ISSN 2616-5775 Vol. 8, Issue 3: 116-125, DOI: 10.25236/AJCIS.2025.080316

YOLOv8-Plus: A Small Object Detection Model Based on Fine Feature Capture and Enhanced Attention Convolution Fusion

Hui Li[1,a,*], Xiaoyan Pang[1,b]

[1]School of Software, Henan Polytechnic University, Jiaozuo, 454000, China
[a]212209020080@home.hpu.edu.cn, [b]pangxy@hpu.edu.cn
**Corresponding author*

Abstract: *Small object detection holds significant value in various practical applications. However, due to their limited pixel coverage, weak feature information, and susceptibility to background noise, YOLOv8 faces challenges in detecting small objects, including low recognition accuracy and missed detections. To address these issues, we propose an improved small object detection model, YOLOv8-Plus. First, to tackle the difficulty in detecting subtle features of small objects in the YOLOv8 model, we add a dedicated output layer, TDLayer, in addition to the original three output layers. This new layer generates larger feature maps, allowing for better differentiation of fine details in small objects. Second, to improve feature processing, we design the C2FDSC module, which adaptively adjusts detection strategies based on the shape and characteristics of small objects, ensuring fine details are captured. Finally, to mitigate the impact of background noise, we introduce the EACF module, which combines the advantages of CNNs and attention mechanisms to effectively reduce noise interference, improving both accuracy and robustness in small object detection. Experimental results on the VisDrone2019 dataset show that the improved YOLOv8-Plus model achieves a 6.7% and 4.7% increase in mAP50, respectively, compared to the baseline model. YOLOv8-Plus outperforms other state-of-the-art models, demonstrating superior performance in small object detection tasks in complex scenarios.*

Keywords: *Small object detection, YOLOv8, Convolutional neural network, Attention mechanism*

1. Introduction

With the rapid development of computer vision and artificial intelligence technologies, the demand for object detection has significantly increased, especially in the detection of small objects. The applications are broad, including intelligent transportation[1], pedestrian detection[2], facial recognition[3], defect detection[4], robotics automation[5], remote sensing imaging[6], security monitoring[7], and medical applications[8]. To meet the demand for efficient and accurate small object detection, optimizing small object detection technology has become a major research challenge.

Object detection algorithms can be categorized into two-stage and one-stage methods. The two-stage methods, such as R-CNN[9], first generate candidate boxes, extract features, and then predict classes and locations, offering high accuracy but lower speed. Fast R-CNN[10] improves upon R-CNN by using a feature extraction network, which increases speed; Faster R-CNN[11] further replaces selective search with a Region Proposal Network (RPN), enabling end-to-end training. In contrast, one-stage methods, such as the YOLO series, transform the object detection problem into a regression problem, enhancing detection speed. YOLOv1[12] divides the image into grids and predicts bounding boxes and classes for each grid, while SSD[13] uses multi-scale feature maps to improve detection accuracy for objects of varying sizes. YOLOv2[14] and YOLOv3[15] optimize the feature extraction network and introduce a Feature Pyramid Network (FPN) to further enhance detection accuracy. YOLOv5[16] optimizes the backbone network and enhances multi-scale prediction capabilities. Subsequently, YOLOv6[17], YOLOv7[18], and YOLOv8[19] have further optimized and improved performance.

However, small objects present challenges due to their limited size and information, and they are often affected by complex backgrounds and noise, which degrade the performance of general object detection models. Improving the accuracy of small object detection remains a major challenge. In recent years, researchers have proposed various methods to improve the performance of small object detection. For instance, in[20-22], multi-scale feature fusion methods were employed by extracting information from

ISSN 2616-5775 Vol. 8, Issue 3: 116-125, DOI: 10.25236/AJCIS.2025.080316

feature maps at different levels and fusing these features to enhance small object detection capability. Other methods[23-25] focus on context-based detection methods that emphasize capturing the surrounding environment (context information) to aid in identifying small objects. Similar to feature fusion, context-based methods aim to provide more information to the final detection network, but small object information may be masked by redundant context information. Furthermore, some researchers have improved the performance of small object detection by increasing the input image resolution. The use of Generative Adversarial Networks (GAN)[26-28] aims to reduce the feature discrepancy between small objects and larger/mid-sized objects by mapping the low-resolution features of small objects to high-resolution object features, thereby enhancing the representation of small object features to achieve detection performance comparable to that of larger objects. However, GAN-based methods face the challenge of maintaining a balance between the discriminator and the generator. Another drawback of GAN-based methods is that the generator often struggles to produce enough samples during the training process.

To improve the accuracy of small object detection, this paper proposes an improved algorithm based on the YOLOv8 model, named YOLOv8-Plus. By adding a dedicated output layer (TDLayer) for small object detection in the network architecture, the model's performance in detecting small objects is enhanced. Additionally, by integrating the C2FDSC module and the Enhanced Attention Convolution Fusion (EACF) module, the model's performance on small object detection tasks is further improved. This method effectively reduces false negatives and false positives for small objects, providing better detection accuracy and real-time performance, making it suitable for a broader range of practical application scenarios.

2. YOLOv8

YOLOv8, released by Ultralytics, is a newer object detection model that inherits the advantages of the YOLO series in real-time detection. It improves inference speed while maintaining high accuracy. The YOLOv8 architecture consists of three main components: Backbone, Neck, and Head. In the Backbone of YOLOv8, CSPNet is used to extract image features, and the C3 module from YOLOv5 is replaced with the C2f module, which contains more residual connections, thereby enhancing detection accuracy. In the Neck, FPN fuses high-level semantic information with low-level detailed information, while PAN enhances the information transfer from low-level positions to high-level ones. The Head of YOLOv8 is responsible for predicting the object category, bounding box coordinates, and confidence score, separating the classification task from the regression task, which improves the accuracy of object detection.

Although YOLOv8 improves the detection capability of small objects by introducing a stronger feature fusion mechanism, precise localization and recognition of small objects in complex backgrounds or under occlusion remains a challenge. Therefore, the model requires more sophisticated design to effectively capture subtle features and enhance the detection performance of small objects.

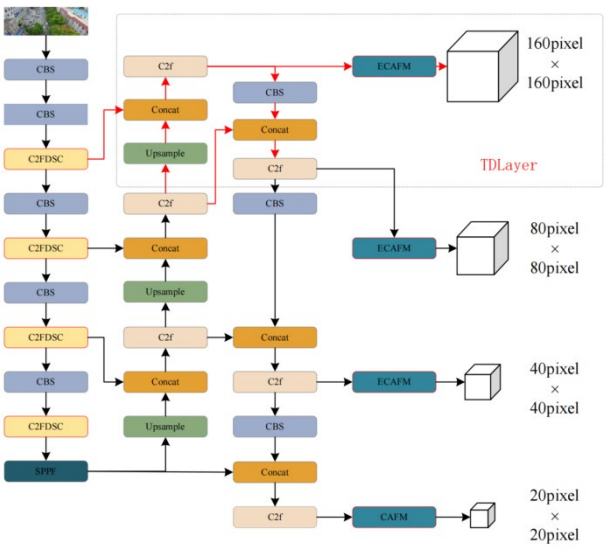

Figure 1: YOLOv8-Plus network structure.

ISSN 2616-5775 Vol. 8, Issue 3: 116-125, DOI: 10.25236/AJCIS.2025.080316

3. YOLOv8-Plus

This chapter proposes the YOLOv8-Plus model, specifically designed for small object detection, based on the YOLOv8s algorithm. First, the TDLayer, which adds a small object detection layer, is used to fully leverage the shallow spatial location information and semantic information of small objects, effectively enhancing the model's sensitivity to small objects. Next, the C2FDSC module is designed to better capture the complex structural features within images. Finally, the EACF module is adopted to integrate both global and local features, strengthening the focus on small objects. The overall model structure is shown in Figure 1.

3.1. TDLayer

YOLOv8 employs a multi-scale feature fusion structure, utilizing three different feature layers (P2, P3, and P4) to achieve multi-scale object detection. For example, when the input image is 640×640, after convolutional downsampling with strides of 8, 16, and 32, the three detection layers output feature maps of sizes 80×80, 40×40, and 20×20, respectively. These feature maps correspond to the detection of small, medium, and large objects. However, the 80×80 feature map, which has the smallest receptive field, represents information from an 8×8 region of the original image. Given that many small objects are even smaller than this, after several convolution operations, the image size gradually reduces, causing the representation of smaller objects on the feature map to become weaker. As a result, YOLOv8 may suffer from missed detections when it comes to small object detection.

The feature map at the P1 level retains a high resolution, which provides the network with more detailed information and enhances the detection of small objects. Therefore, this paper introduces an additional detection layer, TDLayer, based on the original three feature layers, incorporating a 4x downsampling factor. This modification enables a more effective feature fusion framework, as illustrated in Figure 2. The feature map generated by TDLayer has a size of 160×160, with each pixel corresponding to a 4×4 region of the original image. This improvement allows the network to efficiently detect small objects with a resolution of at least 16 pixels. This multi-scale fusion approach significantly improves the detection of objects at various sizes, with particular emphasis on small objects, thereby enhancing the overall performance of the network in real-world object detection tasks.

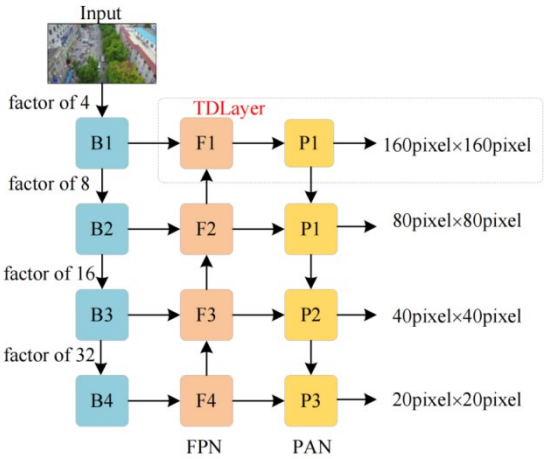

Figure 2: Detection framework.

3.2. C2FDSC module

In the YOLOv8 algorithm, the C2f module acts as a crucial component designed to maintain the model's lightweight nature while providing richer gradient flow information. However, for small object detection, traditional convolution operations often struggle to effectively capture these subtle features, as small objects typically have complex shapes and boundaries and occupy a small proportion of the image's pixels.

To address this, we propose a new module called the C2FDSC, which is designed to capture fine-grained features. The structure of this module is shown in Figure 3. In the C2FDSC module, the C2f component is responsible for cross-stage feature fusion, thus providing the model with comprehensive image information. Meanwhile, DSC[29] dynamically adjusts the receptive field of convolution kernels,

ISSN 2616-5775 Vol. 8, Issue 3: 116-125, DOI: 10.25236/AJCIS.2025.080316

focusing on capturing and processing image structures with complex shapes and backgrounds. This enhances the model's ability to perceive such structures. By combining DSC with the C2f module, we effectively leverage the strengths of both components in feature extraction and fusion.

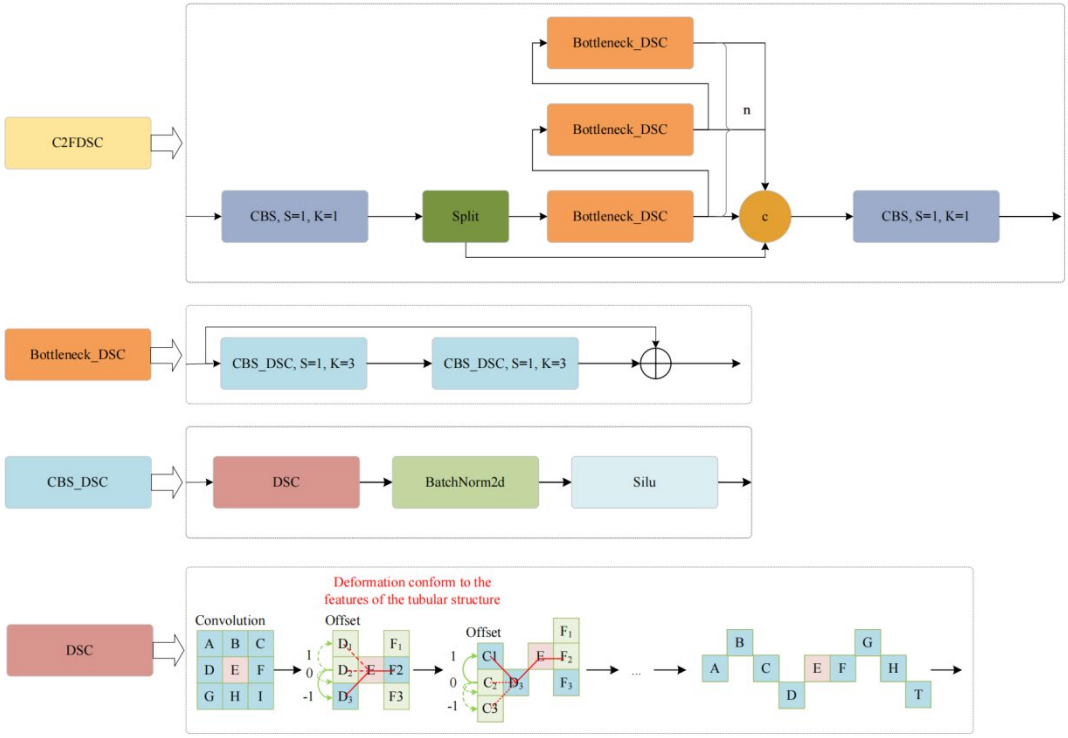

Figure 3: Schematic Diagram of the C2FDSC Module.

To provide greater flexibility to the convolutional kernels and enable them to focus on the complex geometric features of the target, DSC introduces a deformation offset, denoted as Δ. However, if the model is allowed to freely learn the deformation offset, the receptive field may shift away from the target. To address this issue, an iterative strategy is adopted. In this approach, the next position of each target to be processed is sequentially selected for observation. This ensures the continuity of attention and prevents the receptive field from expanding excessively due to large deformation offsets, which could lead to an over-diffusion of the perception area. The standard convolution kernel is straightened along the x and y axes. For example, in the case of a convolution kernel F of size 9 along the x-axis, the specific position of each grid in F is represented as: $F_i \pm k=(x_i \pm k, y_i \pm k)$, where $k = \{0, 1, 2, 3, 4\}$ represents the horizontal distance from the center grid. The se1lection of each grid $F_i \pm k$ in the convolution kernel F is an accumulated process. Starting from the center position F_i, the position F_{i+1} further from the center grid depends on the added offset $\Delta = \{\theta \mid \theta \in [-1,1]\}$ relative to F_i. Therefore, the offset needs to be accumulated, denoted as $\sum$, to ensure that the convolution kernel conforms to a linear structural form. Along the x-axis is transformed as:

$$F_{i\pm k}=\begin{cases}(x_{i+k}, y_{i+k}) = (x_i + k, y_i + \sum_i^{i+k} \Delta y), \\ (x_{i-k}, y_{i-k}) = (x_i - k, y_i + \sum_{i-k}^{i} \Delta y),\end{cases} \tag{1}$$

Along the y-axis direction, it becomes:

$$F_{j\pm k}=\begin{cases}(x_{j+k}, y_{j+k}) = (x_j + \sum_j^{j+k} \Delta x, y_i + k), \\ (x_{j-k}, y_{j-k}) = (x_j + \sum_{j-k}^{j} \Delta x, y_i - k),\end{cases} \tag{2}$$

Due to the variations in the two dimensions (x-axis and y-axis), the DSC covers a 9 × 9 range during the deformation process.

3.3. EACF module

Feature maps contain different feature information in each channel, but convolutional layers primarily compute features from adjacent positions within each map, without considering inter-channel interactions.

ISSN 2616-5775 Vol. 8, Issue 3: 116-125, DOI: 10.25236/AJCIS.2025.080316

Inspired by hyperspectral denoising[30], this paper proposes the Enhanced Attention Convolution Fusion (EACF) module to suppress background noise and enhance the network's attention to small objects.EACF is an enhanced module that integrates convolution operations and attention mechanisms. It consists of a local branch, a global branch, DropPath regularization, GELU activation functions, and shortcut connections, as shown in Figure 4.

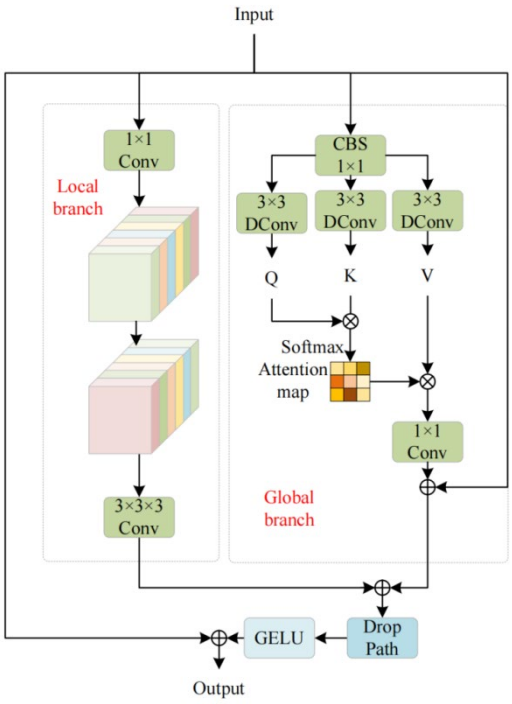

Figure 4: Schematic diagram of the EACF module structure.

In the local branch, convolution and channel rearrangement are used to extract local features and focus on the detailed information in the image. First, a 1×1 convolution is used to adjust the channel dimension, followed by a channel shuffling operation to divide the input tensor into multiple groups along the channel dimension. Depthwise separable convolutions are then applied within each group to shuffle the channels. The output tensors of each group are concatenated along the channel dimension to generate a new output tensor. Finally, a 3×3×3 convolution is used to extract features. The local branch can be represented as:

$$\mathrm{P_{local}} = \mathrm{K_{3\times3\times3}}(\mathrm{CS}(\mathrm{K_{1\times1}}(\mathrm{I}))) \tag{3}$$

In the global branch, an attention mechanism is used to capture long-range dependencies and focus on the global context information in the image. First, three tensors of shape $\hat{H}\times\hat{W}\times\hat{C}$ are generated using a 1×1 convolution and a 3×3 depthwise separable convolution, which serve as the query (Q), key (K), and value (V), respectively. Next, the dimensions of Q and K are rearranged to generate $\hat{Q}\in R^{\hat{H}\hat{W}\times\hat{C}}$ and $\hat{K}\in R^{\hat{C}\times\hat{H}\hat{W}}$. This reduces the computational cost when calculating the attention map for interaction. The output of the global branch computation can be obtained through the following steps:

$$P_{global} = K_{1\times1}Attention(\hat{Q},\ \hat{K},\ \hat{V}) + I \tag{4}$$

Then, after performing a concatenation operation on the outputs of the two branches, the DropPath technique and GELU activation function are applied, and a shortcut residual connection is introduced, resulting in the output computed by the EACF module:

$$\mathrm{P_{output}} = \mathrm{G}(\mathrm{D}(\mathrm{P_{local}} + \mathrm{P_{global}})) + \mathrm{I} \tag{5}$$

ISSN 2616-5775 Vol. 8, Issue 3: 116-125, DOI: 10.25236/AJCIS.2025.080316

4. Experiments

4.1. Dataset

The VisDrone-DET2019 dataset [31] was collected and released by the AISKYEYE team from the Machine Learning and Data Mining Laboratory at Tianjin University. It contains a total of 8,629 images, as shown in Figure 5. Among them, 6,471 images are used for training the model, 548 images are used for model validation, and 1,610 images are used to test the model's performance. This dataset includes 10 different categories of everyday scenes, specifically: pedestrians, people, bicycles, sedans, vans, trucks, tricycles, canopy tricycles, buses, and motorcycles.

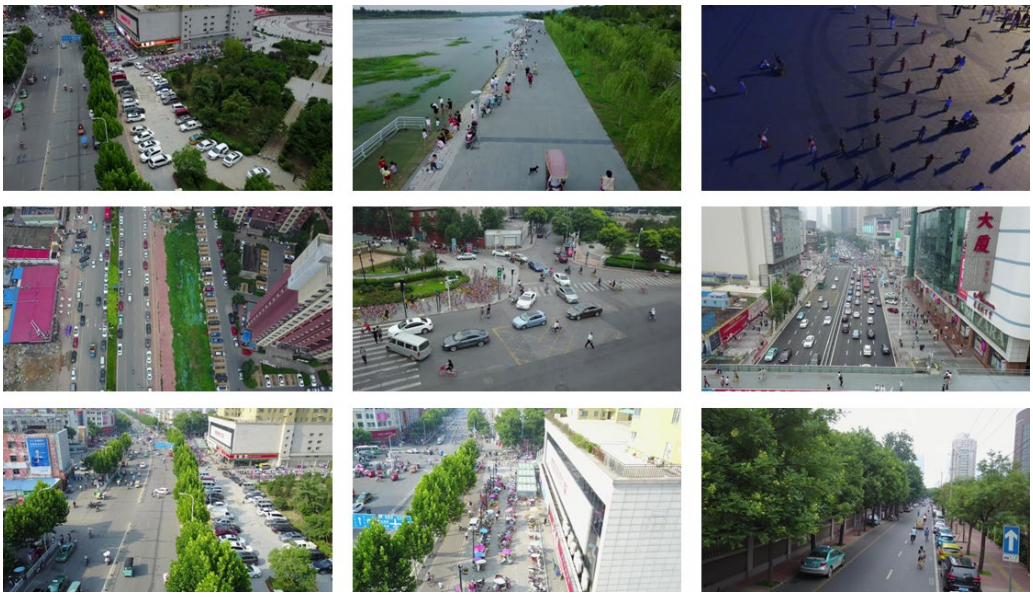

Figure 5: VisDrone-DET2019 dataset.

4.2. Evaluation Metrics

In this experiment, we use Precision (P), Recall (R), mean Average Precision (mAP) for each category, number of model parameters (Parameters), Frames Per Second (FPS), and GFLOPS as evaluation metrics to assess the model's performance.

4.3. Ablation Experiments

To clearly demonstrate the improvement in the model's detection capability, ablation experiments were conducted on the VisDrone dataset. The experiments were performed on the YOLOv8s network model, with the proposed modules and methods added progressively.The experimental results are detailed in Table 1.

Table 1: Ablation Experiment Results on VisDrone dataset

Number	Methods	P(%)	R(%)	mAP50(%)	mAP50-95(%)	Params(M)	FPS
Exp.1	Baseline	49.8	38.8	39	23.2	11.1	1118.52
Exp.2	Baseline+TDLayer	54.3	41.7	43.7	26.5	10.6	713.81
Exp.3	Baseline+C2FDSC	51.3	39.4	40.6	24.3	13	258.94
Exp.4	Baseline+EACF	51.1	38.7	39	23.7	12.8	768.12
Exp.5	Baseline+TDLayer +C2FDSC	55.3	42.8	44.9	27.3	12.5	256.55
Exp.6	Baseline+TDLayer +C2FDSC+EACF	55.5	43.5	45.4	27.9	14.2	244.89

The YOLOv8 model, as shown in Table 1, demonstrates significant improvements in small object detection accuracy with the introduction of various enhancements. Exp.1 presents the baseline YOLOv8s network, while Exp.2 adds a small object detection layer, resulting in a 4.7% improvement in mAP50 with minimal increase in parameters. This indicates that the modified detection head is better at

ISSN 2616-5775 Vol. 8, Issue 3: 116-125, DOI: 10.25236/AJCIS.2025.080316

preserving small object features, reducing missed and false detections. Additionally, the TDLayer improves detection resolution while reducing channel numbers, optimizing efficiency and reducing network redundancy. Despite an increase in layers, the overall number of parameters decreased, maintaining high detection performance and computational efficiency. In Exp.3, replacing the C2F module with the C2FDSC module led to improvements in detection accuracy, recall rate, and mAP50, with increases of 1.5%, 0.6%, and 1.6%, respectively. This suggests the C2FDSC module adapts better to varying object shapes and sizes. Exp.4 introduced the EACF (Attention and Convolution Fusion) module, improving detection accuracy by 1.3%, with a slight decrease in recall rate (0.1%) and a 0.9% increase in mAP50. The EACF module strengthens important features during extraction and reduces interference, enhancing overall detection performance. Exp.5 combined the small object detection layer with the C2FDSC module, achieving a 5.9% improvement in mAP50. The combined effect of these modules outperforms individual implementations. Finally, Exp.6 integrated all three modules into the YOLOv8-Plus algorithm, resulting in a 5.7% improvement in accuracy, a 4.7% increase in recall rate, and a 6.4% rise in mAP50. Although FPS decreased slightly, it remained above 30 frames per second, meeting real-time detection requirements. The YOLOv8-Plus algorithm thus achieved a balance between mAP and FPS, significantly enhancing small object recognition and detection capabilities.

4.4. Comparative Experiments

To validate the YOLOv8-Plus model's performance, we compared it with popular object detection algorithms on the VisDrone dataset (see Table 2).

Table 2: Comparative Experimental Results on VisDrone dataset

Methods	P(%)	R(%)	mAP50(%)	mAP50-95(%)	Parameters(M)	GFLOPS
RetinaNet	24.2	18.9	18.8	10.8	61	145
SSD	42.8	27.9	26.7	14.9	26.3	62.8
Faster-RCNN	36.5	27.7	28.7	16.5	41.2	206.7
YOLOv3-tiny	38.3	25	23.9	13.3	12.1	19.1
YOLOv5n	44	32.4	32.4	18.6	2.5	7.2
YOLOv5s	50.1	38.1	39.1	23.2	9.1	24.1
YOLOv6n	39.7	31.2	30.3	17.7	4.2	11.9
YOLOv6s	47.7	37.3	37.2	22.1	16.3	44.2
YOLOv7-tiny	48.6	37.5	35.7	18.6	6	13.3
YOLOv8n	44.5	32.8	33	19.2	3	8.2
YOLOv8s	49.8	38.8	39	23.2	11.1	28.7
YOLOv8m	53.2	41.9	42.5	25.9	25.9	79.1
YOLOv9c[32]	56.3	42.8	44	27	25.5	103.7
YOLOv8-Plus	55.5	43.5	45.4	27.9	14.2	46.7

The results show that RetinaNet, SSD, and Faster-RCNN perform 26.6%, 15.6%, and 15.8% worse than YOLOv8-Plus, respectively, while having significantly higher parameter counts and GFLOPS. While YOLO models maintain a favorable parameter count for small object detection, their accuracy still lags behind. The mAP50 of YOLOv3-tiny, YOLOv5n, YOLOv5s, YOLOv6n, YOLOv7-tiny, and YOLOv8n are 8.2% to 26.6% lower than YOLOv8-Plus. YOLOv8m and YOLOv9c have mAP50 values only 2.9% and 1.4% lower, but their parameter counts and GFLOPS increase significantly (by 11.7M and 11.3M parameters, and 32.4 and 57 GFLOPS, respectively), which can slow inference speeds. Overall, YOLOv8-Plus outperforms other models in various metrics, especially in instance segmentation. Its small parameter count and low computational burden make it ideal for practical applications on resource-constrained mobile devices or drones, giving it strong potential for future object detection tasks.

4.5. Visualization Analysis

Figure 6 compares the detection performance of the proposed algorithm and the baseline model across various scenarios, including well-lit, dimly lit, complex backgrounds, and dense target environments.In the first set, the baseline model misses many objects, while the proposed algorithm successfully detects distant pedestrians, motorcycles, and occluded tricycles and bicycles. In the second set, with dense targets and dim lighting, the baseline model both misses and misidentifies objects, such as confusing a toy with a motorcycle. In the third set, the proposed algorithm detects densely packed motorcycles and distant pedestrians, whereas the baseline misses several objects. In the final set, in low-light conditions, the baseline model fails to detect a motorcycle hidden under a tree, while the proposed algorithm performs

ISSN 2616-5775 Vol. 8, Issue 3: 116-125, DOI: 10.25236/AJCIS.2025.080316

well in detecting objects in such environments.

Figure 6: The comparison of detection results on the VisDrone dataset: (a) column shows the original image, (b) column shows the detection results using YOLOv8s, and (c) column shows the detection results using YOLOv8-plus.

Overall, the proposed algorithm outperforms the baseline, particularly in detecting small and occluded objects, and reduces missed and misdetections, providing better localization and detection for small-object detection. Figure 6 compares the detection performance of the proposed algorithm and the baseline model in various scenarios, including well-lit, dimly lit, complex backgrounds, and dense target environments.

5. Conclusion

In this paper, we address the issue of small objects occupying a small pixel area in images, which often leads to missed detections and susceptibility to noise. We propose an improvement to the popular YOLOv8 algorithm, adapting it for small object detection, resulting in the YOLOv8-Plus algorithm. Firstly, we introduce a new small object detection layer, TDLayer, into the neck network, allowing the model to fully leverage both shallow spatial location information and deep high-level semantic information for small object detection. Secondly, we integrate the C2FDSC module, designed in this study, enabling the model to flexibly capture the shapes and boundaries of small objects during training. Finally, we introduce the attention convolution fusion module, EACF, which allows the model to focus more on the target region, suppressing noise and ignoring irrelevant background information. Experimental results on the VisDrone 2019 dataset show that the model not only achieves significant improvements in detection accuracy but also demonstrates stronger generalization capability, effectively handling object detection tasks in complex environments. However, there are still some limitations in the proposed detection algorithm, such as minor missed detections for extremely dense and small objects. Although model accuracy has improved, the increase in model size and complexity has led to longer training and inference times. Future research will focus on optimizing the computational efficiency of the model, such as by introducing lighter network architectures or enhancing the parallel processing capability of the algorithm to improve real-time detection performance. Additionally, to address the challenges of small object detection, exploring more advanced feature extraction techniques and loss function designs may further improve the robustness and accuracy of the model.

ISSN 2616-5775 Vol. 8, Issue 3: 116-125, DOI: 10.25236/AJCIS.2025.080316

References

[1] Wang Q , Ye G , Chen S W F .A UAV perspective based lightweight target detection and tracking algorithm for intelligent transportation[J].complex & intelligent systems, 2025, 11(1). DOI:10.1007/s40747-024-01687-7.

[2] Zhang Y Z J .An improved tiny-yolov3 pedestrian detection algorithm[J]. Optik - International Journal for Light and Electron Optics, 2019.

[3] Chen W, Huang H , Peng S ,et al.YOLO-face: a real-time face detector[J].Springer Berlin Heidelberg, 2021(4).DOI:10.1007/s00371-020-01831-7.

[4] Liu S , Li J .EC-PFN: a multiscale woven fusion network for industrial product surface defect detection[J]. complex & intelligent systems, 2025, 11(1).DOI:10.1007/s40747-024-01699-3.

[5] Reis D H D, Welfer D , Cuadros M A D S L ,et al.Mobile Robot Navigation Using an Object Recognition Software with RGBD Images and the YOLO Algorithm[J].Applied Artificial Intelligence [2025-04-04]. DOI:10.1080/08839514.2019.1684778.

[6] Li M, Chen Y, Zhang T, et al. TA-YOLO: a lightweight small object detection model based on multi-dimensional trans-attention module for remote sensing images[J]. Complex & Intelligent Systems, 2024, 10(4): 5459-5473.

[7] Chen F, Ding Q, Hui B, et al. Multi-scale kernel correlation filter algorithm for visual tracking based on the fusion of adaptive features[J]. Acta Optics, 2020, 40: 109-120.

[8] Zhang H, Zhang J, Zhong X, et al. MSM-TDE: multi-scale semantics mining and tiny details enhancement network for retinal vessel segmentation[J]. Complex & Intelligent Systems, 2025, 11(1): 114.

[9] Girshick R , Donahue J , Darrell T ,et al.Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation[J].IEEE Computer Society, 2014.DOI:10.1109/CVPR.2014.81.

[10] Girshick R .Fast R-CNN[J].Computer Science, 2015.DOI:10.1109/ICCV.2015.169.

[11] Ren S , He K , Girshick R ,et al.Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J].IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6):1137-1149.DOI:10.1109/TPAMI.2016.2577031.

[12] 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.

[13] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer International Publishing, 2016: 21-37.

[14] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 7263-7271.

[15] Redmon J, Farhadi A. Yolov3: An incremental improvement[J]. arXiv preprint arXiv:1804.02767, 2018.

[16] ocher G. YOLOv5. GitHub code repository. Available at: https://www.github.com/ultralytics/yolov5.2022.

[17] Li C, Li L, Jiang H, et al. YOLOv6: A single-stage object detection framework for industrial applications[J]. arXiv preprint arXiv:2209.02976, 2022.

[18] Wang C Y, Bochkovskiy A, Liao H Y M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 7464-7475.

[19] Jocher, G., Chaurasia, A., & Qiu, J. Ultralytics YOLO (Version 8.0.0) [Computer software]. https://github.com/ultralytics/ultralytics.2023.

[20] Qiao S, Chen L C, Yuille A. Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 10213-10224.

[21] Hong M, Li S, Yang Y, et al. SSPNet: Scale selection pyramid network for tiny person detection from UAV images[J]. IEEE geoscience and remote sensing letters, 2021, 19: 1-5.

[22] Wang M, Yang W, Wang L, et al. FE-YOLOv5: Feature enhancement network based on YOLOv5 for small object detection[J]. Journal of Visual Communication and Image Representation, 2023, 90: 103752.

[23] Li Y, Zeng J, Shan S, et al. Occlusion aware facial expression recognition using CNN with attention mechanism[J]. IEEE transactions on image processing, 2018, 28(5): 2439-2450.

[24] Tang X, Du D K, He Z, et al. Pyramidbox: A context-assisted single shot face detector[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 797-813.

[25] Hu H, Gu J, Zhang Z, et al. Relation networks for object detection[C]//Proceedings of the IEEE

ISSN 2616-5775 Vol. 8, Issue 3: 116-125, DOI: 10.25236/AJCIS.2025.080316

conference on computer vision and pattern recognition. 2018: 3588-3597.
[26] Wang H, Wang J, Bai K, et al. Centered multi-task generative adversarial network for small object detection[J]. Sensors, 2021, 21(15): 5194.
[27] Liu J, Li C, Liang F, et al. Inception convolution with efficient dilation search[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 11486-11495.
[28] Rabbi J, Ray N, Schubert M, et al. Small-object detection in remote sensing images with end-to-end edge-enhanced GAN and object detector network[J]. Remote Sensing, 2020, 12(9): 1432.
[29] Qi Y, He Y, Qi X, et al. Dynamic snake convolution based on topological geometric constraints for tubular structure segmentation[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2023: 6070-6079.
[30] Hu S, Gao F, Zhou X, et al. Hybrid convolutional and attention network for hyperspectral image denoising[J]. IEEE Geoscience and Remote Sensing Letters, 2024.
[31] Du D , Zhu P , Wen L ,et al.VisDrone-DET2019: The Vision Meets Drone Object Detection in Image Challenge Results[C]//ICCV visdrone workshop.2019.DOI:10.1109/ICCVW.2019.00030.
[32] Wang C Y, Yeh I H, Mark Liao H Y. Yolov9: Learning what you want to learn using programmable gradient information[C]//European conference on computer vision. Cham: Springer Nature Switzerland, 2024: 1-21.