International Journal of Science and Engineering Applications
Volume 15-Issue 06, 06 – 12, 2026, ISSN:- 2319 - 7560
DOI: 10.7753/IJSEA1506.1002

Aerial Small Target Detection Algorithm based on YOLOv8

Zhao Xiang
School of Electronic Information and Electrical Engineering
Yangtze University
Jingzhou, China

Abstract: Drone aerial small target images suffer from defects such as overly dense detection targets, excessively small sizes, and difficulty in extracting feature information. These defects lead to low detection accuracy of existing target detection algorithms on aerial small target images. To address these issues, a drone aerial small target detection algorithm, MGF-YOLOv8, based on YOLOv8s and utilizing the Slice-Assisted Hyper-Inference (SAHI) method is proposed. Firstly, the SAHI slicing method is employed to slice remote sensing small target images, effectively mitigating the defects of overly dense detection targets and excessively small sizes. Secondly, a generalized high-efficiency layer aggregation network (GELAN) is incorporated into the Backbone part of YOLOv8 to replace the C2f module in the backbone network. This simplifies the backbone network structure, enhances feature extraction capabilities, and constructs a lightweight model. Then, a multi-scale structure (MS-block) is adopted in the neck network for feature fusion, reducing parameters while optimizing the model's performance in recognizing targets of different scales, complex backgrounds, and small targets. Finally, an additional P2 detection head is introduced to enhance the resolution of the feature map, thereby better locating small targets. Additionally, the ASFF mechanism is adopted to improve the original detection head, forming a new detection head, FASSF, to filter out information conflicts during multi-scale feature fusion and optimize the detection process, significantly enhancing small target detection capability. Experimental results on the VisDrone2019 dataset show that the mAP@0.5 and mAP@0.5:0.95 of MGF-YOLOv8 reach 56.9% and 36.3%, respectively, representing an improvement of 18.4% and 13.3% compared to the YOLOv8 algorithm. The parameter count is 10.26×10^6, a reduction of 7.82% compared to the original algorithm. The accuracy of this algorithm surpasses other similar algorithms and meets monitoring requirements, making it effectively applicable to target detection tasks on drone aerial platforms.

Keywords: YOLOv8; aerial small target; SAHI; small target detection layer

1. INTRODUCTION

Drones are autonomous aerial devices that possess excellent maneuverability and convenient operational characteristics, enabling precise execution of positioning and navigation tasks. These devices find widespread applications in fields such as geographic mapping, disaster response, and infrastructure inspection and maintenance. Drone target detection technology[1] combines advanced technologies like computer vision and machine learning. Compared to ground-based monitoring systems, it can capture a global view from the air, providing more comprehensive traffic flow analysis and more precise target detection in hard-to-reach areas. This feature effectively reduces labor costs. However, drone-captured images often cover large geographic areas, making small target detection particularly challenging. These small targets (such as people, vehicles, etc.) appear small and lack distinct details in the images. Techniques like edge detection, template matching, and morphological processing, while effective in certain scenarios, generally suffer from poor adaptability, limited robustness, and real-time performance, making them difficult to meet current target detection requirements[2].

In recent years, with the development of artificial intelligence, deep learning-based object detection algorithms have achieved remarkable success in the field of image detection due to their excellent real-time performance and accuracy[3]. Existing deep learning-based object detection algorithms are mainly divided into two-stage algorithms and single-stage algorithms[4]. Two-stage algorithms typically involve generating candidate regions first, then extracting features from each candidate box, and determining whether it contains an object through a classifier. However, they suffer from insufficient accuracy in detecting small objects[5], such as Faster R-CNN[6] and Mask R-CNN[7]. Single-stage algorithms directly transform the object detection task into an end-to-end regression problem, directly locating and classifying objects in the input image without generating candidate regions, saving a significant amount of computational power, such as Single Shot Multibox Detector (SSD)[8] and YOLO[9]. Among them, the YOLO series of algorithms has become one of the mainstream algorithms in the field of object detection due to its high detection speed and accurate detection accuracy. However, when facing the challenge of detecting small objects, the YOLO series of algorithms still exhibit issues such as poor handling of dense targets, low detection accuracy, large parameter count, and susceptibility to missed and false detections.

To address the challenges of detecting small targets and limited computational resources, Wibowo et al.[10] proposed the YOLOv7-MOD algorithm with deformable convolution kernels. This algorithm can adaptively adjust the shape of convolution kernels to extract features more effectively, but there is still a certain degree of false alarm during the detection process. Cao et al.[11] proposed the MSD-YOLO algorithm, which effectively reduces the issues of missed and false detections in dense occlusion scenarios by replacing the downsampling module and introducing the Soft-NMS algorithm. However, in severe occlusion environments, the detection accuracy still needs to be improved. Yang Lei et al. [12] proposed the FAN-YOLOv8n algorithm, which enhances feature fusion by fusing four feature layers of different scales in the backbone network and using parallel computation of large-kernel convolution and small-kernel convolution. Additionally, a module is introduced in the neck to strengthen the interaction between deep and shallow features, improving the detection capability of occluded targets

but resulting in an increase in parameter count. Liu et al.[13] proposed the DF-YOLO algorithm, which efficiently extracts multi-scale feature information through FasterNet and utilizes the DF-PAN module to achieve effective fusion of deep and shallow features, significantly improving the detection capability of targets of different sizes and reducing model parameters. However, this method still faces issues of missed and false detections in complex scenarios.

Despite significant progress in enhancing detection accuracy and optimizing computational complexity in existing research, there remains a challenging balance between detection accuracy and hardware computing power in small object detection tasks. To address this, this paper proposes an improved algorithm based on YOLOv8, named MGF-YOLOv8. It combines the slicing aided hyper inference (SAHI) data augmentation method and integrates the generalized high-efficiency layer aggregation network (GELAN) and multi-scale structure (MS-block) modules into YOLOv8, adding a small object detection layer and designing an FASFF detection head to replace the original one. Experimental results show that MGF-YOLOv8 can effectively improve the detection accuracy of aerial small objects, providing more ideas for small object detection.

2. Introduction to YOLOv8

YOLOv8[14] is the eighth version of the YOLO series of object detection models, proposed by Ultralytic in 2023, specifically designed for image classification, object detection, and instance segmentation tasks. In terms of architecture, YOLOv8 integrates the technical advantages of YOLOv5 to YOLOX and has undergone optimization, achieving significant improvements in accuracy, speed, inference capability, and multi-task support. It strikes a balance between detection speed and detection accuracy. Therefore, this paper chooses YOLOv8 as the basic model for drone small target detection. YOLOv8 is divided into multiple versions based on different network depths and widths, including YOLOv8n, YOLOv8s, YOLOv8m, YOLOv8, and YOLO8x. The network structures of these versions are basically the same, with the only difference being the network depth and width[15]. The YOLOv8 model mainly consists of three key network layers: the backbone network (Backbone), neck network (Neck), and detection head (head). The specific network architecture is shown in Figure 1. During the detection task, the model first performs feature extraction through the backbone network, which includes convolution (Conv), C2f, and SPPF modules, to obtain feature maps of different scales. The extracted features are then input into the neck network [the PAFPN architecture composed of Conv, C2f, concatenation (Concat), upsampling (Upsample), and other modules], which constructs a feature pyramid by fusing features of different scales to obtain richer feature information from the input image. Finally, the head part detects and classifies feature maps of three different sizes: large, medium, and small.

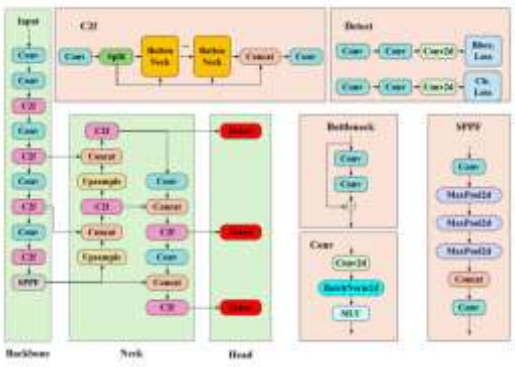

Figure. 1 YOLOv8 Architecture Diagram

3. Improved MGF-YOLOv8 network structure

Compared to other versions of the YOLO series, YOLOv8s strikes a good balance between detection accuracy and detection speed, thus it is selected as the benchmark model. However, due to the presence of numerous small-scale targets in the drone dataset, coupled with the complex and variable geographical environment, there are various types of interference factors. As a basic model, YOLOv8s suffers from issues such as "missed detections" and poor recognition. To achieve efficient detection of small targets in aerial photography environments, the main improvements made to the YOLOv8 structure in this paper are as follows:

1. To address the challenge of detecting small targets in high-resolution images while maintaining high memory utilization, this paper employs the Slice-Assisted High-Resolution Imaging (SAHI) technique. This technique processes input network images by slicing them, generating larger pixel regions for small target objects, thereby enhancing the effectiveness of network inference and fine-tuning, and providing more refined features for subsequent models;

2. A novel lightweight network architecture based on gradient path planning, the Generalized Efficient Layer Aggregation Network (GELAN), is adopted in the backbone network to replace the C2f module in the core network, simplifying the core network structure, enhancing its feature extraction capability, and achieving model lightweighting;

3. Secondly, in the neck network of YOLOv8, MSBlock is integrated into the C2f module to obtain the C2f_MS module. The first stage of the encoder uses the smallest convolution kernel, while the final stage adopts the largest convolution kernel. This design matches the gradual increase in feature resolution. This architecture not only enhances the ability to extract fine-grained and coarse-grained semantic information, improves the multi-scale feature representation ability of the encoder, but also further optimizes the performance of the model in recognizing targets of different scales, complex backgrounds, and small targets;

4. To address the deficiency of the model in small object recognition, a small object detection layer is added to the original model, and the original detection head is improved using ASFF to form a new detection head, FASSF.

The improved network structure is shown in Figure 2.

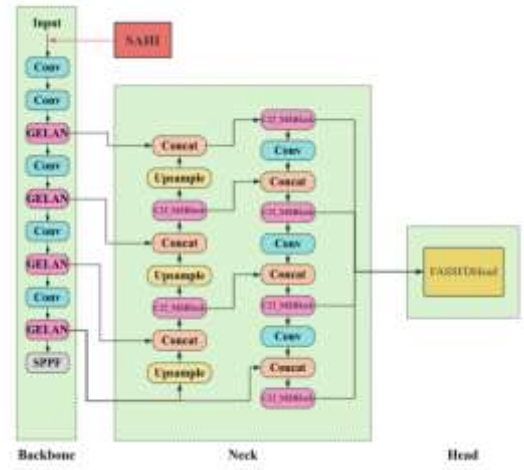

Figure. 2 MGF-YOLOv8 Architecture Diagram

International Journal of Science and Engineering Applications
Volume 15-Issue 06, 06 – 12, 2026, ISSN:- 2319 - 7560
DOI: 10.7753/IJSEA1506.1002

3.1 Slicing Assisted High-Resolution Imaging (SAHI)

SAHI is a general computational framework designed to address the challenge of detecting small targets in large-sized images and complex background images, and is commonly used in the model inference and prediction stage. Its core functionality is to automatically segment large-sized images or video streams into multiple overlapping small images (slices). The slicing process can be uniform or dynamically adjusted based on image content. Then, object detection is independently run on each slice, and the detection results of each slice are subsequently aggregated. SAHI eliminates redundant boxes generated by slice overlap through intelligent merging algorithms (such as non-maximum suppression based on intersection over union and confidence), stitches the slices, and finally outputs complete global detection results. This effectively avoids issues of repeated inspection and missed detection. The specific structure is shown in Figure 3.

Addressing the limitations of the VisDrone2019 dataset, such as significant background interference, numerous occlusions, and small targets, this paper employs the SAHI slicing algorithm for preprocessing. By setting a slicing size of 640×640 and an overlap rate of 0.15, an optimized dataset is generated. The original image and the sliced sub-image are shown in Figures 4 and 5, respectively. This processing effectively mitigates the defects of the VisDrone2019 dataset, including dense detection targets and excessively small sizes.

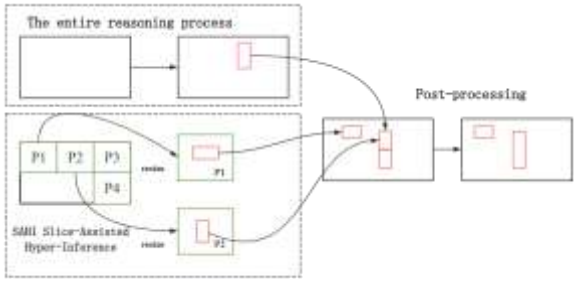

Figure. 3 SAHI slice-assisted super-inference method

Figure. 4 Original image of VisDrone2019 dataset

Figure. 5 Sub-image of VisDrone2019 dataset after slicing

3.2 RepNCSPELAN module

In deep neural networks, information bottlenecks lead to irreversible information loss during forward propagation of input data, thereby affecting model accuracy and convergence efficiency. Although existing reversible architectures, reconstruction loss functions, and deep supervision methods attempt to alleviate this issue, they often come with limitations such as increased inference cost, error accumulation, or loss of shallow information, especially in complex tasks and small object detection. To address these challenges, researchers have proposed the Generalized Efficient Layer Aggregation Network (GELAN) based on the CSPNe[16] and ELAN[17] architectures. In the CSPNet architecture, the input is split into two parts through a transformation layer and then passes through arbitrary computation blocks separately. Afterwards, these branches are re-merged (via concatenation) and pass through the transformation layer again. Compared to CSPNet, ELAN employs stacked convolutional layers, where the output of each layer is combined with the input of the next layer and then processed through convolution. GELAN combines the designs of CSPNet and ELAN, adopting the concept of segmentation and recombination from CSPNet and introducing the hierarchical convolutional processing method from ELAN in each part. The difference lies in that GELAN not only uses convolutional layers but also any computation blocks, making the network more flexible and customizable according to different application requirements. This design not only reduces the number of model parameters and computational complexity but also enhances model flexibility, making it suitable for a wide range of hardware and application scenarios. The network results of CSPNet, ELAN, and GELAN are shown in Figure 6.

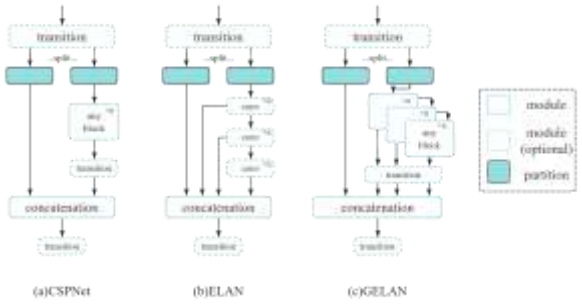

Figure. 6 The architecture of GELAN comprises: (a) CSIPNet, (b) ELAN, and (c) GELAN.

Based on the design philosophy of the GELAN network, this paper replaces the original C2f module with its key module RepNCSPELAN to further optimize the network structure. RepNCSPELAN performs feature extraction and fusion through a series of convolutional layers (Conv) and RepNCSP (Replicated Cross Stage Partial Network) modules. The main components of this module include the number of input channels (c1), the number of output channels (c2), and the number of intermediate channels (c3, c4), which determine the size and complexity of the feature maps. Through the synergistic effect of

International Journal of Science and Engineering Applications
Volume 15-Issue 06, 06 – 12, 2026, ISSN:- 2319 - 7560
DOI: 10.7753/IJSEA1506.1002

these convolutional layers and RepNCSP modules, RepNCSPELAN can effectively fuse local and global features, thereby enhancing the accuracy and robustness of object detection. While maintaining the efficiency of feature extraction, the RepNCSPELAN module significantly reduces the number of network parameters and computational complexity, optimizes the use of hardware resources, and achieves a better balance between performance and resource consumption for the entire detection model. The specific structure is shown in Figure 7.

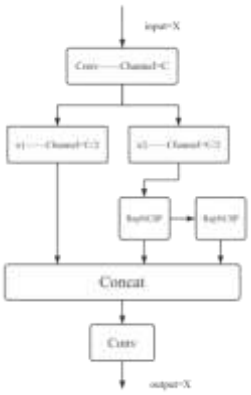

Figure. 7 RepNCSPELAN structure diagram

From the comparison of the success rates of the two algorithms with changes in signal-to-noise ratio and number of snapshots in Figures 8 and 9, it can be seen intuitively that as the signal-to-noise ratio or number of snapshots increases, the success rates of both algorithms improve. However, under the same conditions of number of snapshots or signal-to-noise ratio, the performance of our algorithm is better than that of traditional algorithms.

3.3 C2f_MSBlock module

In the YOLOv8 network, the C2f module serves as the core feature extraction unit. In the Neck, it primarily integrates multi-scale features through cross-stage partial connections to enhance gradient flow and semantic expression. However, in drone aerial photography scenarios, its multiple downsampling and deep fusion mechanisms tend to cause the loss of shallow-level details. Moreover, the fixed receptive field struggles to adapt to extremely small targets and drastic scale variations, leading to missed detection of small targets and reduced localization accuracy. To address this, this paper draws inspiration from advanced designs in the multi-scale detection domain, replacing the Neck with an MSBlock featuring heterogeneous convolution kernels and a hierarchical feature fusion mechanism, thus forming an improved C2f_MSBlock structure, as illustrated in Figure 8.

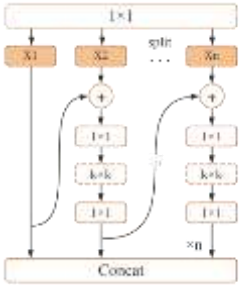

Figure. 8 MSBlock structure diagram

The MSBlock module, proposed by YOLO-MS[18], is based on the fundamental principle of enhancing the multi-scale feature representation capability of real-time object detectors. By adopting a hierarchical feature fusion strategy and a heterogeneous convolution kernel selection protocol, MSBlock effectively processes features of different scales at different stages of the network. This design enables the detector to better recognize and handle objects of varying sizes. Its core idea is to enhance the model's performance in handling multi-scale information by improving the size and structure of convolution kernels, as well as optimizing the feature fusion method, thereby improving the overall accuracy and efficiency of object detection.

3.4 Adaptive spatial feature fusion detection head

In the YOLOv8s model, the input image is resized to 640×640×3 and generates three feature maps with dimensions of 80×80×128, 40×40×256, and 20×20×512, respectively. However, when detecting small targets, the minimum size of the feature map is reduced to 20×20×512, which may lead to the loss of a large amount of feature information of small targets. In addition, due to deep convolution operations, the feature information of small targets may be further weakened, thereby reducing detection accuracy. To address this issue, this paper introduces an output from the first C2f module in the Backbone (with dimensions of 160×160×64), adds upsampling, Concat, and C2f modules in the Neck, and adds a Detect module in the Head, which introduces a high-resolution P2 head specifically designed for small targets, while maintaining the original P3-P5 heads to achieve comprehensive scale coverage.

Single-stage detectors suffer from deficiencies in handling scale variations, which stem from the fundamental trade-off between resolution and semantic information across different feature levels. Shallow feature maps retain the fine spatial details necessary for small object detection, but lack semantic richness; whereas deep features provide powerful semantic representations, but at the cost of losing crucial spatial information. This inconsistency leads to small object features being diluted or treated as background in the deep layers, generating conflicting information during multi-scale feature fusion and suboptimal gradient propagation during training. To address this, as shown in Figure 9, we introduce an Adaptive Spatial Feature Fusion (ASFF) mechanism to integrate with the original detection head, effectively filtering out conflicting information and enhancing scale invariance.

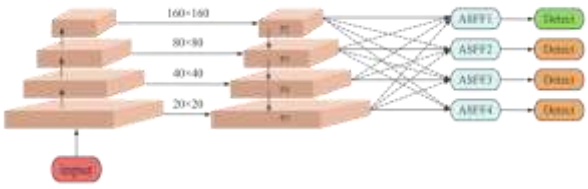

Figure. 9 ASFF structure diagram

4. Experimental data and result analysis

4.1 Dataset and evaluation metrics

To evaluate the MGF-YOLOv8 algorithm, the classic remote sensing small target dataset, VisDrone2019 dataset[32], was adopted. The VisDrone2019 dataset, released by Tianjin University, is a large benchmark dataset specifically designed for drone vision tasks. It is one of the classic remote sensing small target datasets, encompassing extensive data collected from 14 different cities in China. These images and videos were captured by various drone cameras, documenting scenes under diverse weather and lighting conditions. The VisDrone2019 dataset comprises 8,629 images, with 6,471 for training, 548 for

International Journal of Science and Engineering Applications
Volume 15-Issue 06, 06 – 12, 2026, ISSN:- 2319 - 7560
DOI: 10.7753/IJSEA1506.1002

Yolov8s	module A	module B	module C	SAHI	mAP@0.5/%	mAP@0.5 ：0.95/%	Precision/%	Recall/%	Parameter/ ($\times 10^6$)
√					38.5	23.0	49.2	38.3	11.13
√	√				40.4	23.9	51.5	40.5	8.56
√		√			40.6	24.1	50.7	39.8	10.04
√			√		44.8	27.2	53.8	43.3	14.24
√	√	√			41.3	24.5	51.9	41.2	8.23
√	√		√		45.2	28.3	54.4	43.8	11.07
√		√	√		45.1	28.5	53.2	43.2	12.73
√	√	√	√		46.3	29.5	57.6	44.7	10.26
√				√	56.2	35.7	64.4	52.6	11.13
√	√	√	√	√	58.9	38.3	66.4	54.8	10.26

Table. 1 Ablation experiment on VisDrone2019 dataset

validation, and 1,610 for testing. The images contain 10 annotated object classes, namely pedestrian, person, car, van, bus, truck, motorcycle, bicycle, awning-tricycle, and tricycle, totaling over 260×104 annotated boxes. Due to issues such as severe target occlusion, partially small targets, and uneven data distribution, the VisDrone2019 dataset remains one of the more challenging remote sensing small target datasets currently. To assess the detection accuracy of the model for remote sensing small target objects, evaluation metrics including precision (P), recall (R), mean Average Precision (mAP), and parameter count (Params) were primarily used to evaluate the model size.

4.2 Ablation experiment

To verify the effectiveness of the adopted data augmentation method, SAHI image slicing, and the three modules utilized (GELAN (Module A), MS-block (Module B), and FASFF (Module C)), ablation experiments were conducted on the VisDrone2019 dataset. A " √ " indicates the addition of that module or the adoption of that method. The results of the ablation experiments are presented in Table 1.

Table 1 shows that after slicing the VisDrone2019 dataset using the SAHI method, the average precision (mAP @ 0.5) of the original YOLOv8 algorithm increased by 15.7%, while the mAP @ 0.5 of the MGF-YOLOv8 algorithm increased by 18.4%. This proves that the SAHI slicing method can effectively improve the problem of dense and small-sized small target images. After combining the GELAN module with YOLOv8s, the mAP @ 0.5 increased by 1.9%, and the parameter count decreased from 11.13×10^6 to 8.56×10^6, a reduction of 23.1%. This effectively improved the feature extraction efficiency of the backbone network and reduced the model complexity. After introducing the MS-Block module, the mAP @ 0.5 increased by 2.1%. This module effectively improved the problem of insufficient multi-scale feature representation through hierarchical multi-branch feature extraction and global query dynamic guidance mechanism. After replacing the original detection head with the FASFF detection head (adding a P2 detection layer combined with ASFF technology), the mAP @ 0.5 increased from 38.5% to 43.8%, an improvement of 5.3%, with a relative improvement of 12.1%. The mAP @ 0.5:0.95 increased by 3.2%, significantly improving the detection accuracy. This proves that adding a small target detection layer and adaptive spatial feature fusion technology can effectively improve the problem of small target feature information being easily lost. Integrating the GELAN module on the basis of the FASFF detection head further improved the mAP @ 0.5 to 44.2%, an increase of 5.7%. The GELAN module ensures that the model extracts more effective feature information with fewer parameters, while the FASFF detection head retains richer feature information of small targets. After combining the MS-Block module with the FASFF detection head, the mAP @ 0.5 increased to 44.1%, an improvement of 5.6%, further enhancing the model's multi-scale feature fusion ability.

After combining the GELAN module, MS-Block module, and FASFF detection head, without using SAHI, the mAP@0.5 increased to 45.3%, an improvement of 6.8%, with a relative improvement of 17.7%. The parameter count was 10.26×10^6, a decrease of 7.8% compared to the baseline model, further enhancing the model's feature extraction capability while maintaining its lightweight advantage. Simultaneously, after adopting three improvements and the SAHI slicing method, the mAP@0.5 reached 56.9%, an improvement of 18.4%, with a relative improvement of 47.8%. The mAP@0.5:0.95 reached 36.3%, an improvement of 13.3%. The precision and recall rates reached 64.4% and 52.8%, respectively, with the parameter count remaining at 10.26×10^6, a decrease of 7.8% compared to the baseline model. This significantly improved the detection accuracy of small remote sensing targets and reduced the model's parameter count, demonstrating the effectiveness of the SAHI slicing method and the improved MGF-YOLOv8s algorithm.

4.3 Comparative experiments with other algorithms

From Table 2, it can be concluded that the classic object detection algorithm Faster R-CNN has high detection accuracy, but its parameter count is too high, resulting in an excessively large network size; RetinaNet also faces the problem of a large network size, and its detection accuracy is not high; while YOLOv5 has extremely low parameter and computation counts, as well as a lightweight network structure, it cannot balance detection accuracy, and its detection accuracy is too low when facing small object detection tasks; YOLOv3 performs decently in terms of detection accuracy, but its network size is also too large, making it difficult to meet real-time requirements; YOLOv3-tiny significantly reduces the parameter count based on YOLOv3, but also loses a large amount of accuracy; YOLOv8 and YOLOv10 achieve a good balance between detection accuracy and computational cost, but their model architecture is prone to false positives and missed detections when facing small object detection in dense scenes, resulting in low detection accuracy. The MGF-YOLOv8 algorithm has extremely high accuracy and low parameter count. Compared to the original YOLOv8 algorithm, it improves mAP @ 0.5 by 7.8%, mAP @ 0.5:0.95 by 6.5%, and reduces the parameter count by 7.82%. MGF-YOLOv8

effectively improves the detection performance of small target images in aerial photography.

模型	mAP@0.5/%	mAP@0.5 ：0.95/%	参数量/($\times 10^6$)
SSD	23.2	12.6	
RetinaNet	21.6	12.1	36.41
FasterR-CNN	33.9	18.4	42.39
YOLOv3	35.2	21.1	60.78
YOLOv3-tiny	15.4	6.5	12.74
YOLOv5s	34.7	19.2	7.33
YOLOv8n	31.8	18.1	3.10
YOLOv8s	38.5	23.0	11.13
YOLOv10s	31.5	17.2	8.04
MGF-YOLOv8	46.3	29.5	10.26

Table. 2 Comparative experiment on VisDrone2019 dataset

5. Conclusion

To address issues such as dense target images, excessively small target sizes, difficulties in extracting feature information, and low detection accuracy in aerial small target images, an improved aerial small target detection algorithm based on YOLOv8, named MGF-YOLOv8, is proposed. Firstly, the SAHI method is employed to slice the dataset, effectively improving issues such as dense targets and excessively small target sizes in aerial small target images. Secondly, the YOLOv8 algorithm is improved by incorporating a generalized high-efficiency layer aggregation network (GELAN) to replace the C2f module in the backbone network, simplifying the backbone network structure, enhancing feature extraction capabilities, and constructing a lightweight model. Then, a multi-scale structure (MS-block) is adopted in the neck network for feature fusion, reducing parameters while optimizing the model's performance in recognizing targets of different scales, complex backgrounds, and small targets. Finally, an additional P2 detection head is introduced to enhance the resolution of the feature map, thereby better locating small targets. Additionally, the ASFF mechanism is adopted to improve the original detection head, forming a new detection head named FASSF to filter out information conflicts during multi-scale feature fusion and optimize the detection process, thus significantly enhancing the small target detection capability. Compared with other advanced models, it can be concluded that the MGF-YOLOv8 algorithm exhibits superior small target recognition capabilities and effectively improves the detection accuracy of aerial small target images. Although the MGF-YOLOv8 algorithm achieves high detection accuracy, it also faces the issue of an overly large network volume. In the future, research on model lightweighting will continue, exploring ways to reduce the network volume and computational resource consumption of the model while maintaining its detection accuracy.

6. References

[1] WU Xin, LI Wei, HONG Danfeng, et al. Deep learning for unmanned aerial vehicle-based object detection and tracking: a survey[J]. IEEE Geoscience and Remote Sensing Magazine, 2022, 10(1): 91. DOI:10.1109/MGRS.2021.3115137

[2] BKUANG X Y, CHENG F J, WU C Q, et al. Efficientand lightweight remote sensing image object detectionmethod based on improved YOLOv7-tiny[J]. Journal ofElectronic Measurement and Instrumentation,2024,38(7): 22-33.

[3] Xue Linyan, Li Xuanang, Qi Chaoyi, et al. Real-time detection method for multi-classification of intestinal polyps based on improved YOLOv5s[J]. Journal of Hebei University (Natural Science Edition),2024,44(04): 424-432

[4] XU X Y,ZHAO M,SHI P X,et al.Crack Detection and Comparison Study Based on Faster R-CNN and Mask R-CNN[J].Sensors,2022,22(3):1215-1232.

[5] Liu Tao, Ju Shihong, Gao Yimeng. Small Target Detection Algorithm Based on Improved YOLOv8n from a Drone Perspective[J].Computer Applications, 2024, 44(11): 3603-3609

[6] Ren S Q, He K M, Girshick R, et al. Faster R-CNN:towards real-time object detection with region proposal networks[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2017,39(6): 1137-1149.

[7] HE K, GKIOXARI G, DOLL R P, et al. Mask R-CNN[C].Proceedings of the IEEE International Conference on Computer Vision, 2017: 2961-2969.

[8] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//Computer Vision–ECCV 2016: 14th European Conference. Amsterdam, October 11-14, 2016: 21-37.

[9] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection [C].Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2016: 779-788.

[10] Wibowo A, Trilaksono B R, Hidayat E M I, et al. Object detection in dense and mixed traffic for autonomous vehicles with modified yolo[J]. IEEE Access,2023,11: 134866-134877..

[11] Cao J, Zhang T, Hou L, et al. An improved yolov8 algorithm for small object detection in autonomous driving[J]. Journal of Real-Time Image Processing,2024, 21(4): 138.

[12] Yang Lei, Chen Yanfei, Li Haiming, et al. Object Detection Algorithm for Autonomous Driving Scenarios Based on Improved YOLOv8[J]. Computer Engineering and Applications, 2025, 61(1): 131-141

[13] Liu H, Lu G, Li M, et al. High-precision real-time autonomous driving target detection based on YOLOv8[J].Journal of Real-Time Image Processing, 2024, 21(5): 174.

[14] Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection[C]‖2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA.New York: IEEE Press, 2016: 779-788.

[15] Chen, Y.; Zheng, W.; Zhao, Y.; Song, T.H.; Shin, H. DW-yolo: An efficient object detector for drones and self-driving vehicles. Arab.J. Sci. Eng.2023,48, 1427–1436. [CrossRef]

[16] WANG C Y，LIAO H Y M，WU Y H，et al. CSPNet: A new backbone that can enhance learning capability of CNN [C].Proceedings of the IEEE/CVF conference on

computer vision and pattern recognition workshops.2020: 390-391.

[17] Wang, L.; Lee, C.-Y.; Tu, Z.; Lazebnik, S. Training deeper convolutional networks with deep supervision. arXiv 2015,arXiv:1505.02496.

[18] YOLO-MS:Rethinking Multi-Scale Representation Learning for Real-time Object Detection.