ISSN 2616-5775 Vol. 8, Issue 2: 37-44, DOI: 10.25236/AJCIS.2025.080205

Research on Small Target Detection Algorithm Based on Improved YOLOv5

Yi Shi[a,*], Lei Ding[b], Shan Li[c], Xin Wang[d]

School of Electronic Information and Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an, Shaanxi, China
[a]lding@sust.edu.cn, [b]sy1235611@163.com
**Corresponding author*

Abstract: *In the field of computer vision, small target detection is extremely challenging due to the small size, low pixel density, and lack of contextual information of targets, making it difficult for traditional algorithms to effectively identify small targets because of frequent missed and false detections. Therefore, designing efficient and accurate detection algorithms is crucial. This paper proposes a novel small target detection method named YOLO_GF, which is based on the YOLOv5s framework. It introduces an improved DAMO-YOLO dense association memory mechanism to address the insufficient exchange of information between high- and low-level features during feature fusion. Furthermore, by integrating the improved efficient multi-scale attention EMA, the method utilizes the wide receptive field of parallel subnetworks to collect multi-scale spatial information and establish interdependencies among different spatial locations, thereby realizing a cross-spatial learning mechanism to enhance model accuracy. Additionally, the method employs a novel convolutional CSPStage to construct the feature fusion module, reducing model inference latency while improving detection accuracy. YOLO_GF achieves an end-to-end small target detection process and has been validated on the VisDrone2019 dataset, with experimental results demonstrating its high detection accuracy and good detection performance.*

Keywords: *Small Target Detection; DAMO-YOLO; Feature Fusion; Cross-Spatial Learning*

1. Introduction

In the field of computer vision, the detection of small targets stands as an immensely challenging issue. Owing to their diminutive size, low pixel density, and insufficiency of contextual information, the detection process is highly susceptible to environmental factors and the inherent characteristics of the targets themselves. This significantly diminishes the efficiency of target detection. Traditional target detection algorithms often fall short in effectively detecting small targets due to issues such as false positives and missed detections. Consequently, the design of an efficient and precise algorithm for small target detection has become a pressing necessity[1].

To address the aforementioned issues, this paper proposes a novel small object detection method termed YOLO_GF. This approach builds upon the fundamental framework of YOLOv5s and incorporates an enhanced densely-associated memory mechanism from DAMO-YOLO. It effectively resolves the insufficient exchange of information between high- and low-level features during the feature fusion process. Furthermore, an improved efficient multi-scale attention EMA is integrated, enabling cross-spatial learning to mitigate the issue of missing weak and small targets. The feature fusion module is constructed based on the novel convolutional CSPStage, which not only reduces the model inference latency but also enhances the detection accuracy[2]. This method achieves end-to-end small object detection and has been experimentally validated on the VisDrone2019 dataset, demonstrating high detection accuracy and satisfactory detection results.

To summarize, the main contributions of this paper are as follows:

In response to the issues of false detection and missed detection of small targets, this paper proposes an enhanced densely-associated memory mechanism based on DAMO-YOLO. It effectively addresses the insufficient exchange of information between high- and low-level features during the feature fusion process. Building upon this, an efficient multi-scale attention EMA is incorporated. By leveraging the large receptive field of parallel sub-networks, multi-scale spatial information is collected, and

ISSN 2616-5775 Vol. 8, Issue 2: 37-44, DOI: 10.25236/AJCIS.2025.080205

interdependencies are established among different spatial locations. This enables a cross-spatial learning mechanism, thereby enhancing the accuracy of the model.

The essence of single-stage object detection algorithms lies in simplifying the object detection task into an end-to-end classification and regression problem. These algorithms directly predict the category probability values and bounding box offsets in the image to obtain classification and regression results.In 2015, Joseph et al. proposed the original version of YOLO (You Only Look Once)[3], which achieved object detection by dividing the image into a fixed grid that predicts the probability of containing an object, the category, and the position information of the bounding box.In 2016, Liu et al. proposed the SSD (Single Shot MultiBox Detector) algorithm.[4] It predicts targets of different scales at various positions on the feature map and introduces multi-scale prior boxes to accomplish the object detection task.In 2017, Wu et al. proposed RetinaNet[5], which introduced Focal Loss to adjust the weights of positive and negative samples. This effectively addressed the issue of being easily affected by a large number of background categories during training and improved the performance of object detection[6].In 2020, Tan et al. proposed EfficientDet, using EfficientNet as the backbone network and presenting a Weighted Bi-directional Features Pyramid Network (BiFPN)[7]. Through composite scaling methods, it achieved better detection accuracy and smaller computational cost under resource-constrained conditions.In 2021, Chen et al. proposed YOLOF, which adopted an expanded encoder and unified matching, significantly enhancing the model's performance[8].In 2022, Alexey et al. proposed the YOLOv7 algorithm, which introduced more data augmentation methods, thereby improving the robustness of the model to some extent.Although such methods have a relatively fast detection speed, their detection accuracy is relatively low[9].

2. YOLO_GF Network Model

2.1. Overview of YOLO_GF

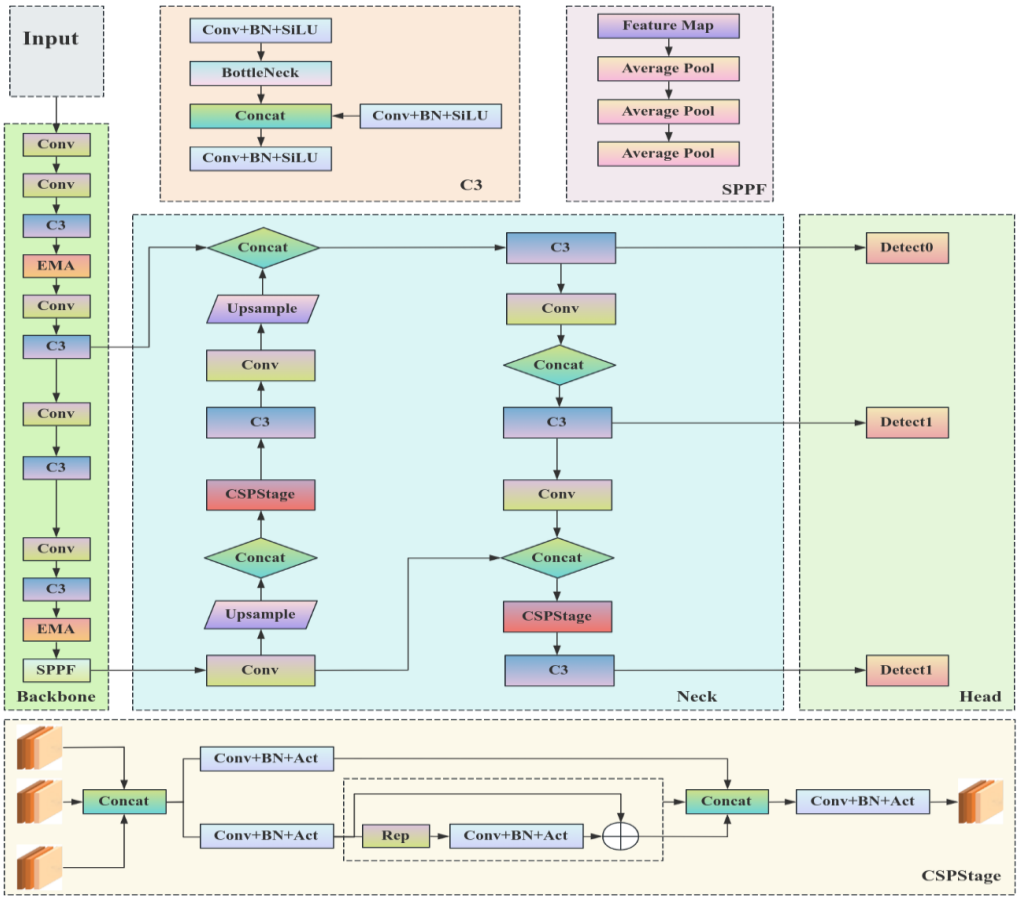

Figure 1: YOLO_GF overall algorithmic structure

ISSN 2616-5775 Vol. 8, Issue 2: 37-44, DOI: 10.25236/AJCIS.2025.080205

This paper proposes the YOLO_GF object detection algorithm, which addresses issues such as indistinct feature extraction due to numerous small target samples, complex backgrounds, false detections, and missed detections in object detection tasks. Firstly, this algorithm is based on the basic framework of YOLOv5s. On this basis, it incorporates an improved efficient multi-scale attention mechanism (EMA) to achieve cross-spatial learning and ameliorate the problem of missed detection of weak and small targets.Secondly, YOLO_GF constructs a feature fusion module based on the novel convolutional CSPStage and adopts an improved dense associative memory mechanism from DAMO-YOLO to resolve the insufficient exchange of information between high-level and low-level features during feature fusion in the model.The strategy proposed in this paper, considering multiple factors, makes YOLO_GF an efficient algorithm for real-time small object detection in dense images. The network model structure of YOLO_GF is mainly divided into four parts: Input, Backbone, Neck, and Head[10]. The overall algorithm structure is shown in Figure 1.

The Backbone section is composed of modules such as Conv, C3, and SPPF, which are responsible for extracting multi-scale deep feature representations B2, B3, B4, and B5 from the input image. The corresponding feature map sizes are R, 1/2R, 1/4R, and 1/8R respectively, where R is the product of the width and height of the feature map. The Neck section takes these as input and performs feature alignment and fusion, and then distributes them to the corresponding network levels. The multi-scale feature maps from the Backbone are input through C1, C2, C3, and then further fused locally by a novel convolution-based feature fusion module (CSPStage). Finally, the Head section regresses the position and class information of the target bounding box based on the fused feature representation output by the Neck section.

2.2 Improve the Dense Association Memory Mechanism of DAMO-YOLO

To address the difficulties associated with small object detection, this paper enhances the dense association memory mechanism of DAMO-YOLO, thereby improving issues related to false positives and missed detections of small objects.

Feature pyramid networks aim to aggregate features of different resolutions extracted from the backbone. Traditional FPN introduces a top-down pathway to fuse multi-scale features. Considering the limitations of unidirectional information flow, PAFPN adds an additional bottom-up pathway aggregation network, albeit with higher computational costs. BiFPN removes nodes with only one input edge and adds skip links among the original inputs at the same level. Employing Generalized-FPN (GFPN) as the neck and achieving SOTA performance enables full exchange of high-level semantic information and low-level spatial information. In GFPN, multi-scale feature fusion occurs among level features of previous and current layers, with log2(n) skip-layer connections providing more efficient information transmission, allowing for extension to deeper networks and thus higher accuracy. However, the latency of models based on GFPN is significantly higher than that of models based on the improved PANet[11].

Based on GFPN, this paper proposes a novel Efficient-RepGFPN to meet the design of real-time object detection, mainly including the following contents, the Efficient-RepGFPN architecture diagram is shown in Figure 2:

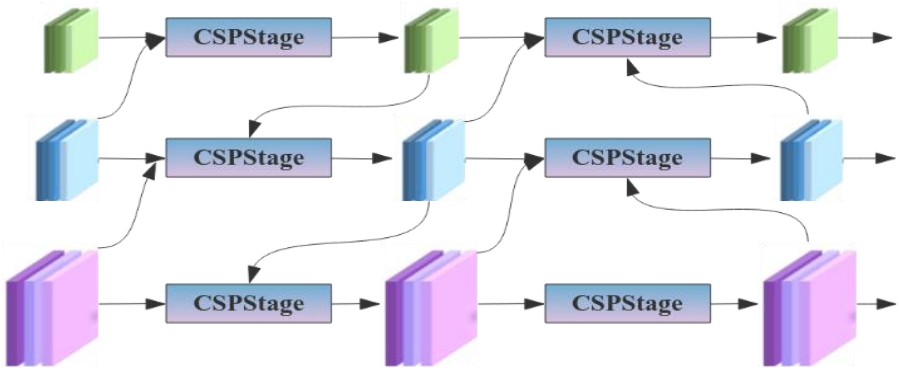

Figure 2: Efficient-RepGFPN architecture diagram

ISSN 2616-5775 Vol. 8, Issue 2: 37-44, DOI: 10.25236/AJCIS.2025.080205

Due to the significant differences in FLOPs among feature maps of different scales, it is challenging to control the scenario where each scale feature map shares the same dimensionality channel within the constraints of limited computational cost. Therefore, in the neck's feature fusion, a setting of different scale feature maps with different-sized channels is adopted. A comparison was made between the performance of identical and different channels, as well as the accuracy advantages brought by the trade-offs in neck depth and width. Experiments show that by flexibly controlling the number of channels at different scales, higher accuracy can be obtained than sharing the same channel at all scales. The best performance is achieved when the depth equals the width, which are (96, 192, 384).

GFPN enhances feature interactions through queen-fusion, but it also introduces a large number of additional up-sampling and down-sampling operators. A comparison of the advantages of these up-sampling and down-sampling operators shows that the additional up-sampling operators result in an increase in latency of 0.6ms, while the accuracy improvement is only 0.3mAP, which is far less than the benefits brought by the additional down-sampling operators. Therefore, under the constraints of real-time detection, the additional up-sampling operations in queen-fusion are removed.[12]

ZeroHead can maximize the reduction of computational cost in the RepGFPN neck. It is noteworthy that ZeroHead can essentially be considered as a coupled head, which is significantly different from the decoupled heads in other works. In the loss after the head, following GFocal, Quality Focal Loss (QFL) is used for classification supervision, and Distribution Focal Loss (DFL) and GIOU loss are used for regression supervision. QFL encourages the learning of a joint representation of classification and localization quality. DFL provides a more informative and precise bounding box estimation by modeling the bounding box location as a general distribution. The training loss formula for the proposed DAMO-YOLO is:

$$Loss = \alpha loss_{QFL} + \beta loss_{DFL} + \gamma loss_{GIOU} \tag{1}$$

Apart from the head and loss, label assignment is an important component during the detector's training, which is responsible for assigning classification and regression targets to predefined anchor points. The misalignment of classification and regression is a common issue in static assignment methods. Although dynamic assignment alleviates this problem, it still exists due to the imbalance between classification and regression losses, such as CrossEntropy and IoU Loss. To address this issue, this paper introduces focal loss into the classification cost and uses the IoU between the predicted and ground truth boxes as soft labels. The formula is as follows:

$$\text{AssignCost} = \text{C}_{\text{reg}} + \text{C}_{\text{cls}}$$

$$\propto = \text{IoU}\left(\text{reg}_{\text{gt}}, \text{reg}_{\text{pred}}\right)$$

$$C_{reg} = -ln(\propto)$$

$$C_{cls} = \left(\propto - cls_{pred}\right)^2 \times CE\left(cls_{pred}, \propto\right) \tag{2}$$

2.3 Highly Efficient Multi-scale Attention Module

To address the issues of complex background and indistinct feature extraction in small object detection, this paper presents an improvement based on the Highly Efficient Multi-scale Attention Module (EMA). The Highly Efficient Multi-scale Attention (EMA) mechanism is a parallel attention mechanism specifically designed to solve computer vision tasks, aiming to enhance model performance and processing speed. Compared with traditional Convolutional Neural Networks (CNN), EMA adopts a parallel convolution structure to process input data, thereby accelerating model training speed, especially having significant advantages when dealing with large-scale data. Additionally, EMA effectively enhances the model's pairwise relationships at the feature pixel level by aggregating information from different spatial dimensions and fusing the attention maps of parallel subnetworks. This helps to improve the localization and detection accuracy of small objects, thereby enhancing the model's perception capability and performance[13].

Let the input tensor be represented as, where \(C\) denotes the number of input channels, and \(H\) and \(W\) represent the dimensions of the input feature space. After grouping the channel dimension, the branch and the branch are composed into parallel subnetworks, which are used to capture the dependencies between channels and extract the attention weight descriptors of the grouped feature

ISSN 2616-5775 Vol. 8, Issue 2: 37-44, DOI: 10.25236/AJCIS.2025.080205

maps. The branch realizes cross-channel interactive features, and one-dimensional global average pooling outputs in both the horizontal and vertical dimensions are achieved through formulas (3) and (4).

$$z_c^H(H) = \frac{1}{W}\sum_{0\le i\le W} x_c(H,i) \tag{3}$$

$$z_c^H(W) = \frac{1}{H}\sum_{0\le J\le H} x_c(j,W) \tag{4}$$

In the formula, is the input feature at the \(c\)-th channel, and \(i\) and \(j\) represent the positions of the feature in the horizontal and vertical directions of \(H\), and is the output of the global average pooling. The branch captures local cross-channel interactions through convolution to expand the feature space. It encodes the inter-channel feature information and adjusts the importance between different channels, preserving the spatial information in the channels. The parallel subnetwork block helps to capture cross-dimensional interactions and establish inter-dimensional dependencies. Using the cross-spatial information aggregation method, the features are aggregated under different spatial dimensions. At this stage, two-dimensional global average pooling is used for different branches, and the formula is:

$$Z_c = \frac{1}{H\times W}\sum_i^H\sum_i^W x_c(i,j) \tag{5}$$

In the formula, is the input feature at the \(c\)-th channel, and \((i,j)\) represents the position where the feature is located, and is the output of the two-dimensional global average pooling. The output of each group of feature maps is calculated as the aggregation of two generated spatial sub-attentions, and finally, the feature is output through the Sigmoid function[14]，The structure diagram of the Efficient Multi-scale Attention Module (EMA) is shown in Figure 3：

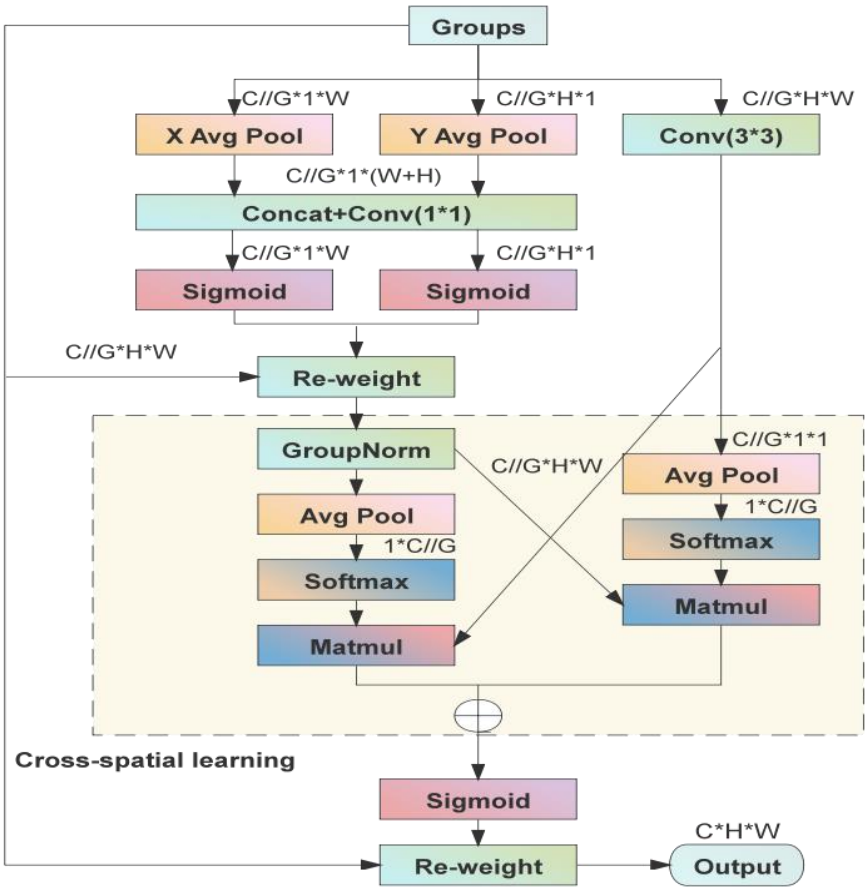

Figure 3: Structure Diagram of EMA (Efficient Multi-scale Attention Module)

To validate the effectiveness of the multi-dimensional collaborative attention module, it's necessary to show the distribution of attention across different regions when processing input images. Therefore,

ISSN 2616-5775 Vol. 8, Issue 2: 37-44, DOI: 10.25236/AJCIS.2025.080205

the distribution of attention information can be visualized. The results are shown in Figure 4:

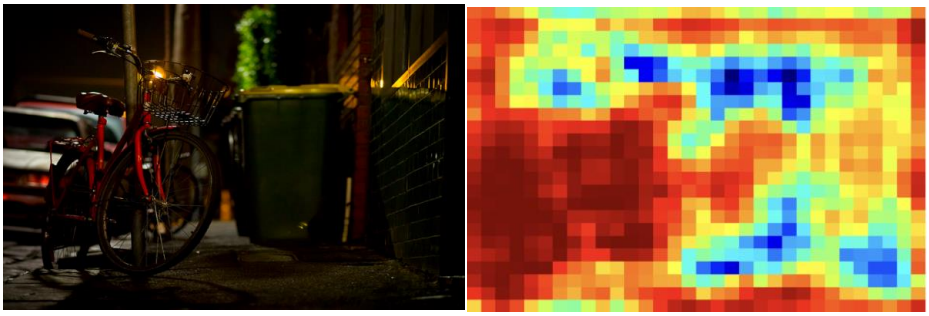

Figure 4: Attention Information Distribution Map

3. Experimental Analysis

3.1. Experimental Dataset and Experimental Details

This paper conducts experiments on the small object detection dataset to evaluate the YOLO_GF method. The VisDrone2019 dataset consists of a total of 8629 images, which are divided into training and validation sets and a test set at an 8:2 ratio. Furthermore, 10% of the images are separated from the training and validation sets as the validation set. Therefore, the dataset is divided into a training set, a validation set, and a test set with 6471, 548, and 1610 images respectively. The dataset is labeled with a total of 10 predefined categories including pedestrian, bicycle, car, van, truck, tricycle, awning-tricycle, bus, and motor.

To verify the effectiveness of the algorithm, this paper compares it with several existing popular algorithms. The experiments are conducted on a Tesla V100 GPU with 32GB VRAM in the Ubuntu 16.04 environment. During the model training process, the input image size is set to 640×640, the batch size is set to 64, the SGD optimizer is used for training, the initial learning rate is 0.01, the weight decay is 0.0005, and the total number of training epochs is set to 100[14].

This paper adopts Precision (P), the number of Parameters (Params), and the mean Average Precision of all samples (mAP) as evaluation indicators. Precision refers to the proportion of correctly predicted positive samples among all samples predicted as positive after predicting all samples, as shown in formula 6:

$$Percision = \frac{TP}{TP+FP} \tag{6}$$

TP refers to the condition where the predicted value and the true value are both positive samples. FP refers to the condition where the predicted value and the true value are different, with the predicted value being a positive sample and the true value being a negative sample. The number of parameters (Params) refers to the learnable weights and biases in the algorithm. The quantity of parameters can be used to assess the complexity and storage requirements of the model. The Average Precision (AP) is calculated as the area enclosed by the P-R curve and the coordinate axes. The detection accuracy under different categories within the dataset is mainly evaluated using mAP, which is calculated by averaging the AP values for each detection category, as shown in formulas (7) and (8). A higher mAP value indicates better object detection performance.

$$AP = \int_0^1 P(R)\, dR \tag{7}$$

$$mAP = \frac{1}{class\_num} \int_0^1 P(R)\, dR \tag{8}$$

3.2. Analysis of Experimental Results

3.2.1. Quantitative Analysis

This section conducts experiments using the divided VisDrone2019 dataset to evaluate the effectiveness of the YOLO_GF method and compare it with several mainstream object detection methods in terms of objective evaluation indicators and visualized results. As shown in Table 1, which displays the comparison of their mAP values, it can be seen that the YOLO_GF method has higher

ISSN 2616-5775 Vol. 8, Issue 2: 37-44, DOI: 10.25236/AJCIS.2025.080205

accuracy. The experimental results indicate that the YOLO_GF method achieves a 10.4% improvement in accuracy compared to YOLOv5[15].

Table 1: Results of Different Methods on the VisDrone2019 Dataset

Object Detection Model	Backbone Network	P	mAP
Faster R-CNN	ResNet-50	-	30.8%
EfficientDet	Efficient-B0	-	32.3%
YOLOv3	DarkNet-53	-	31.8%
YOLOv5	DarkNet-53	91.2%	33.6%
YOLOv7	DarkNet-53	90.9%	35.7%
SSD	VGG-16	-	31.5%
DAMO-YOLO	CSPDarkNet-53	-	36.2%
YOLO_GF	DarkNet-53	91.1%	39.2%

3.3. Ablation Experiment

3.3.1. Quantitative comparisons

By comparing the YOLO_GF algorithm with different improved algorithms, it can be concluded that two improvement schemes have enhanced the effect of small object detection. As shown in Table 2, compared with the original algorithm, in Experiment 2, from the perspective of the model's average precision (mAP), the average precision is increased by 3.65%, but the detection precision (P) is decreased by 0.01%. Therefore, Experiment 2 is conducted, and it is found that the novel low-light-level detection algorithm YOLO_GF has improved both the average precision (mAP) and the detection precision (P) compared to the original YOLOv5s algorithm. Compared with the original model, the accuracy and average precision are increased by 0.04% and 5.60% respectively, resulting in better detection performance.

Table 2: Ablation Experiment Results on the VisDrone2019 Dataset

Algorithm	DAMO-YOLO	MCA	P	mAP
YOLOv5s			91.06%	33.60%
Experiment 1	√		91.05%	37.25%
Experiment 2	√	√	91.10%	**39.20%**

4. Conclusion

Due to the inaccurate detection results, low precision, and easy missed detection of general small object detection algorithms, this paper proposes a new small object detection algorithm called YOLO_GF, which mainly focuses on the three issues of missed detection of small targets, false detection, indistinctive feature extraction, and missed detection of occluded objects in the aerial environment. Firstly, the DAMO-YOLO network is used for feature fusion, and an attention module is proposed to reduce the computational complexity. Secondly, combined with CSPStage, the network Neck part is designed to reduce the influence of interference among dense objects. Compared with the mainstream object detection methods on the low-light-level target detection dataset VisDrone2019, the

ISSN 2616-5775 Vol. 8, Issue 2: 37-44, DOI: 10.25236/AJCIS.2025.080205

YOLO_GF method has higher detection accuracy in low-light scenes, but the detection speed needs to be improved. The next step will be to study the lightweight network structure model, explore the data sharing methods among modules, so as to reduce the parameters and computation, and ultimately improve the target detection speed.

References

[1] Ge Z, Liu S, Wang F, et al. Yolox: Exceeding yolo series in 2021[J]. arXiv preprint arXiv:2107.08430, 2021.
[2] Rahmati M, Pompili D. UNISeC: Inspection, separation, and classification of underwater acoustic noise point sources[J]. IEEE Journal of Oceanic Engineering, 2017, 43(3): 777-791.
[3] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[J]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016: 779-788.
[4] Liu K, Sun Q, Sun D, et al. Underwater target detection based on improved YOLOv7[J]. Journal of Marine Science and Engineering, 2023, 11(3): 677.
[5] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[J]. Springer, Cham: European conference on computer vision, 2016: 21-37.
[6] Tan M X, Pang R M, Le Q V. EfficientDet: scalable and effi-cient object detection[C] //Proceedings of the IEEE/CVFConference on Computer Vision and Pattern Recognition.Los Alamitos: IEEE Computer Society Press, 2020:10778-10787.
[7] Tan M, Le Q. Efficient Net: rethinking model scaling for con-volutional neural networks[C] //Proceedings of International Conference on Machine Learning. New York: PMLR, 2019:6105-6114.
[8] Chen Q, Wang Y M, Yang T M, et al. You only look one-level feature[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 13034-13043.
[9] Wang, Chengcheng, et al. "Gold-YOLO: Efficient Object Detector via Gather-and-Distribute Mechanism." arXiv preprint arXiv:2309.11331 (2023).
[10] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 7132-7141.
[11] Xu X, Jiang Y, Chen W, et al. DAMO-YOLO : A Report on Real-Time Object Detection Design[J].ArXiv, 2022, abs/2211.15444.DOI:10.48550/arXiv.2211.15444.
[12] Wang F Y, Hu H T, Shen C. BAM: a lightweight and efficient balanced attention mechanism for single image super resolution [OL]. [2021-09-10]. https://arxiv.org/abs/2104.07566.
[13] Wang, Wenhai, et al. "Internimage: Exploring large-scale vision foundation models with deformable convolutions." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.
[14] Zhu L, Wang X, Ke Z, et al. BiFormer: Vision Transformer with Bi-Level Routing Attention[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 10323-10333.
[15] Deng W, Yuan H, Deng L, et al. Reparameterized Residual Feature Network for Lightweight Image Super-Resolution[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 1712-1721.