Proceedings of the 5th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/108/2025.20813

Enhancing Small Target Detection in Aerial Imagery Based on YOLOv8

Cheng Huang[1,2,a,*], Beihai Tan[1,b,*]

[1]*School of Integrated Circuits, Guangdong University of Technology, Guangzhou, 510006, China*
[2]*Advanced Manufacturing School, Guangdong Polytechnic of Environmental Protection Engineering, Guangzhou, 510500, China*
a. atchessgd@163.com, b. bhtan@gdut.edu.cn
**corresponding author*

Abstract: Among the numerous deep learning-based object detection algorithms, the YOLO (You Only Look Once) series has become a preferred choice for various detection scenarios due to its fast detection speed, high accuracy, and strong adaptability. However, its performance remains suboptimal when applied to unconventional images, such as aerial images, which often feature complex content and small detection targets. To overcome this limitation, a novel model called YOLOv8-TDE (Tiny Detection Enhanced) is introduced in this work. First, to more effectively differentiate between key features and noise and to comprehensively capture multi-scale target features, the feature extraction network is improved by employing pooling convolution kernels of various sizes and incorporating a lightweight attention mechanism. Second, the GeoIoU (Geometric Intersection over Union) loss function is introduced to reduce sensitivity to aspect ratio and center point distance, addressing the limitations of the original CIoU loss function, which is overly sensitive to small changes in these parameters. Finally, a novel detection head, the FAD Head, is proposed to dynamically generate detection head parameters based on the input image's features, enabling better feature extraction for targets of different sizes in complex scenes. This enhancement improves the model's adaptability across various scenarios, as well as detection accuracy and stability. Experiments on the VisDrone2019 dataset demonstrate that the proposed model outperforms the original YOLOv8n, achieving a 15.5% improvement in mAP@0.5 and a 17.6% improvement in mAP@0.5:0.95.

Keywords: YOLOv8, UAV, small target detection, attention mechanism.

1. Introduction

Thanks to advancements in science and technology, drones have found widespread applications in fields such as aerial photography, environmental monitoring, communication relay, disaster rescue, agricultural protection, and logistics [1]. However, aerial images often exhibit large backgrounds, complex content, and small targets that are prone to occlusion by objects such as buildings or trees. These challenges, combined with the influence of lighting and weather conditions, make small object detection in aerial imagery a critical area of focus in computer vision [2].

Deep learning has significantly advanced object detection, leveraging its powerful feature extraction capabilities and data-driven learning paradigm. Convolutional neural networks (CNNs), in

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/108/2025.20813

particular, have emerged as the core framework driving performance improvements in object detection tasks [3-7]. Current methods are generally divided into two groups: single-stage models, like OverFeat [8] and SSD [9], that prioritize speed and simplicity; and two-stage models, such as R-CNN [10] and SPPNet [11], which emphasize precision but incur high computational costs. Despite their success, these models are often optimized for natural scenes, leading to suboptimal performance when applied to aerial imagery.

To address these limitations, recent studies have concentrated on enhancing methods for detecting small objects in aerial images. For example, Betti et al. [12] proposed YOLO-S, a lightweight network optimized for speed, albeit with limited feature extraction capabilities. Yang et al. [13] enhanced YOLOv5 by adding a specialized layer for detecting tiny objects and fine-tuning the IoU, achieving improved accuracy on the VisDrone2019 dataset, though at the cost of increased model size. Wang et al. [14] embedded a small target detection structure and a global attention mechanism into YOLOv8, resulting in higher accuracy but with a significant increase in parameter counts. Ma et al. [15] developed SP-YOLOv8s by substituting traditional cross-layer convolutions with a new type of convolution, improving feature fusion. This modification resulted in higher mAP but also increased computational complexity. Similarly, Zhu et al. [16] enhanced YOLOv8n with SPD-Conv and an EMA attention mechanism, improving mAP@50 but falling short in fine-grained localization (mAP@50:95).

Building on these advancements, this paper proposes an improved YOLOv8n model specifically designed for small object detection in aerial imagery, addressing key limitations through the following contributions:

- E-SPPF: A multi-scale pooling strategy is introduced to better capture small object features, complemented by a lightweight attention mechanism to dynamically adjust feature weights. The use of multiple pooling kernel sizes enhances feature diversity and comprehensiveness, enabling the model to more effectively distinguish key features from noise, thereby improving precision and robustness in feature extraction.
- GeoIoU : A novel loss function is employed to reduce sensitivity to minor changes in aspect ratios and center points. By leveraging pseudo-differential techniques, GeoIoU improves localization accuracy and stability, particularly for small objects, by addressing the limitations of the original CIoU loss function.
- FAD Head: A flexible detection head is proposed to dynamically adapt to input features. This design enhances the model's adaptability across different scenarios and provides more precise detection by incorporating scale awareness, spatial awareness, and task awareness. These capabilities allow the model to better analyze object details, resulting in higher detection accuracy and stability in complex environments.

These improvements collectively enable more accurate small object detection while addressing the unique challenges posed by aerial imagery.

2. Overview of the YOLOv8 Algorithm

YOLO [17], introduced in 2016, quickly gained prominence in the field because of its capacity to provide high detection accuracy while maintaining fast processing speeds [18]. This study focuses on YOLOv8, which offers different scales, ranging from smaller to larger configurations. All variants follow a unified architecture but vary in their depth and width, with these dimensions expanding progressively through the different configurations.

One of YOLOv8's key innovations is the replacement of all C3 modules with C2f modules [19]. This modification introduces additional skip connections, which enhance feature fusion across

DOI: 10.54254/2755-2721/108/2025.20813

different feature layers. As a result, the gradient flow is enriched throughout the backbone and neck networks, providing a robust foundation for efficient learning and feature extraction.

YOLOv8 also employs a bidirectional Path Aggregation Network (PANet) structure [20], which significantly improves feature extraction efficiency. The bidirectional design reduces the information transmission path between the top and bottom layers, facilitating efficient information flow. This structure allows higher layers to access low-level features more effectively, enabling greater interaction and integration of feature information across network layers. Consequently, the model demonstrates an improved capacity for perceiving and analyzing complex features.

During the prediction phase, YOLOv8 introduces an innovative head design that separates the tasks of classification and regression. This design minimizes task interference, enhancing the accuracy of both. Furthermore, YOLOv8 replaces the traditional anchor-based approach with an anchor-free strategy, allowing the model to locate target objects more flexibly and accurately. This anchor-free design simplifies the detection process and improves recognition capabilities, particularly for complex scenarios.

For classification tasks, YOLOv8 employs binary cross-entropy to evaluate the differences between predicted and true labels. For regression tasks, the model employs a combination of Distribution Focal Loss (DFL) and Complete Intersection over Union (CIoU) loss [21]. This combination accounts for factors such as bounding box location, shape, and distribution, enabling precise quantification and optimization of regression errors. These enhancements contribute to more accurate regression predictions, thereby boosting the model's effectiveness in object detection.

With these innovations in network modules, feature extraction, prediction structure, and loss functions, YOLOv8 represents a significant advancement in object detection, achieving high accuracy and efficiency.

3. Methodology of the Proposed Algorithm

3.1. Improvement of the SPPF Layer

In YOLOv8, the Spatial Pyramid Pooling - Fast (SPPF) layer plays a pivotal role in efficiently integrating features, serving as an effective implementation of spatial pyramid pooling. However, in practical applications—particularly for aerial imagery—challenges arise due to the increasing down-sampling ratio and repeated down-sampling operations. These operations gradually lead to a reduction in small-object details within the feature maps. Aerial images typically contain a variety of small objects, whose features are easily weakened or even overlooked during the hierarchical transmission across network layers. Consequently, the model struggles to distinguish important features from complex background noise, which adversely affects detection accuracy.

To address this issue, this study proposes an enhanced version of the SPPF layer, termed E-SPPF. The E-SPPF layer utilizes multi-scale pooling kernels along with an attention mechanism, enhancing the efficiency of small-object feature extraction. By incorporating multi-scale pooling, the layer effectively captures features of objects at different scales, while the attention mechanism dynamically adjusts feature weights, enabling the model to focus on key features and suppress background noise. The structures of the original SPPF layer and the proposed E-SPPF layer are illustrated in figure 1.

Through the incorporation of multi-scale pooling, the E-SPPF layer captures a broader range of feature details across varying object scales. Simultaneously, the attention mechanism dynamically adjusts feature weights, emphasizing critical features while suppressing background noise. This synergy enables the model to preserve small-object features more effectively during feature integration, resulting in improved accuracy and robustness in object detection for aerial imagery.

DOI: 10.54254/2755-2721/108/2025.20813

3.1.1. Multi-Scale Pooling Kernels

Due to the varying target sizes and scales in aerial imagery, the single pooling kernel size used in the original SPPF model presents significant limitations. A single pooling size struggles to balance the fine-grained details of small objects with the broader context required for large objects, which severely constrains the model's generalization capability and adaptability.

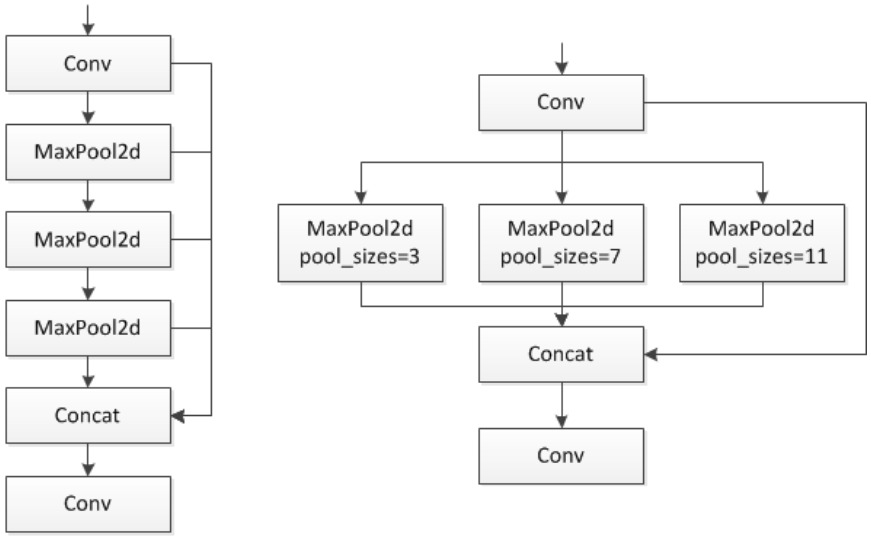


Figure 1: Structural diagram of the original SPPF layer(Left) and E-SPPF layer(Right).

To address this limitation, the proposed E-SPPF layer incorporates a multi-scale pooling kernel strategy. Specifically, three carefully selected pooling kernel sizes (3×3, 7×7, and 11×11) are utilized to construct a multi-dimensional feature extraction framework:

- 3×3 pooling kernel: Focuses on capturing fine-grained features of tiny targets. This is especially useful for aerial images, where many small targets require accurate feature extraction.
- 7×7 pooling kernel: Targets the extraction of contextual information for medium-sized objects, enriching the surrounding information to support accurate object recognition and localization.
- 11×11 pooling kernel: Primarily captures broad contextual information for large objects and backgrounds, helping the model understand the relationship between objects and their environment from a macro perspective.

This collaborative multi-scale pooling method allows the model to capture features from multiple layers and scales simultaneously. It notably increases the model's detection range for various sizes of objects, while also boosting its sensitivity and accuracy, particularly for small-object features.

3.1.2. Attention Mechanism

Aerial imagery often features highly complex and variable backgrounds, with targets displaying diversity in shape, color, and size. These characteristics impose higher demands on the model's feature extraction and representation capabilities. To further enhance the model's feature extraction efficiency, adaptability, and performance in challenging detection settings, this research integrates a carefully designed attention mechanism into the E-SPPF layer. Figure 2 demonstrates the configuration of this component.

The attention module dynamically adjusts the weights of the feature representations, prioritizing key features while filtering out irrelevant noise. By integrating linear transformations, activation functions, convolutional operations, and normalization techniques, the mechanism effectively

DOI: 10.54254/2755-2721/108/2025.20813

highlights important spatial regions, improving the model's capacity to detect and accurately locate targets. This ensures reliable detection results even in complex backgrounds with multiple interferences.

By combining multi-scale pooling kernels and the attention mechanism within the E-SPPF layer, the improved model efficiently captures features of objects at varying scales and emphasizes critical information in dynamic environments. These enhancements significantly improve detection accuracy and overall performance, providing a practical solution for small-object detection in aerial imagery.

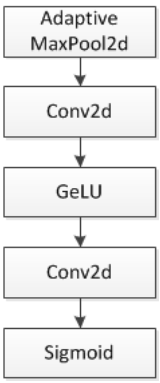


Figure 2: Attention Module

3.2. GeoIoU(Geometric IoU)

The original YOLOv8 utilizes CIoU as its loss function. CIoU, as an advanced loss function, is widely used in object detection models. It comprehensively accounts for the intersection over union (IoU), the offset of bounding box center points, and the consistency of aspect ratios. Additionally, it introduces penalty terms for center point distance and aspect ratio consistency, enabling a more holistic evaluation of bounding box quality. However, CIoU exhibits several shortcomings:

- Insufficient localization accuracy: CIoU, by considering multiple factors such as overlapping area, center point distance, and aspect ratio, involves a complex computation process. This complexity hampers its precision in small-object localization tasks. Specifically, for smaller objects, the distance penalty term becomes relatively weak, limiting its ability to ensure precise localization for small targets.
- Sensitivity to minor variations: Even slight changes in the aspect ratio or center point of small objects can result in significant variations in the CIoU value. This heightened sensitivity negatively impacts the stability of evaluation results, leading to noticeable fluctuations during training. These fluctuations hinder stable model optimization and reduce the reliability of bounding box assessments.
- Suboptimal performance with occlusion and overlap: CIoU struggles to effectively handle common issues in aerial imagery, such as occlusion and overlap. Small objects in aerial images are often prone to occlusion or overlap with one another. In such scenarios, CIoU fails to optimize and adjust bounding boxes effectively, resulting in significantly reduced detection accuracy.

To address these challenges, this study proposes GeoIoU (Geometric IoU), which eliminates the reliance on the center point of bounding boxes in CIoU. Instead, it innovatively introduces the concept of proportional distance squared sum (i.e., the proportional distance differences at the two diagonal corner points between the predicted and ground truth bounding boxes), creating a more accurate constraint mechanism that enhances the model's ability to adapt to bounding boxes of varying sizes and shapes.

DOI: 10.54254/2755-2721/108/2025.20813

Additionally, GeoIoU draws inspiration from the essence of MPDIoU proposed by Siliang Ma et al. [22], incorporating refined adjustments to the width and height parameters. These adjustments enable a more accurate calculation and correction of the loss, enhancing the stability and reliability of results. Through this series of optimizations, GeoIoU achieves better alignment between predicted and ground truth bounding boxes in various application settings, especially in challenging environments with a wide range of target sizes.

In addition, GeoIoU greatly simplifies the similarity comparison process between bounding boxes. This simplification enhances the model's robustness and adaptability in scenarios involving complex or atypical overlapping situations, enabling stable and reliable detection outputs. The main principles of GeoIoU are as follows:

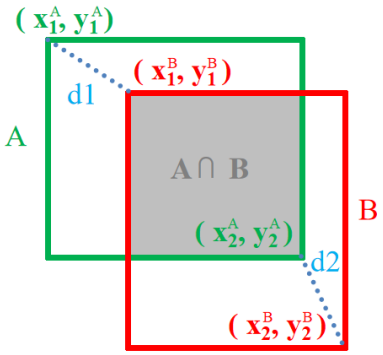


Figure 3: The relationship between the ground truth bounding box and the predicted bounding box

$$GeoIoU = IoU - \frac{d_1^2 + d_2^2}{w_p^2 + h_p^2} - \frac{S}{4} - \alpha \times v \tag{1}$$

Among them:

The ground truth bounding box is represented by 'A', while 'B' corresponds to the predicted bounding box.

$$IoU = \frac{A \cap B}{A \cup B} \tag{2}$$

d_1^2 and d_2^2 are the squared Euclidean distances between the top-left corner and the bottom-right corner of the detection bounding box and the ground truth bounding box respectively.

$$d_1^2 = \left(x_1^B - x_1^A\right)^2 + \left(y_1^B - y_1^A\right)^2 \tag{3}$$

$$d_2^2 = \left(x_2^B - x_2^A\right)^2 + \left(y_2^B - y_2^A\right)^2 \tag{4}$$

- "w" is the width of the predicted or actual box.
- "h" is the height of predicted box or real box.
- *"S"* is the sum of squared proportional distances, calculated using the following formula:

$$S = \left(\frac{x_1^B - x_1^A}{w_A}\right) + \left(\frac{y_1^B - y_1^A}{h_A}\right) + \left(\frac{x_2^B - x_2^A}{w_A}\right) + \left(\frac{y_2^B - y_2^A}{h_A}\right) \tag{5}$$

v is the consistency term for aspect ratio, calculated as follows:

DOI: 10.54254/2755-2721/108/2025.20813

$$v = \left(\frac{4}{\pi^2}\right) \bullet \left(\arctan\left(\frac{w_A}{h_A}\right) - \arctan\left(\frac{w_B}{h_B}\right)\right)^2 \tag{6}$$

α is an adjustment parameter:

$$\alpha = \frac{v}{1 + v - IoU + \text{eps}} \tag{7}$$

Due to the value range of v being [0~4), a very small non-zero parameter value named "eps" has been added to the denominator to avoid division errors during the operation.

3.3. Introducing the Dynamic Detection Head

In traditional object detection networks, detection heads are typically designed with fixed structures. This rigidity limits their flexibility and adaptability when handling targets of varying sizes, complex backgrounds, and diverse detection scenarios. This issue is particularly evident in aerial imagery, where the complexity and diversity of content make it difficult for fixed detection heads to fully leverage multi-scale feature information. As a result, detection performance, especially for small objects and in complex backgrounds, is often suboptimal. To address this challenge, inspired by Dynamic Head [23], this study proposes a novel dynamic detection head, FAD Head (Fast-Adaptive Detection Head), aimed at significantly enhancing the model's adaptability to diverse targets and scenarios while improving detection accuracy.

Aerial images exhibit unique complexity, often containing objects of vastly different scales within the same frame. For instance, typical aerial scenes may include large-scale objects such as buildings and buses, alongside smaller targets like people and animals. To address this characteristic, the proposed detection head incorporates a scale-aware attention mechanism, effectively addressing the challenges of coexisting multi-scale targets.

Moreover, since aerial images are captured by drones from varying altitudes and angles, objects often appear in drastically different shapes due to height variations and perspective effects. To handle this issue, the detection head is enhanced with a spatial-aware attention mechanism, improving the model’s ability to adapt to shape variations and refine feature extraction precision.

Finally, to boost overall performance in scenarios requiring high classification accuracy, a task-aware attention mechanism is integrated. This mechanism allows the model to more effectively grasp the requirements of detection tasks and optimize its performance on specific objectives, leading to a noticeable improvement in detection precision.

Unlike traditional fixed-structure detection heads, the proposed FAD Head has the remarkable capability to adaptively adjust feature selection and weight allocation based on the characteristics of the input image. This dynamic adaptability enables precise detection of targets across varying scales, significantly enhancing detection accuracy. The core advantage of the FAD Head lies in its unique mechanisms for dynamic feature selection and weighting, comprising the following key components:

3.3.1. Multi-Scale Feature Aggregation

The FAD Head efficiently extracts information from different feature levels and dynamically selects the most suitable features according to the detection task's needs. By employing a multi-scale feature extraction strategy, it ensures that features of various scales are effectively captured and aggregated, enabling the model to accurately detect targets of varying dimensions.

DOI: 10.54254/2755-2721/108/2025.20813

3.3.2. Dynamic Routing

The dynamic routing mechanism flexibly adjusts the weight distribution across feature channels based on the input image's characteristics. By emphasizing important channels and suppressing irrelevant noise, this mechanism significantly improves detection accuracy and stability, enabling the model to reliably identify target features even in complex environments.

3.3.3. Adaptive Feature Weighting

By integrating attention mechanisms, this component dynamically adjusts the weights of feature maps across channels. The adjustment process combines both channel and spatial attention, enhancing the representation of target regions while suppressing background noise. This allows the model to focus more effectively on critical target features, improving the overall accuracy and reliability of detection. In the specific implementation, the proposed model embeds the FAD Head module into the detection head part. The overall structure of YOLOv8-TDE is illustrated in figure 5.

Given the concatenated features $F_{in} = \{f_i\}_{i=1}^{L}$ at various levels within the feature pyramid, upsampling or downsampling operations are employed to adjust the features at different levels to match the scale of the median-level features. The rescaled feature pyramid can be represented as a 4D tensor $F \in R^{L \times H \times W \times C}$, where L denotes the number of levels in the pyramid, while H,W and C correspond to the height, width, and channel dimensions of the intermediate features, respectively. By setting S = H×W, the tensor can be reshaped into a 3D tensor $F \in R^{L \times S \times C}$. The role of each dimension of the tensor is as follows:

①Cross-Level Feature Enhancement. Features across different pyramid levels capture information distributed at various object scales. Strengthening the interaction and representation of F across different levels can better capture the global characteristics of large objects and the local details of small objects, thereby improving the model's adaptability to multi-scale targets.

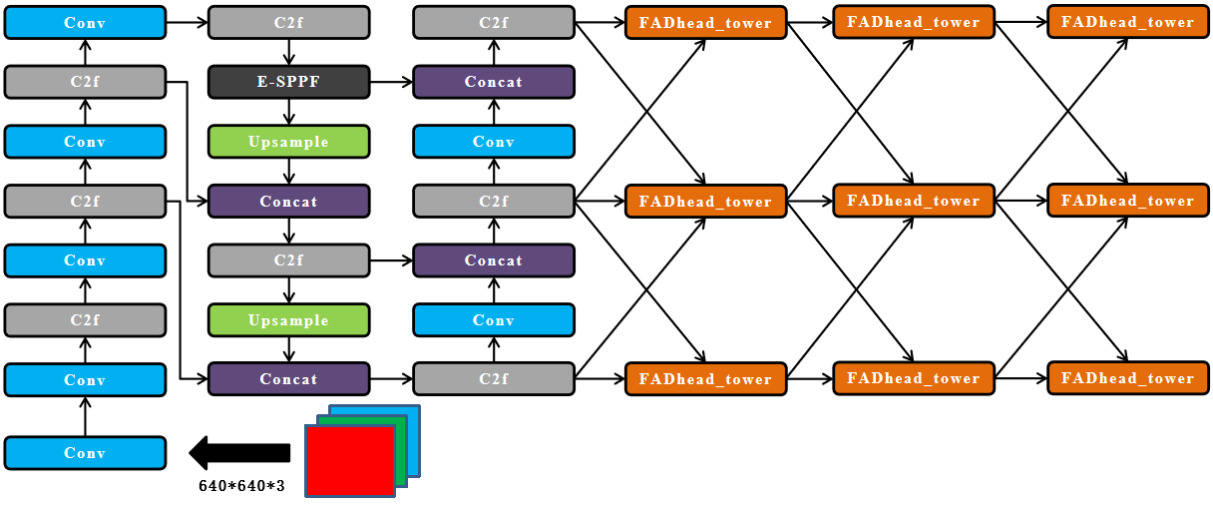


Figure 4: YOLOv8-TDE

②Spatial Region Perception Optimization. The feature map's spatial dimensions encompass regions that may include both relevant and irrelevant information about the target. By optimizing F for feature representation at different spatial locations, the model can better focus on key areas of the target, reducing the interference of irrelevant background and improving detection stability and accuracy in complex scenes.

DOI: 10.54254/2755-2721/108/2025.20813

③Task-Specific Channel Allocation. Features along the channel dimension represent the semantic information needed for different tasks. By adjusting F in the channel dimension, the model can better meet the specific requirements of tasks such as classification and localization, improving task collaboration efficiency and overall detection performance.

For the feature tensor $F \in R^{L \times S \times C}$, the general formula for applying self-attention is:

$$W(F) = A(F) \cdot F \tag{8}$$

Here, $A(\ )$represents an attention function, and $A(F)$ denotes the application of this attention weight matrix to the input feature F. A simple approach to apply this attention mechanism is using fully connected layers. However, because of the tensor's high dimensionality, directly learning an attention function across all its dimensions would demand considerable computational resources. To address this, this study adopts a decoupled learning approach, which converts the attention function into three sequential attentions, each focusing on a single perspective:

$$W(F) = A_{\mathrm{C}}\left(A_{\mathrm{S}}\left(A_{\mathrm{L}}(F) \cdot F\right) \cdot F\right) \cdot F \tag{9}$$

Here, A_{L}, A_{S} and A_{C} represent three distinct attention functions applied to dimensions L,S and C respectively.

In the cross-level feature attention mechanism $A_{\mathrm{L}}(\ )$,a scale-sensitive attention mechanism is introduced to adaptively combine features from different scales according to their semantic relevance.

$$A_{\mathrm{L}}(F) \cdot F = \sigma\left(\phi\left(\frac{1}{SC}\sum_{S,C} F\right)\right) \cdot F \tag{10}$$

Here, $\phi()$ is a linear function approximated by a 1×1 convolution layer, while $\sigma(\mathrm{x}) = \max\left(0, \min\left(1, \frac{x+1}{2}\right)\right)$ represents the Hard-Sigmoid function.

Next, a spatial region-aware attention module is applied to the integrated features to emphasize important regions that are present across both spatial positions and feature levels. Given the high dimensionality of S, we break this module into two steps: first, deformable convolution is used to facilitate sparse attention learning, and then cross-level features are aggregated at corresponding spatial locations.

$$A_{\mathrm{S}}(F) \cdot F = \frac{1}{L}\sum_{l=1}^{L}\sum_{k=1}^{K} \omega_{l,k} \cdot F\left(l; p_k + \Delta p_k; c\right) \cdot \Delta m_k \tag{11}$$

Where K is the number of sparse sampling locations, $p_k+\Delta p_k$ is a shifted location by the self-learned spatial offset Δp_k to focus on a discriminative region and Δm_k is a self-learned importance scalar at location p_k. Both are learned from the input feature from the median level of F.

To facilitate joint learning and capture diverse object representations, we apply a task-specific attention mechanism at the end. It selectively activates and deactivates feature channels to prioritize different tasks:

$$A_{\mathrm{C}}(F) \cdot F = \max\left(\alpha^1(F) \cdot F_{\mathrm{C}} + \beta^1(F), \alpha^2(F) \cdot F_{\mathrm{C}} + \beta^2(F)\right) \tag{12}$$

Where F_{C} is the feature slice at the c-th channel and $[\alpha^1, \alpha^2, \beta^1, \beta^2]^{\mathrm{T}} = \theta(\cdot)$ is a hyper function that learns to control the activation thresholds. $\theta(\cdot)$ is based on [24], where global average pooling is first applied to the L x S dimensions to reduce dimensionality, followed by two fully connected layers and a normalization layer. Finally, a shifted sigmoid function is used to normalize the output to the range of [-1, 1].

DOI: 10.54254/2755-2721/108/2025.20813

Ultimately, as the three attention mechanisms are applied in sequence, multiple A_L,A_S and A_C modules can be effectively stacked together through repeated nesting of Equation (7). Figure 6 illustrates the previously mentioned detection module.

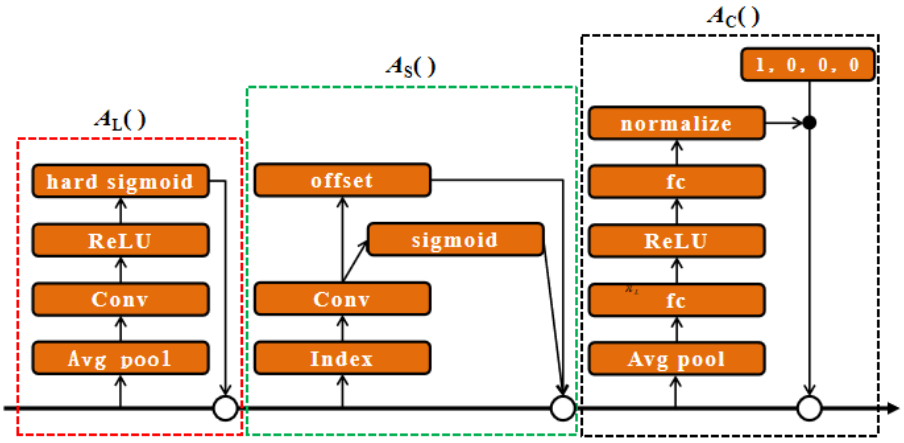


Figure 5: FAD Head module

By incorporating the dynamic detection module, the model's flexibility and adaptability are greatly improved, tackling major challenges in object detection. By leveraging a dynamic selection mechanism, the model can intelligently choose optimal feature maps for detection tasks, especially when handling small targets or complex backgrounds. This improves accuracy by focusing on the most valuable feature information, minimizing errors caused by poor feature selection or missed key features, and ensuring reliable detection results.

The adaptive feature weighting mechanism, as a core component, adjusts attention weights dynamically across channels and spatial dimensions. This enables the model to precisely capture and highlight target features while suppressing background interference. As a result, the quality of feature representation is enhanced, providing a solid foundation for subsequent classification and localization tasks.

Furthermore, the dynamic routing mechanism allows the model to adjust detection strategies based on input image characteristics, making it highly adaptable to diverse scenarios. Whether faced with large variations in target scale, occlusions, or cluttered backgrounds, the model can effectively align its detection strategy with the input, ensuring robust performance across various real-world applications.

These advantages collectively enable the dynamic detection head to enhance detection accuracy, optimize feature representation, and adapt to complex scenarios, making it a practical and robust solution for object detection in challenging environments.

4. Experiment And Analysis

4.1. Experimental Dataset

This study evaluates and validates the improved YOLOv8 model using the VisDrone2019 dataset [25]. The dataset is specifically designed for drone vision tasks, with a primary focus on object detection and tracking in aerial images. The dataset comprises real-world scene images captured by various drone cameras, covering a wide range of environments such as urban areas, rural areas, and highways.

The VisDrone2019 dataset consists of 8,599 images, which are split into training, validation, and test sets. It covers 10 object types, including bicycles, tricycle, van and people. Each image is

DOI: 10.54254/2755-2721/108/2025.20813

annotated with details about the target's position, dimensions, and shape, offering standardized data for model training and assessment.

The dataset is characterized by its high diversity and complexity, encompassing various weather conditions, lighting scenarios, shooting angles, and altitudes. This allows it to accurately simulate challenging scenarios, serving as an effective benchmark to assess the robustness and effectiveness of algorithms. As a publicly available dataset, VisDrone2019 is widely used in object detection, tracking, and remote sensing research, enabling researchers to test their algorithms, compare performance, and identify strengths and weaknesses for further optimization and refinement.

4.2. Ablation Experiment

To verify the effectiveness of the proposed E-SPPF, GeoIoU, and FAD head algorithms, YOLOv8n is used as the baseline model. Ablation experiments are conducted under identical conditions on the VisDrone2019 dataset to evaluate the impact of different module combinations on the model's small object detection performance. The test results for each improved module are summarized in table 1, where " √ " denotes the presence of improvement and "×" denotes its absence.

Table 1: Ablation experiment

Original model	E-SPPF	GeoIoU	FAD head	mAP50	mAP50:95	Total Inference Time/ms	FPS
√	×	×	×	0.348	0.204	5.8	172
√	√	×	×	0.355	0.207	5.9	169
√	×	√	×	0.356	0.208	5.8	172
√	×	×	√	0.393	0.234	8.3	120
√	√	√	×	0.358	0.209	6	167
√	√	×	√	0.397	0.236	8.4	119
√	×	√	√	0.397	0.236	7.7	130
√	√	√	√	0.402	0.24	8.1	123

Based on the experimental findings, it is clear that all three proposed modules, whether used individually or in combination, contribute significantly to the performance improvements of the baseline model. The key findings are as follows:

- E-SPPF layer improvement: When the E-SPPF layer is individually improved in YOLOv8n, the mAP@0.5 increases by 2%, and mAP@0.5:0.95 improves by 1.5%.
- GeoIoU loss function improvement: By replacing the CIoU with GeoIoU, the mAP@0.5 increases by 2.3%, and mAP@0.5:0.95 improves by 2.0%.
- FAD Head improvement: Integrating the FAD head achieves the most significant individual improvement, with mAP@0.5 and mAP@0.5:0.95 increasing by 12.9% and 14.7%, respectively.
- E-SPPF Layer + GeoIoU: When the E-SPPF layer and GeoIoU loss function are combined, mAP@0.5 increases by 2.8%, and mAP@0.5:0.95 improves by 2.4%.
- E-SPPF Layer + FAD Head: Combining the E-SPPF layer and FAD head results in an increase of 14.1% in mAP@0.5 and 15.6% in mAP@0.5:0.95.
- GeoIoU + FAD Head: When the GeoIoU loss function and FAD head are combined, mAP@0.5 and mAP@0.5:0.95 increase by 14.1% and 15.6%, respectively.

DOI: 10.54254/2755-2721/108/2025.20813

- All three modules: Integrating all three modules into YOLOv8n achieves the highest performance improvement, with mAP@0.5 increasing by 15.5% and mAP@0.5:0.95 increasing by 17.6%.

These findings indicate that the refined algorithm improves the model's capability to capture fine-grained details from remote sensing images, resulting in a substantial boost in detection accuracy.

Figure 7 and figure 8 illustrate the visualization results of the original YOLOv8n and the proposed improved model on the VisDrone2019 detection dataset, respectively. Through comparing these two sets of images, it is evident that the proposed model, in contrast to the original, not only minimizes false positives and false negatives but also achieves higher recognition accuracy.

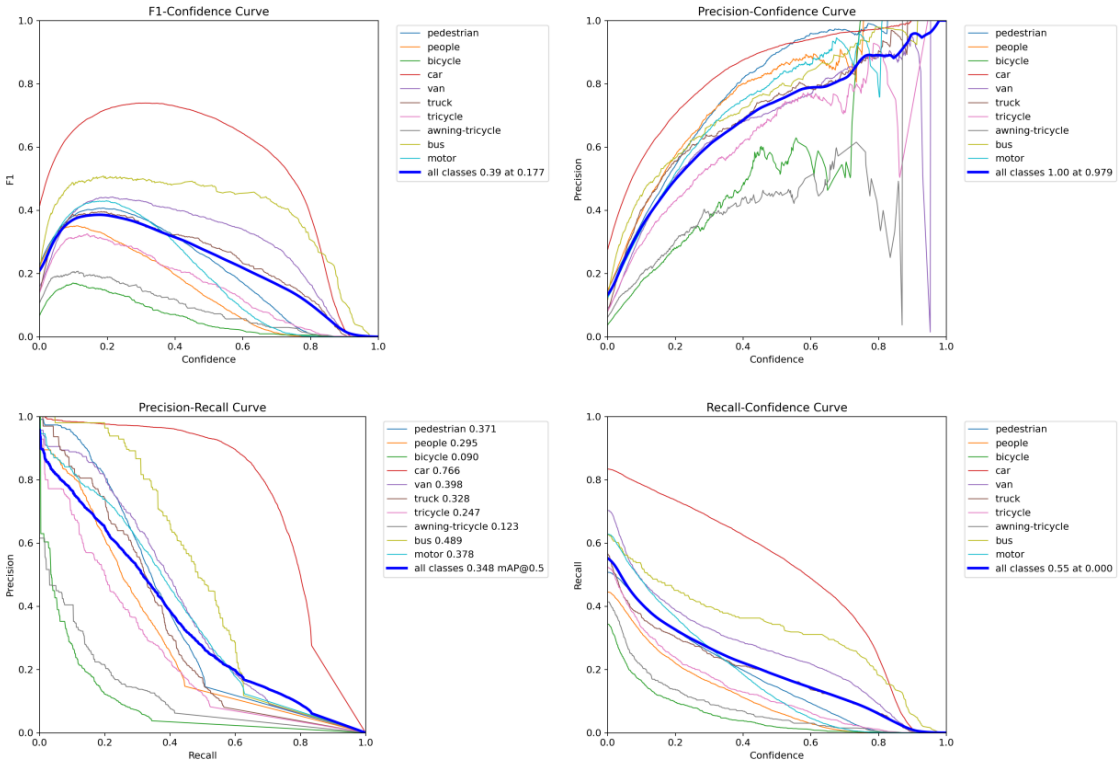
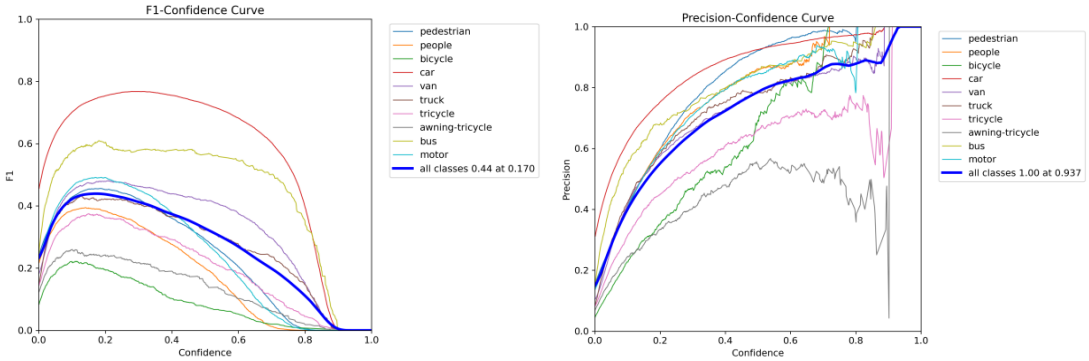


Figure 6: Performance evaluation curves of YOLOv8n. The curve types are labeled in each subplot.



DOI: 10.54254/2755-2721/108/2025.20813

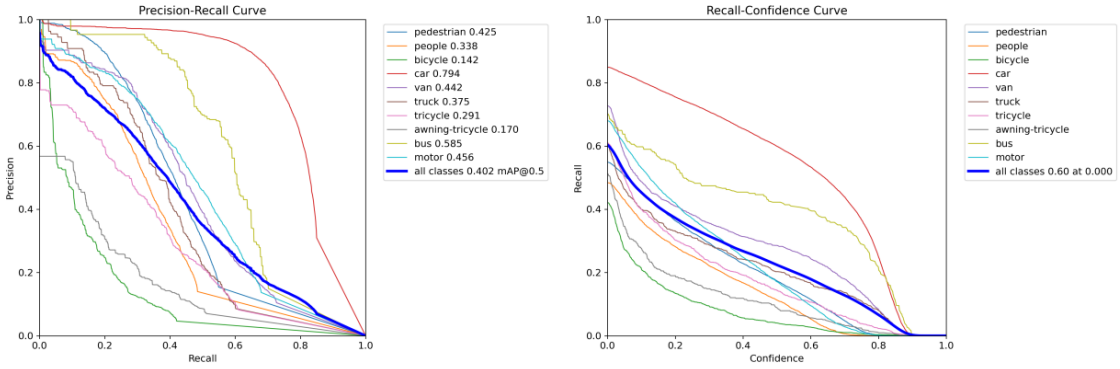


Figure 7: Performance evaluation curves of YOLOv8-TDE. The curve types are labeled in each subplot.

To further evaluate the model proposed in this study, figure 9 and figure 10 present the detection results of the original YOLOv8n and the improved model on aerial imagery.



Figure 8: Detection results of YOLOv8n.



Figure 9: Detection results of YOLOv8-TDE.

5. Conclusion

This study proposes an improved YOLOv8n model, which focuses on enhancing the detection accuracy of small objects in aerial images by innovatively introducing three core modules: the E-SPPF layer, FAD head, and GeoIoU. Rigorous experimental validation and in-depth analysis reveal that these improvements effectively optimize and upgrade the model's detection performance in complex aerial image scenarios.

Specifically, using the VisDrone2019 dataset, the enhanced model shows notable performance improvements, with mAP@0.5 and mAP@0.5:0.95 increasing by 15.5% and 17.6%, respectively, over the original YOLOv8n model. Each of the modules—E-SPPF, GeoIoU, and FAD head—demonstrates independent performance improvements when applied individually. Moreover, when these three modules are organically integrated and work synergistically, they yield even greater performance enhancements, validating their effectiveness in improving small-object feature extraction and precise detection.

The E-SPPF layer, through its multi-scale pooling kernels and attention mechanisms, overcomes the limitations of traditional feature extraction, enabling more comprehensive, precise, and efficient capture and refinement of small-object features. This greatly enriches the representation and transmission of small-object feature information.

DOI: 10.54254/2755-2721/108/2025.20813

The GeoIoU module, by replacing and optimizing the traditional CIoU loss function, fundamentally improves the accuracy and stability of geometric attribute matching for detection boxes. It effectively addresses many drawbacks and limitations of conventional loss functions in handling small objects and complex scenarios.

The FAD head, with its multi-dimensional attention mechanisms integrating scale-aware, spatial-aware, and task-aware attentions, comprehensively optimizes the model's adaptive detection capabilities for complex background interference and multi-scale object variations. This significantly enhances the model's robustness and accuracy in complex and dynamic aerial image environments.

In summary, this study contributes an effective approach to enhance detection performance for small targets in aerial images captured by drones, achieving notable advancements in detection precision. Future research could focus on further optimizing these modules and exploring their application potential in other visual tasks. This would expand the model's applicability within the field of computer vision, providing valuable technical solutions and insights for related research and practical applications.

Author Contributions

Conceptualization, C.H., and B.T.;Investigation, C.H., and B.T.;Writing original draft, C.H.;Writing review and editing, C.H., and B.T. All authors have read and agreed to the published version ofthe manuscript.

Funding

This research was funded by the 2024 Guangdong Graduate Education Innovation Project and the 2024 Key Research Platforms and Projects of Ordinary Universities in Guangdong Province (Grant Number: 2024ZDZX4152).

Data Availability Statement

The datasets used or analyzed during the current study are availablefrom the corresponding author upon reasonable request.

References

[1] *H. Shen, "Development and application of drones," Science News, vol. 25, no. 1, pp. 42–45, 2023. (in Chinese)*
[2] *S. Zhong and L. Wang, "A review of object detection technology in UAV aerial images," Advances in Laser and Optoelectronics [J/OL], pp. 1–32, Dec. 17, 2024. (in Chinese)*
[3] *K. Mo, L. Chu, and X. Zhang, "DRAL: Deep Reinforcement Adaptive Learning for Multi-UAVs Navigation in Unknown Indoor Environment," arXiv preprint arXiv:2409.03930, 2024.*
[4] *Y. Liu, J. Wu, S. Sun, X. Wang, and H. Wang, "Contrastive Self-supervised Learning in Recommender Systems: A Survey," arXiv preprint arXiv:2205.01593, 2022.*
[5] *T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated Learning: Challenges, Methods, and Future Directions," IEEE Signal Processing Magazine, vol. 37, no. 3, pp. 50–60, May 2020, doi: 10.1109/MSP.2020.2975749.*
[6] *T. Baltrušaitis, C. Ahuja, and L. P. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 2, pp. 423–443, Feb. 2019.*
[7] *J. Lü and Q. Fu, "A joint algorithm based on improved active learning and self-training," Journal of Beijing Normal University (Natural Science Edition), vol. 58, no. 1, pp. 25–32, 2022. (in Chinese)*
[8] *P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," arXiv preprint arXiv:1312.6229, 2013.*
[9] *W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, and C.-Y. Fu, "SSD: Single shot multi-box detector," in Computer Vision – ECCV 2016, Springer, 2016, pp. 21–37.*
[10] *R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 580–587.*

DOI: 10.54254/2755-2721/108/2025.20813

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.
[12] A. Betti and M. Tesei, "YOLO-S: A lightweight and accurate YOLO-like network for small target selection in aerial imagery," Sensors, vol. 23, no. 4, pp. 1865–1865, 2023.
[13] H. Yang and H. Li, "A single-aggregation YOLO algorithm for detecting small aerial targets," Foreign Electronics Measurement Technology, vol. 42, no. 4, pp. 131–140, 2023. (in Chinese)
[14] F. Wang, H. Wang, Z. Qin, and M. Zhang, "UAV target detection algorithm based on improved YOLOv8," IEEE Access, vol. 11, pp. 116534–116544, 2023.
[15] M. Ma and H. Pang, "SP-YOLOv8s: An improved YOLOv8s model for remote sensing image tiny object detection," Applied Sciences, vol. 13, no. 18, p. 8161, 2023, doi: 10.3390/app13148161.
[16] Y. Zhu and Y. Zhang, "SEP-YOLO: An improved YOLOv8-based road target detection algorithm," Computer Applications and Software, pp. 1–8. (in Chinese)
[17] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified real-time object detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 779–788.
[18] J. Zhou and J. Wang, "A review of YOLO object detection algorithms," Journal of Changzhou Institute of Technology, vol. 36, no. 1, pp. 18–23, 88, 2023. (in Chinese)
[19] B. J. Xiao, M. Nguyen, and W. Q. Yan, "Fruit ripeness identification using YOLOv8 model," Multimedia Tools and Applications, pp. 1–18, 2023.
[20] P. Liu, P. Dollár, K. He, R. Girshick, and P. Dollár, "Path aggregation network for instance segmentation," arXiv preprint arXiv:1803.01534, 2018.
[21] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Enhancing Geometric Factors in Model Learning and Inference for Object Detection and Instance Segmentation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9937–9946, doi: 10.1109/CVPR42600.2020.00996.
[22] S. Ma and Y. Xiong, "MPDIoU: A loss for efficient and accurate bounding box regression," arXiv preprint arXiv:2307.07662, 2023.
[23] X. Dai, Y. Chen, B. Xiao, D. Chen, L. Yuan, and L. Zhang, "Dynamic Head: Unifying Object Detection Heads with Attentions," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7373–7382, doi: 10.1109/CVPR46437.2021.00729.
[24] Y. Chen, X. Dai, M. Liu, D. Chen, L. Yuan, and Z. Liu, "Dynamic ReLU," arXiv preprint arXiv:2003.10027, 2020. [Online]. Available: https://arxiv.org/abs/2003.10027.
[25] Z. Zhu, L. Wen, D. Du, X. Bian, H. Ling, and Q. Hu, "Detection and Tracking Meet Drones Challenge," in Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), 2019, pp. 1901–1910, doi: 10.1109/ICCVW.2019.00244.