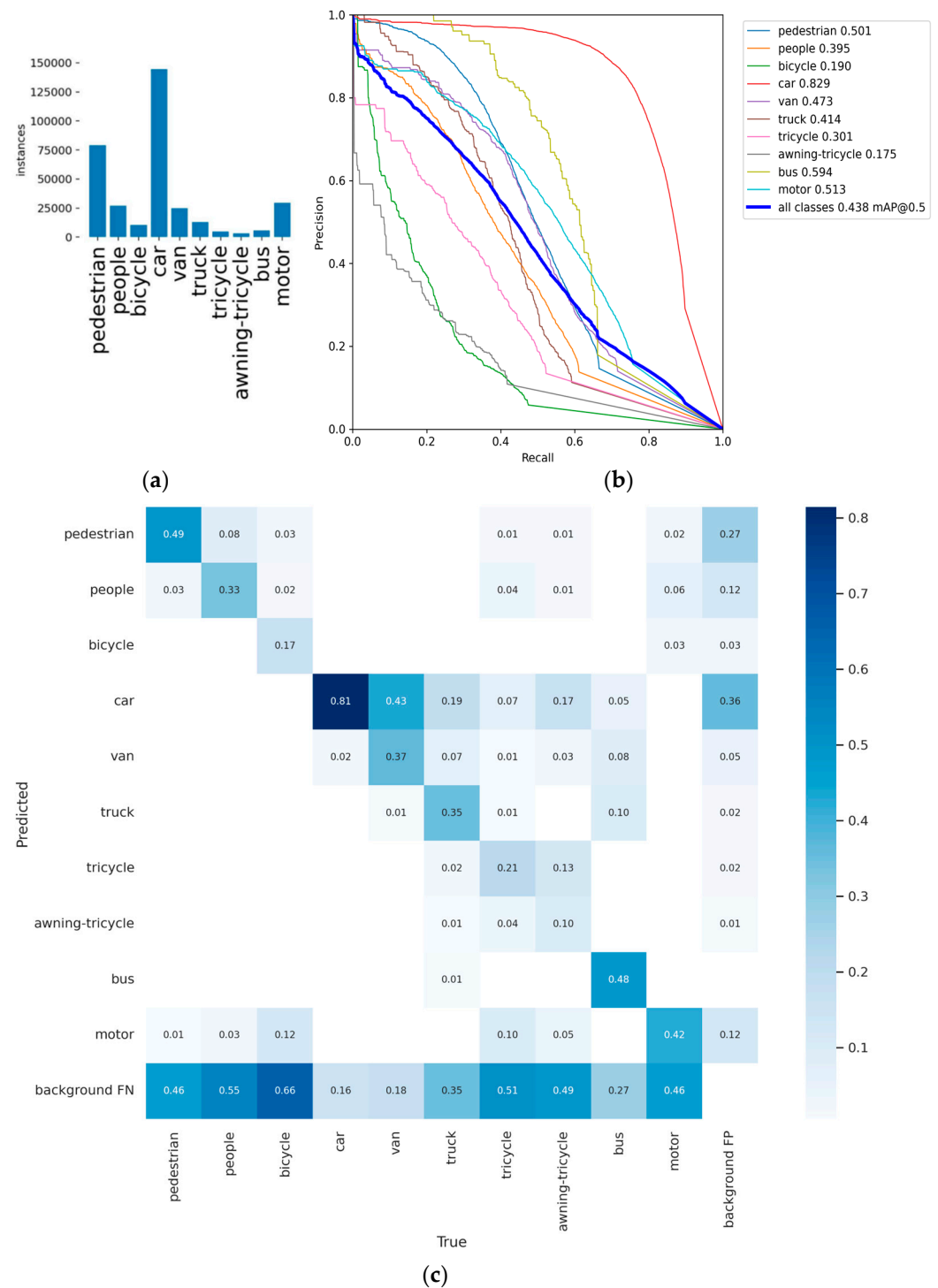


Figure 10. Instance distribution of VisDrone2019 dataset: (a) instances, (b) PR_curve, and (c) confusion matrix.

4.4. Ablation Study

We performed some ablation experiments, as Table 2 and Figure 11 illustrate. It can be seen from Table 2 that these changes, including adding a detection head for small targets, an involution block, the BiFormer attention mechanism, and a CAM context enhancement module, and modifying the loss function to SIoU, are the most effective detecting method for small targets. From Figure 11, we are able to compare the results of several enhanced methods. The performance of the improved technique is better than that of the other improved methods, which is more favorable for the detection of small targets. It is higher than other improved methods in P, R, mAP@.5, and mAP@[.5:.95].

Table 2. Comparison of various variants on the VisDrone2019 dataset.

Model	Precision	Recall	mAP@.5	mAP@[.5:.95]
YOLOv5	39.89%	30.87%	29.97%	17.13%
+Head	44.52%	34.87%	33.81%	18.38%
+Involution	43.59%	33.9%	33.1%	17.69%
+BiFormer	45.09%	34.24%	33.56%	18.03%
+CAM	47.69%	40.07%	38.57%	22.37%
+SIoU	41.33%	32.7%	30.35%	17.26%
+Head+Involution+BiFormer	50.31%	42.71%	42.58%	24.78%
+Head+Involution+BiFormer+SIoU	50.88%	42.62%	42.82%	24.91%
+Head+Involution+BiFormer+SIoU+CAM(Ours)	53.32%	42.15%	43.85%	26.14%

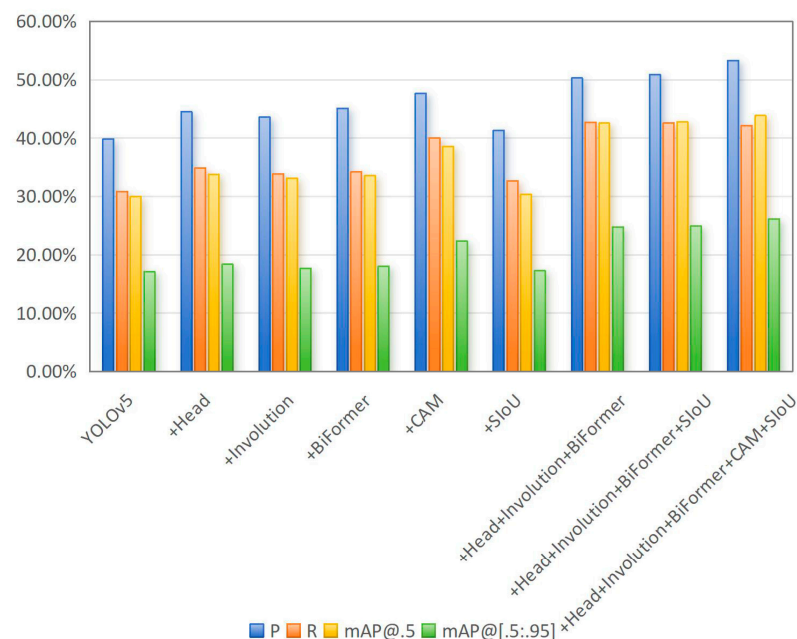


Figure 11. Comparison of the effects of the different improvement methods.

5. Conclusions

In order to solve the problems of low resolution, high density, complex background, high positioning accuracy, and susceptibility to environmental noise in small targets, this paper proposes an improved model based on YOLOv5. First, a detection head specifically meant for small targets is added to extract the features of small targets at a higher resolution. Second, an involution block is added between the neck and the trunk to reduce the disappearance of channel data. Third, in order to maximize the feature fusion structure of the

neck network and to gather additional feature data, CAM and BiFormer are added to the neck part. Finally, SIOU is used to replace the original loss function CIOU of YOLOv5. SIOU fully considers the problem of mismatch between the required real box and the predicted box. The experimental results demonstrate that when compared to the YOLOv5 algorithm, the proposed model has P increased by 13.43%, R increased by 11.28%, and mAP@.5 and mAP@[.5:.95] increased by 13.88% and 9.01%, respectively, and the detection effect of small targets was significantly enhanced.

The small-target detection model proposed in this paper has a wide range of applications in the fields of intelligent video surveillance, autonomous driving, industrial detection, medical image analysis, etc., where the target is usually small in scale and difficult to be accurately detected by traditional detection methods. In the field of intelligent video surveillance, precision agriculture management is achieved by monitoring the distribution of crops in large areas of farmland. In the field of autonomous driving, it is possible to accurately identify traffic signs and obstacles that are small or partially obscured in front of you, improving the safety of autonomous driving. In the field of industrial inspection, it can be used to monitor and detect small defects on the production line or in equipment to improve product quality and production efficiency. In the field of medical impact analysis, it can help to detect small tumors (such as lung nodules, small tumors in the brain, etc.) at early stages in CT or MRI images for early diagnoses.

Although the improved model proposed in this paper has achieved satisfactory results on the VisDrone2019 dataset and other small target datasets, there is still room for improvement in computational complexity and energy consumption. Future research work will be devoted to designing lightweight neural network models in the hope that, by reducing network parameters and calculation volume, the model will require fewer computing resources during inference, run faster, and be better applied to real-time applications. In addition, its low memory footprint will also significantly reduce the model's energy consumption, making it more suitable for devices that require long battery life.

Author Contributions: Methodology, R.C.; Investigation, R.C.; Writing—original draft, R.C.; Supervision, Z.L., W.O. and K.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by NSFC of China (U1504610, 62262005, 61962010) and the High-Level Innovative Talents in Guizhou Province (No. GCC [2023]033).

Data Availability Statement: <https://github.com/VisDrone/VisDrone-Dataset> (accessed on 21 September 2024).

Conflicts of Interest: The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

1. Yang, R.; Li, W.; Shang, X.; Zhu, D.; Man, X. KPE-YOLOv5: An improved small target detection algorithm based on YOLOv5. *Electronics* **2023**, *12*, 817. [\[CrossRef\]](#)
2. Huang, M.; Zhang, Y.; Chen, Y. Small target detection model in aerial images based on TCA-YOLOv5m. *IEEE Access* **2022**, *11*, 3352–3366. [\[CrossRef\]](#)
3. Cao, S.; Wang, T.; Li, T.; Mao, Z. UAV small target detection algorithm based on an improved YOLOv5s model. *J. Vis. Commun. Image Represent.* **2023**, *97*, 103936. [\[CrossRef\]](#)
4. Liu, L.; Ouyang, W.; Wang, X.; Fieguth, P.; Chen, J.; Liu, X.; Pietikäinen, M. Deep learning for generic object detection: A survey. *Int. J. Comput. Vis.* **2020**, *128*, 261–318. [\[CrossRef\]](#)
5. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 886–893.
6. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [\[CrossRef\]](#)
7. Viola, P.; Jones, M. Rapid object detection using a boosted cascade of simple features. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Kauai, HI, USA, 8–14 December 2001; p. 1.
8. Ojala, T.; Pietikainen, M.; Maenpää, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 971–987. [\[CrossRef\]](#)
9. Pisner, D.A.; Schnyer, D.M. Support vector machine. In *Machine Learning*; Academic Press: Cambridge, MA, USA, 2020; pp. 101–121.

10. Freund, Y.; Schapire, R.E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* **1997**, *55*, 119–139. [\[CrossRef\]](#)
11. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**. [\[CrossRef\]](#)
12. Zou, Z.; Chen, K.; Shi, Z.; Guo, Y.; Ye, J. Object detection in 20 years: A survey. *Proc. IEEE* **2023**, *111*, 257–276. [\[CrossRef\]](#)
13. Xu, X.; Zhao, M.; Shi, P.; Ren, R.; He, X.; Wei, X.; Yang, H. Crack detection and comparison study based on faster R-CNN and mask R-CNN. *Sensors* **2022**, *22*, 1215. [\[CrossRef\]](#)
14. Zeng, L.; Sun, B.; Zhu, D. Underwater target detection based on Faster R-CNN and adversarial occlusion network. *Eng. Appl. Artif. Intell.* **2021**, *100*, 104190. [\[CrossRef\]](#)
15. Yi, Z.; Yao, D.; Li, G.; Ai, J.; Xie, W. Detection and localization for lake floating objects based on CA-faster R-CNN. *Multimed. Tools Appl.* **2022**, *81*, 17263–17281. [\[CrossRef\]](#)
16. Hussain, M. YOLO-v1 to YOLO-v8, the rise of YOLO and its complementary nature toward digital manufacturing and industrial defect detection. *Machines* **2023**, *11*, 677. [\[CrossRef\]](#)
17. Liu, Q.; Li, Y.; Dong, Q.; Ye, F. Scene-Specialized Multitarget Detector with an SMC-PHD Filter and a YOLO Network. *Comput. Intell. Neurosci.* **2022**, *2022*, 1010767. [\[CrossRef\]](#)
18. Chen, Y.; Xu, H.; Zhang, X.; Gao, P.; Xu, Z.; Huang, X. An object detection method for bayberry trees based on an improved YOLO algorithm. *Int. J. Digit. Earth* **2023**, *16*, 781–805. [\[CrossRef\]](#)
19. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016; pp. 21–37.
20. Abdusalomov, A.; Baratov, N.; Kutlimuratov, A.; Whangbo, T.K. An improvement of the fire detection and classification method using YOLOv3 for surveillance systems. *Sensors* **2021**, *21*, 6519. [\[CrossRef\]](#)
21. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
22. Yu, H.; Li, X.; Feng, Y.; Han, S. Multiple attentional path aggregation network for marine object detection. *Appl. Intell.* **2023**, *53*, 2434–2451. [\[CrossRef\]](#)
23. Xie, J.; Pang, Y.; Nie, J.; Cao, J.; Han, J. Latent feature pyramid network for object detection. *IEEE Trans. Multimed.* **2022**, *25*, 2153–2163. [\[CrossRef\]](#)
24. Lian, X.; Pang, Y.; Han, J.; Pan, J. Cascaded hierarchical atrous spatial pyramid pooling module for semantic segmentation. *Pattern Recognit.* **2021**, *110*, 107622. [\[CrossRef\]](#)
25. Zhu, L.; Wang, X.; Ke, Z.; Zhang, W.; Lau, R.W. Biformer: Vision transformer with bi-level routing attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 10323–10333.
26. Xiao, J.; Zhao, T.; Yao, Y.; Yu, Q.; Chen, Y. Context Augmentation and Feature Refinement Network for Tiny Object Detection. ICLR. 2022. Available online: <https://paperswithcode.com/paper/context-augmentation-and-feature-refinement> (accessed on 21 September 2024).
27. Tang, S.; Fang, Y.; Zhang, S. HIC-YOLOv5: Improved YOLOv5 For Small Object Detection. In Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA), Yokohama, Japan, 13–17 May 2023.
28. Li, D.; Hu, J.; Wang, C.; Li, X.; She, Q.; Zhu, L.; Zhang, T.; Chen, Q. Involution: Inverting the inheritance of convolution for visual recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 12321–12330.
29. Gevorgyan, Z. SloU loss: More powerful learning for bounding box regression. *arXiv* **2022**, arXiv:2205.12740.
30. Wu, S.; Yang, J.; Wang, X.; Li, X. Iou-balanced loss functions for single-stage object detection. *Pattern Recognit. Lett.* **2022**, *156*, 96–103. [\[CrossRef\]](#)
31. Qian, X.; Zhang, N.; Wang, W. Smooth giou loss for oriented object detection in remote sensing images. *Remote Sens.* **2023**, *15*, 1259. [\[CrossRef\]](#)
32. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU loss: Faster and better learning for bounding box regression. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 12993–13000. [\[CrossRef\]](#)
33. Xue, J.; Cheng, F.; Li, Y.; Song, Y.; Mao, T. Detection of farmland obstacles based on an improved YOLOv5s algorithm by using CIoU and anchor box scale clustering. *Sensors* **2022**, *22*, 1790. [\[CrossRef\]](#)
34. Zhang, Y.; Ren, W.; Zhang, Z.; Jia, Z.; Wang, L.; Tan, T. Focal and efficient IOU loss for accurate bounding box regression. *Neurocomputing* **2022**, *506*, 146–157. [\[CrossRef\]](#)

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.