2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

STE-YOLO: A Dual Enhancement Architecture for Small Object Detection in Aerial Imagery

Wancheng He[1], Yihang Shen[1], Jun Wan[1], Peitao Wang[1], Haohan Ding[1,2,*], and Song Shen[1]

[1] School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, China
[2]Science Center for Future Foods, Jiangnan University, Wuxi, China
6233115010@stu.jiangnan.edu.cn

Abstract. Object detection in aerial imagery faces unique challenges due to small object scales, ambiguous textures, and dense distributions. Traditional detection methods often struggle with preserving structural information of small targets and effectively utilizing both global context and local details. To address these limitations, we propose STE-YOLO (Small Target Enhancement YOLO), featuring two key innovations: a Multi-level Content-Aware Feature Enhancement Module (MCAFE) that dynamically adjusts feature fusion strategies, and a Local-Enhance Global Attention (LEGA) module that effectively balances global context and local feature representation. Extensive experiments on VisDrone2019 dataset demonstrate that STE-YOLO significantly outperforms baseline models. Compared to YOLOv10, our method achieves improvements of 10.8% in mAP@0.5 and 11.0% in mAP@0.5:0.95 on VisDrone2019. Additionally, we conducted generalization experiments on DOTAv1.5 dataset, where our method also shows strong performance, demonstrating the robustness and adaptability of our approach across different aerial imagery scenarios while maintaining acceptable computational overhead.

Keywords: Small object detection, target recognition, Deep learning, Computer vision, YOLO.

1 Introduction

Object detection in aerial imagery has emerged as a crucial task in computer vision, with extensive applications in urban planning, environmental monitoring, and surveillance. With the rapid advancement of unmanned aerial vehicle (UAV) technology, the demand for aerial image information extraction has grown substantially [1]. High-precision object detection algorithms serve as the foundation for accurate object localization and tracking, driving the intelligent development of UAV systems [2]. However, compared to conventional natural image processing, aerial object detection faces unique challenges: limited recognition features, small object scales, ambiguous textures, dense object distributions, and complex backgrounds—all of which significantly challenge algorithm design and optimization[3,4].

In recent years, deep learning-based object detection algorithms have become predominant in this field [5]. These algorithms can be categorized into two-stage and single-stage approaches. Two-stage detection algorithms [7] first generate potential object proposals using region proposal techniques, followed by classification and regression [8]. In contrast, single-stage detection algorithms like YOLO and SSD employ a unified architecture that directly predicts object locations and categories[9,10].While single-stage approaches offer improved real-time performance, YOLO faces significant limitations when applied to aerial imagery. Standard convolutional operations and Feature Pyramid Networks [11] struggle to balance broad contextual information with crucial local details, leading to information loss in small object detection [12]. Furthermore, YOLO's feature extraction mechanisms inadequately capture complex spatial relationships in aerial imagery, particularly for small targets where the balance between global context and local details is critical. Although attention mechanisms [13,14] have been introduced to address these issues, they still exhibit limitations in modeling complex local spatial relationships and determining the relative importance of different regions.

To address these challenges, we propose Small Target Enhancement YOLO(STE-YOLO), an enhanced object detection algorithm specifically designed for aerial imagery. Our main contributions are threefold:

(1) We introduce a Multi-level Content-Aware Feature Enhancement Module (MCAFE) that dynamically adjusts feature fusion strategies based on input con-tent, significantly improving the model's capability in preserving structural in-formation of small objects.

(2) We design a novel Local-Enhance Global Attention (LEGA) module that ef-fectively balances global context and local feature representation, achieving su-perior performance in capturing both long-range dependencies and fine-grained details.

(3) Extensive experiments on VisDrone2019 datasets demonstrate the effective-ness of our approach in small object detection while maintaining computational efficiency. Specifically, STE-YOLO shows significant improvements compared to the baseline YOLOv10: on VisDrone2019, the improvements are 10.8% in mAP@0.5 and 11.0% in mAP@0.5:0.95;

2 Related Work

2.1 YOLO

The You Only Look Once (YOLO) model is widely recognized in object detection for its high accuracy and fast inference speed. Since its development in 2015[15], YOLO has evolved through multiple versions, each introducing significant improvements. YOLOv3 (2018) incorporated more efficient backbone structures [16], multiple anchors, and spatial pyramid pooling [17]. Recent versions have introduced innovations including mosaic data augmentation (YOLOv4)[18], practical deployment functions (YOLOv5)[19], and NMS-free detection architecture (YOLOv10)[20,21,22,23].

In aerial imagery applications, several specialized adaptations have emerged. FFCA-YOLO introduced modules for feature enhancement and context awareness to address small object detection challenges. LA-YOLO integrated the SimAM attention

mechanism and a fusion block with normalized Wasserstein distance to improve detection in low-altitude scenarios [24]. YOLO-SS incorporated an optimized backbone and restructured loss function specifically for small object detection [25]. More recently, ARF-YOLO integrated a coordinate-based attention module into YOLOv8, boosting both accuracy and speed [26].

Our work builds upon these foundations while addressing their limitations in small object detection through novel attention and feature enhancement mechanisms designed specifically for aerial imagery challenges.

2.2 Attention Mechanism

The attention mechanism was first applied to the NLP field by Bahdanau et al. in 2014, and it has been over a decade since then [30]. Now, the attention mechanism is widely used in many fields, including machine translation, speech recognition, and computer vision. In deep learning, attention mechanisms give models the ability to differentiate and focus on important parts of the input. There have been many variants in recent years. For example, SENet first introduced channel attention to explicitly model dependencies between channels [31]. CBAM further expanded on SENet's foundation by combining channel and spatial attention to learn "what" and "where" information separately [32]. GAM addressed the cross-dimensional interaction problem that traditional attention mechanisms overlooked, preserving more information by eliminating pooling operations [33]. SimAM explored a more lightweight direction by proposing a parameter-free attention mechanism, while EMA reduced computational complexity through expectation maximization algorithm.

3 Methodology

This chapter will introduce the overall architecture of the STE-YOLO algorithmic model and the improved modules we developed to enhance small object detection performance, including the Multi-level Content-Aware Feature Enhancement Module (MCAFE) and Local-Enhance Global Attention (LEGA). These innovative modules are designed to address the challenges faced by traditional object detection algorithms when processing small objects in aerial imagery, including insufficient feature representation, loss of target features, and inadequate utilization of contextual information. Through the organic integration of these modules, our model can significantly improve the detection performance of small objects.

3.1 Overall Structure of our Model

To balance detection accuracy and real-time performance, we choose YOLOv10 as the foundation architecture of STE-YOLO. STE-YOLO consists of three key components: a feature extraction backbone network, an enhanced neck network, and a detection head. In the backbone network, we adopt YOLOv10's architecture: utilizing C2f and

SCDown modules for feature extraction, and integrating SPPF and PSA modules at the end to enhance feature representation.

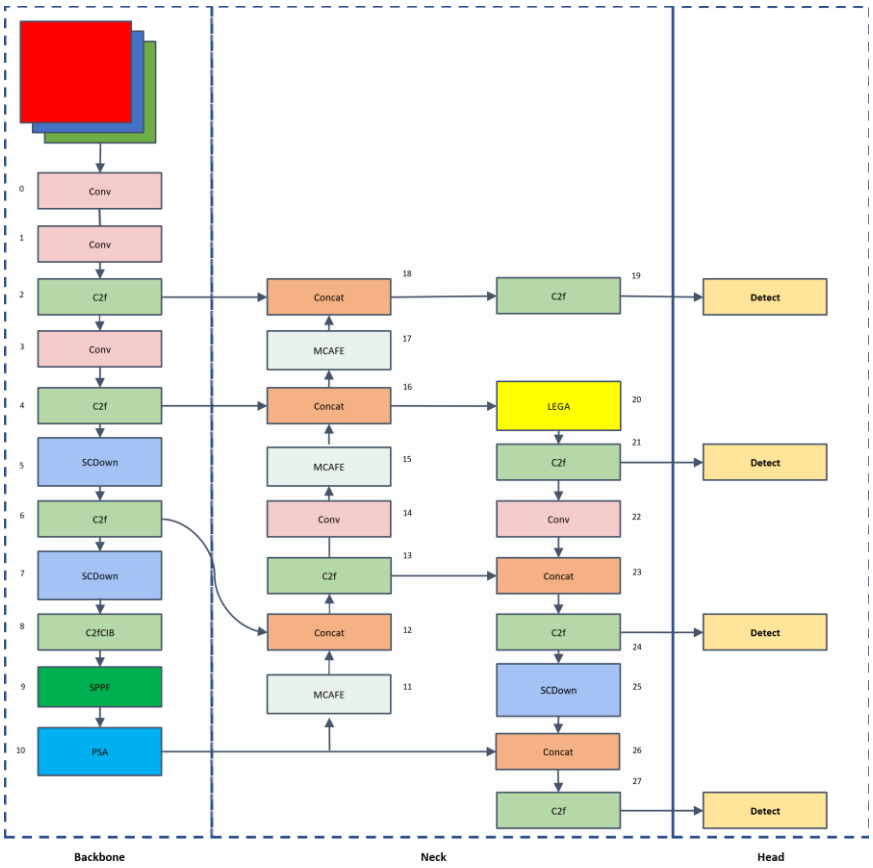

Fig. 1. Structure of STE-YOLO.

For the neck network, we have made significant improvemeEnhancents to the original structure. First, we introduce a custom-designed Local-enhance Global Attention Module, which significantly enhances the model's capability to detect small objects by processing global and local feature information in parallel. Additionally, we replace the traditional simple upsampling operation with a Multi-level Content-Aware Feature Enhancement Module, which improves feature representation through adaptive feature reorganization. The detection head employs the v8Detect structure and adds a lightweight tiny object detection head, responsible for transforming the enhanced multi-scale features into predictions of object location, category, and bounding box coordinates, achieving efficient and accurate object detection.

3.2 Local-Enhance Global Attention (LEGA)

Recent advances in computer vision have highlighted the importance of effectively balancing global and local feature representation. While convolutional neural networks (CNNs) excel at capturing local patterns, their limited receptive fields can hinder the modeling of long-range dependencies. Conversely, attention mechanisms are effective at capturing global context but may overlook crucial local details. To address this problem, inspired by GLSA\cite{b14}, we propose the Local-Enhance Global Attention (LEGA) module. Simultaneously, this module can effectively aggregate and represent both global and local spatial features, thereby enhancing model performance in both large-scale and small-scale target localization tasks through the enhancement of task-relevant information and suppression of redundant information.

As shown in

Fig. **2**, LEGA adopts a three-stream architecture design, including Local Stream, Global Stream, and Gate Stream. This three-stream architecture enables effective fusion of local-global features. The design simultaneously captures global and local spatial features while further optimizing feature representation and information flow. Through the incorporation of adaptive attention mechanisms and deep convolutional feature extraction, the architecture achieves an effective balance between model performance and computational efficiency. The overall structure can be expressed as:

$$LEGA(X) = X \cdot F\big(L(X), G(X)\big) \cdot T(X) \tag{1}$$

where the local spatial features L(X) are primarily modeled through the Local Stream, while the global spatial features G(X) are captured through the Global Stream. After non-linear transformation, these features are combined to form the fusion feature F(L(X),G(X)). The gating function T(X) is implemented as a simple sigmoid activation on the first channel of the input feature, which acts as a lightweight attention mechanism to modulate the feature response.

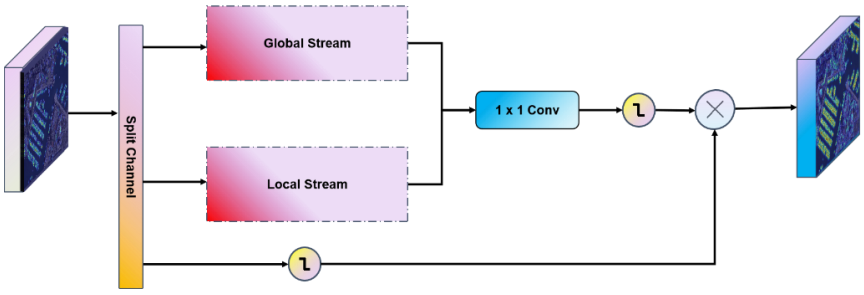

Fig. 2. Overview of the LEGA module.

Local Stream. Although the GLSA\cite{b14} module effectively captures global-local features, the Local Spatial Attention (LSA) component has inherent limitations in modeling local spatial relationships.To enhance the capture of local spatial dependencies,

we propose a Local Stream that performs weighted feature aggregation through SoftPooling operation. Unlike LSA which relies on cascaded convolutional layers with sigmoid activation, our approach implements an efficient pooling-based attention mechanism. The Local Stream first projects features through a 1×1 convolution to reduce channel dimensions, followed by SoftPooling operation that generates importance-weighted feature maps. Specifically, SoftPooling computes weighted averages within local regions using an exponential weighting scheme:

$$SoftPool(x) = \frac{\sum_{i \in R} x_i e^{x_i}}{\sum_{i \in R} e^{x_i}} \tag{2}$$

where x_i represents the input feature at position i, and R denotes the local pooling region. This formulation naturally emphasizes stronger feature responses while suppressing weaker ones through the exponential term e^{x_i}. The weighted features are then processed through two convolutional layers at reduced spatial resolution for computational efficiency, before being transformed back to the original channel dimensions. A final sigmoid activation normalizes the attention weights to [0,1]. Additionally, we incorporate a lightweight channel gate that operates on the first channel to modulate the attention response, helping to prevent over-emphasis of local patterns.

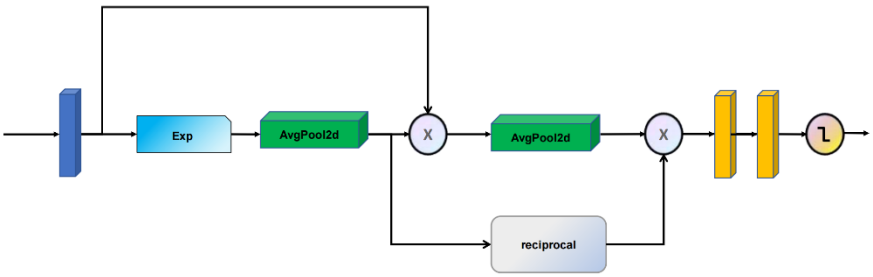

Fig. 3. Structure of Local Stream.

Global stream. Global stream is an adaptive feature enhancement mechanism, which optimizes feature representation through spatial attention and channel modulation. The unit includes two main steps: spatial context information acquisition and context-based feature enhancement.

Given input feature F_1, the feature first undergoes channel dimension adjustment:

$$F_g = \mathrm{Conv}_{1\times1}(F_1) \tag{3}$$

The spatial context information is obtained using an attention mechanism. The module first calculates a spatial attention map, then performs feature aggregation:

$$Context = \mathrm{MatMul}\left(X, \mathrm{Softmax}\left(\mathrm{Conv}_{1\times1}(F_g)\right)\right) \tag{4}$$

where X is the reshaped matrix of input feature F_g with dimensions $(N, C, H \times W)$. The attention mechanism achieves adaptive focus on important regions through learnable spatial weights.After obtaining the context information, the module optimizes feature representation through a channel multiplication branch:

$$G(F_1) = \sigma\big(\mathrm{MLP}(Context)\big) \odot F_g \tag{5}$$

where σ represents the sigmoid activation function and $\odot$ denotes element-wise multiplication. MLP represents a multi-layer perceptron, which is composed of two 1×1 convolution layers with LayerNorm and ReLU activation functions in the middle. To ensure training stability, the last layer of MLP adopts a zero initialization strategy, which makes the module's behavior closer to identity mapping in the early stages of training. Overall, this design not only effectively captures global context information but also enhances feature expressiveness through its feature enhancement strategy.

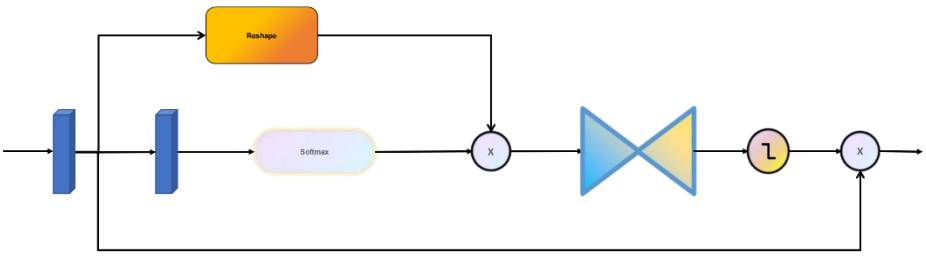

Fig. 4. Structure of Global Stream.

3.3 Multi-level Content-Aware Feature Enhancement Module

To address the limitations of traditional methods while simultaneously enhancing feature representation and fusion capabilities, we propose a Multi-level Content-Aware Feature Enhancement Module (MCAFE). This module not only improves the quality of feature upsampling through content-aware mechanisms, but also significantly enhances feature discrimination and semantic understanding through multi-level processing and channel attention. Unlike traditional methods that focus solely on spatial resolution recovery, MCAFE integrates dynamic feature processing, adaptive fusion, and attention enhancement in a unified framework. This comprehensive approach enables the module to better handle the complex scenarios in aerial imagery where objects exhibit diverse scales and dense distributions. The specific structure of the module is as follows:

Feature Adjustment Layer. Employs 1×1 convolution to adjust the channel dimension of input features, establishing an appropriate feature representation foundation for

subsequent processing. This lightweight preprocessing step can flexibly adjust the number of feature channels and optimize computational efficiency.

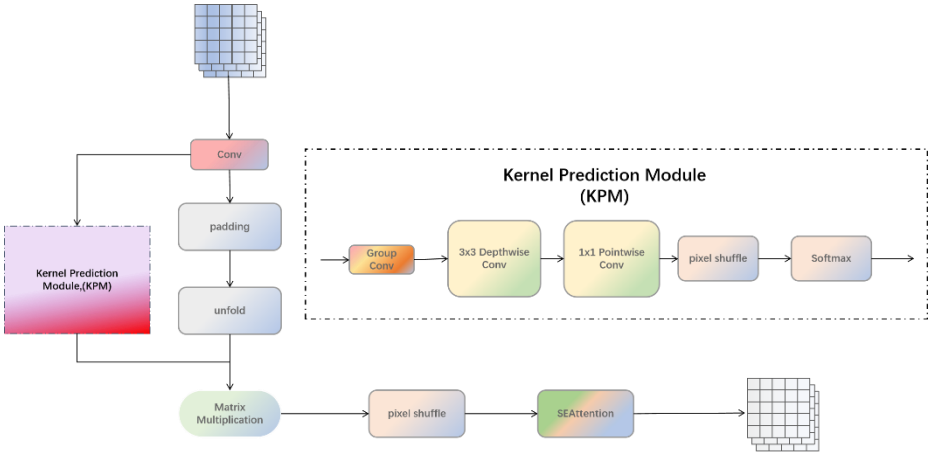

Fig. 5. Structure of Multi-level Content-Aware Feature Enhancement Module (MCAFE).

Content-Aware Feature Reorganization. As the core component of the module, it dynamically generates content-aware upsampling weights for adaptive weighted combination of neighborhood features at each pixel location. It specifically includes two key steps:

— Kernel Prediction: Predicts dynamic reorganization kernels through a lightweight convolutional neural network, which can be represented as:

$$\boldsymbol{K} = \text{Softmax}\mathcal{P}\left(\boldsymbol{F}_{\text{input}}\right) \in R^{k^2 \times H \times W} \tag{6}$$

where k is the reorganization kernel size, and $\mathcal{P}$ is the kernel prediction network.

— Feature Reorganization: Performs feature reorganization and upsampling using the predicted dynamic kernels:

$$\boldsymbol{F_{out}}(p) = \sum_{q \in \Omega_k(p)} \boldsymbol{K}(p,q) \cdot \boldsymbol{F}(q) \tag{7}$$

where p represents the target position, and $\Omega_k(p)$ denotes the $k \times k$ neighborhood centered at p.

Channel Attention Enhancement. Introduces a channel attention mechanism after feature reorganization, which captures inter-channel dependencies through global information modeling to adaptively adjust the weight distribution of feature channels. This design can highlight important features while suppressing secondary information, thereby improving the discriminative ability of features. Through the utilization of global contextual information, the attention mechanism can better understand and

leverage semantic associations between feature channels, further enhancing the effectiveness of feature representation.

The multi-level content aware feature enhancement module not only improves the spatial resolution of feature maps with high quality, but also significantly enhances the model's feature expression and discrimination ability through multi-level feature processing. Especially by introducing channel attention mechanism, the module can better capture and utilize the correlations between features. More importantly, by adopting a lightweight network structure and optimized computational implementation, this module maintains low computational overhead and has good practicality, enabling it to better cope with complex scenes and small object detection tasks in aerial images.

4 Experiments

4.1 Dataset and Metric

In this study, we primarily conducted experiments on the VisDrone 2019 dataset and performed a generalization test on the DOTAv1.5 dataset. The VisDrone 2019 dataset consists of 7,019 high-quality UAV-captured images (6,471 for training and 548 for validation). The objects are divided into ten categories: pedestrians, people, bicycles, cars, vans, trucks, tricycles, awning tricycles, buses, and motorcycles. To further evaluate our model's generalization capability to new scenarios, we conducted a generalization experiment on the DOTAv1.5 dataset, which contains 2,806 images with 188,282 instance annotations across 16 object categories (including aircraft, ships, storage tanks, and various facilities). By testing on these two datasets with different viewing angles, object scales, and scene complexities, we were able to comprehensively assess our algorithm's performance and adaptability in aerial imagery analysis.

We evaluate our model using several metrics, with mAP (Mean Average Precision) at two IoU thresholds (0.5 and 0.5:0.95) as our primary accuracy metrics. The mAP is calculated as:

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{8}$$

$$AP = \int_0^1 P(R)\ dR \tag{9}$$

where N is the number of classes, P denotes precision, and R represents recall. To assess model efficiency, we also measure parameter count (Params) and inference speed (FPS).

4.2 Implementation Details

All experiments were conducted in the Linux operating system environment, using NVIDIA A100 GPU equipped with CUDA 12.2. The software environment is based on Python 3.10 and PyTorch 2.0.1 framework to achieve efficient model training and inference. For training, we used a batch size of 8. The maximum number of training epochs was set to 250, with an early stopping strategy implemented - training terminates

when validation metrics show no improvement for 50 consecutive epochs.Following YOLOv8's default training strategy, we used pretrained weights on the MS COCO dataset to initialize our models. We employed the SGD optimizer with an initial learning rate of 0.01, momentum of 0.937, and weight decay of 0.0005. The learning rate was scheduled using a cosine annealing strategy that gradually decreases to 1e-4 by the end of training, with a 3-epoch linear warmup period at the beginning.For data augmentation, we utilized YOLOv8's comprehensive augmentation pipeline, which includes:

1. Mosaic augmentation: combining four training images into one
2. Random affine transformations: rotation (±10°), scaling (0.5-1.5), and translation (±10% of dimensions)
3. HSV color space augmentation: random adjustments to hue (±0.015), saturation (0.7-1.3), and value (0.4-1.6)
4. Random horizontal flips with probability 0.5
5. Mixup: blending two images with coefficients sampled from beta distribution B (8.0, 8.0)
6. Adaptive padding to maintain aspect ratio while resizing to 640×640 pixels

We used the default YOLOv8 loss function, which combines classification, objectness, and box regression losses with automatic balancing. The inference confidence threshold was set to 0.001 during validation for comprehensive evaluation, with NMS using an IoU threshold of 0.7. Most experiments followed YOLOv10's default parameter settings to maintain experimental reproducibility and comparability, with minor adjustments made for specific scenarios.

4.3 Ablation experiment

To evaluate the effectiveness of our proposed improvement strategies for drone aerial image detection tasks, we conducted a series of ablation experiments on the VisDrone2019 dataset using YOLOv10m as the baseline model. During the experimental process, we sequentially integrated the LEGA module into the neck of the YOLOv10m model, introduced the MCAFE module in the upsampling process, and adopted the YOLOv8 detection head to optimize the model's precise localization capability for tiny target objects. On this basis, we compared the specific impact of incorporating different modules on the detection accuracy of the baseline model. The experimental results are shown in Table 1.

Table 1. Ablation study of different components in STE-YOLO.

YOLO	MCAFE	LEGA	Head	mAP@0.5	mAP@0.5:0.95
√				0.418	0.254
√	√			0.422	0.258
√		√		0.433	0.264
√	√	√		0.438	0.268
√	√	√	√	0.463	0.283

4.4 Overall Performance

We conducted comprehensive experiments comparing our proposed STE-YOLO model against state-of-the-art YOLO variants (YOLOv5, YOLOv6, YOLOv8, YOLOv9, and YOLOv10) as well as Faster R-CNN and RT-DETR using the VisDrone2019 benchmark dataset. The evaluation metrics included mean Average Precision (mAP) at different thresholds (0.5 and 0.5:0.95), model parameters (Params), precision, recall, and inference speed in Frames Per Second (FPS).

Table 2. Performance comparison of different models on VisDrone2019 dataset.

Methods	Params(M)	mAP@0.5	mAP@0.5:0.95	Precision	Recall	FPS
Faster R-CNN	41.53	0.327	0.185	0.462	0.318	22.5
RT-DETR	32.97	0.453	0.275	0.541	0.428	78.0
YOLOv10	16.46	0.418	0.254	0.540	0.406	181.3
YOLOv9	20.02	0.439	0.268	0.556	0.422	147.0
YOLOv8	25.85	0.422	0.256	0.543	0.413	161.3
YOLOv6	51.98	0.411	0.249	0.530	0.395	153.8
YOLOv5	25.05	0.422	0.255	0.533	0.415	166.7
STE-YOLO	27.52	0.463	0.283	0.557	0.440	123.8

As shown in Table 2, our STE-YOLO demonstrates superior performance on the VisDrone2019 dataset compared to all tested models. Our model achieved the highest mAP@0.5 of 0.463 and mAP@0.5:0.95 of 0.283, surpassing the baseline YOLOv10 by significant margins of 10.8% and 11.4% respectively. Even when compared to RT-DETR, which showed the second strongest performance among the tested models, STE-YOLO still outperformed it by 2.2% in mAP@0.5 and 2.9% in mAP@0.5:0.95. Notably, our model also achieved the best detection quality metrics with the highest precision (0.557) and recall (0.440), indicating superior ability to reduce false positives while detecting more true targets. In terms of model efficiency, STE-YOLO has a moderate parameter count of 27.52M, comparable to YOLOv5 (25.05M) and YOLOv8 (25.85M). While YOLOv10 achieves the highest FPS at 181.3, our STE-YOLO maintains real-time processing capabilities at 123.8 FPS while significantly enhancing detection performance, making our approach suitable for practical aerial small target detection applications where accuracy is paramount.

4.5 Comparative experiment

As shown in Table 3, STE-YOLO demonstrates significant improvements in detecting small objects such as pedestrians and motor, while also showing enhanced detection capability for large objects. This improvement can be attributed to LEGA's ability to capture both global and local detail information.

To further validate the generalization capability of the STE-YOLO model, we conducted additional experiments on the DOTAv1.5 (Dataset for Object deTection in Aerial images v1.5) dataset. DOTAv1.5 is a large-scale aerial image object detection dataset, comprising 2,806 images and 188,282 instances, covering 16 common object categories. Compared to VisDrone2019, the DOTAv1.5 dataset features higher image

resolutions and greater variations in object scales, providing a challenging testing environment for our model.

Table 3. Performance comparison of different models on VisDrone2019 dataset.

Model	Target class (AP%)									
	Ped.	Person	Bike	Car	Van	Truck	Tri.	Awn. Tri.	Bus	Motor
YOLOv5	44.5	35.4	17.1	80.7	47.2	40.3	31.3	17.2	61.3	46.8
YOLOv6	42.9	33.9	13.1	80.2	47.1	40.5	30.9	17.5	58.8	46.7
YOLOv8	45.8	35.5	16.1	81.1	47.0	38.9	32.2	15.9	62.0	47.3
YOLOv9	46.5	36.1	19.4	81.4	48.8	43.1	34.6	18.5	61.5	49.5
YOLOv10	44.0	35.4	17.0	80.9	47.1	40.6	30.8	15.0	60.2	47.2
Ours	54.1	43.5	18.9	85.1	50.7	41.7	35.9	18.8	59.0	55.2

As shown in Table 4, STE-YOLO demonstrates excellent performance on the DOTA dataset, achieving an mAP of 41.7%, surpassing several benchmark models. This result proves that our proposed model not only excels in scenarios with densely distributed small objects (such as VisDrone2019) but also exhibits strong adaptability in processing aerial images with large-scale variations and complex backgrounds.

Table 4. Generalization performance of different models on DOTAv1.5 dataset.

Methods	Params(M)	mAP@0.5	mAP@0.5:0.95	FPS
YOLOv10	16.46	0.381	0.236	99.0
YOLOv9	20.02	0.408	0.256	62.1
YOLOv8	25.85	0.400	0.252	80.6
YOLOv6	51.98	0.375	0.233	61.7
YOLOv5	25.05	0.391	0.244	66.7
STE-YOLO	27.52	**0.417**	**0.259**	64.5

In order to verify LEGA's performance, we configured it on YOLOv10 for comparative experiments against other advanced attention mechanism modules. As shown in Table 5, our module improves the performance of YOLOv10 from baseline (mAP@0.5: 0.418, mAP@0.5:0.95: 0.255) to 0.433 and 0.264 respectively. Unlike SEAttention which primarily focuses on channel relationships through global average pooling, LEGA explicitly models spatial dependencies at both local and global levels through its dual-stream architecture. Compared to CBAM which applies sequential channel and spatial attention, our module processes these dimensions in parallel, preserving more spatial information critical for small objects. LEGA's unique advantage for small object detection lies in its SoftPooling operation, which preserves fine-grained details through exponential weighting of feature responses—particularly beneficial for small objects that occupy few pixels with limited distinctive features. Additionally, the dynamic fusion of local and global information helps distinguish small targets from complex backgrounds common in aerial imagery. The experimental results confirm these advantages, showing that LEGA outperforms other attention modules such as SEAttention (0.424/0.260), EMA (0.422/0.257), and CBAM (0.424/0.257).

Table 5. Ablation study of different attention mechanisms on YOLOv10.

Methods	mAP@0.5	mAP@0.5:0.95
YOLOv10	0.418	0.255
YOLOv10+GLSA	0.435	0.265
YOLOv10+SE	0.424	0.260
YOLOv10+EMA	0.422	0.257
YOLOv10+CBAM	0.424	0.257
YOLOv10+LEGA	0.433	0.264

4.6 Visualization comparison with baseline models

To visually demonstrate the detection performance of STE-YOLO, we selected a set of representative images from the VisDrone2019 test set. We used YOLOv10m as a baseline for comparison.

Fig. 6. Visualization results comparison between STE-YOLO and baseline models on low-altitude scenes from the VisDrone2019 dataset.

Fig. 6 illustrates the detection results in both crowded low-altitude scenes (left column) and sparse low-altitude scenarios (right column). The figure is organized in three rows: the top row shows the original input images, the middle row displays the results from the baseline model, and the bottom row presents the outputs of our proposed STE-

YOLO model. These results provide a vivid demonstration of our model's capabilities in handling objects of various types and scales. From densely parked small vehicles to large industrial facilities, STE-YOLO consistently exhibits accurate localization and classification, showing noticeable improvements over the baseline model across different scene complexities.

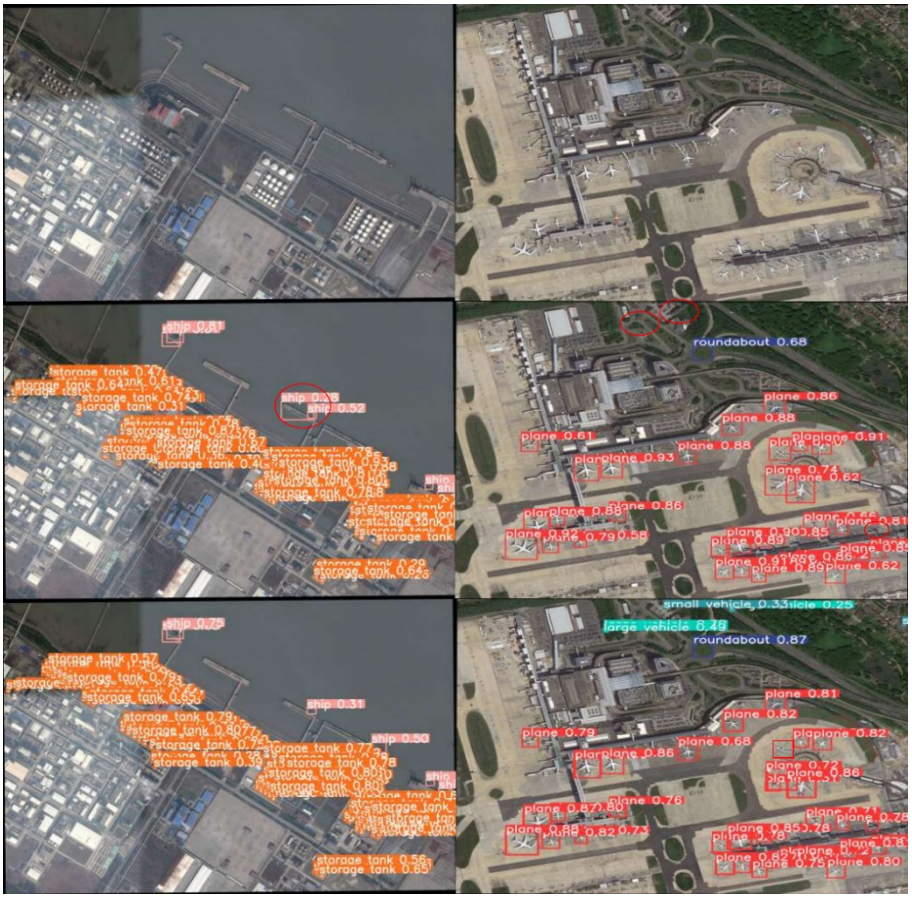

Fig. 7. Visualization results comparison between STE-YOLO and baseline models on the DATAV1.5 dataset.

We evaluated our method on representative images from the DOTAv1.5 test set, using YOLOv10m as baseline. Fig. 7 shows detection results in complex aerial scenes. Our model exhibits significant advantages in detecting densely distributed multi-scale objects, particularly for small and partially occluded targets, demonstrating stronger recall under complex backgrounds - a crucial capability for remote sensing applications.

5 Conclusions

In this paper, we proposed STE-YOLO, an enhanced object detection framework for aerial imagery. Our main contributions include two innovative modules: MCAFE and LEGA. MCAFE dynamically adjusts feature fusion strategies based on input content, improving small object detection through better structural information preservation. LEGA enhances feature representation through a three-stream architecture that balances global context and local details.

Experiments on VisDrone2019 and DOTAv1.5 datasets demonstrate the effec-tiveness of our approach. STE-YOLO achieves significant improvements over baseline models, with mAP@0.5 increasing from 0.418 to 0.463 on Vis-Drone2019 and from 0.381 to 0.417 on DOTAv1.5. While introducing moderate computational overhead, our model maintains practical inference speed for real-world aerial object detection applications.

References

1. Yu, C., Shin, Y.: MCG-RTDETR: Multi-Convolution and Context-Guided Network with Cascaded Group Attention for Object Detection in Unmanned Aerial Vehicle Imagery. Remote Sensing 16(17), 3169 (2024)
2. Lu, W.J., et al.: High-Quality Object Detection Method for UAV Images Based on Improved DINO and Masked Image Modeling. Remote Sensing 15(19), 4740 (2023)
3. Zhang, J., et al.: MBAB-YOLO: A Modified Lightweight Architecture for Real-Time Small Target Detection. IEEE Access 11, 78384–78401 (2023)
4. Zhang, Y.J., Gao, G.F., Chen, Y.D., Yang, Z.J.: ODD-YOLOv8: An Algorithm for Small Object Detection in UAV Imagery. Journal of Supercomputing 81(1), 202 (2025)
5. Zaidi, S.S.A., Ansari, M.S., Aslam, A., Kanwal, N., Asghar, M., Lee, B.: A Survey of Modern Deep Learning Based Object Detection Models. Digital Signal Processing 126, 103514 (2022)
6. Li, Z.H., et al.: Development and Challenges of Object Detection: A Survey. Neurocomputing 598, 128102 (2024)
7. Li, Z.S., et al.: Toward Effective Traffic Sign Detection via Two-Stage Fusion Neural Networks. IEEE Transactions on Intelligent Transportation Systems 25(8), 8283–8294 (2024)
8. Jiang, D., Li, G.F., Tan, C., Huang, L., Sun, Y., Kong, J.Y.: Semantic Segmentation for Multiscale Target Based on Object Recognition Using the Improved Faster-RCNN Model. Future Generation Computer Systems 123, 94–104 (2021)
9. Archana, R., Jeevaraj, P.S.E.: Deep Learning Models for Digital Image Processing: A Review. Artificial Intelligence Review 57(1), 11 (2024)
10. Hussain, M.: YOLO-v1 to YOLO-v8, the Rise of YOLO and Its Complementary Nature toward Digital Manufacturing and Industrial Defect Detection. Machines 11(7), 677 (2023)
11. Yu, X.X., Zhang, Y., Zhou, K.Q.: A Method for Detecting Small Target Weld Defects Based on Feature Reorganization Network. Measurement Science and Technology 36(1), 016046 (2025)
12. Lyu, J., Xu, P.: Image Super-resolution Reconstruction Algorithm Based on Multi-scale Adaptive Upsampling. Journal of Frontiers of Computer Science & Technology 17(4), 879–891 (2023)

13. Qi, X., Zhi, M.: Review of Attention Mechanisms in Image Processing. Journal of Frontiers of Computer Science & Technology 18(2), 345–362 (2024)
14. Tang, F., Huang, Q., Wang, J., Hou, X., Su, J., Liu, J.: DuAT: Dual-Aggregation Transformer Network for Medical Image Segmentation. arXiv:2212.11677 (2022)
15. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You Only Look Once: Unified, Real-Time Object Detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779–788 (2016)
16. Redmon, J., Farhadi, A.: YOLO9000: Better, Faster, Stronger. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6517–6525 (2017)
17. Redmon, J., Farhadi, A.: YOLOv3: An Incremental Improvement. arXiv:1804.02767 (2018)
18. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv:2004.10934 (2020)
19. Jocher, G., et al.: Ultralytics YOLOv5 (2023). https://github.com/ultralytics/yolov5
20. Huang, C.X., et al.: YOLOv6: A Single-Stage Object Detection Framework for Industrial Applications. arXiv:2209.02976 (2022)
21. Ge, C.R., et al.: YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. arXiv:2207.02696 (2022)
22. Jocher, G., Chaurasia, A.: YOLOv8: A Real-Time Object Detection Architecture (2023). https://github.com/ultralytics/ultralytics
23. Wang, A., Chen, H., Liu, L., et al.: YOLOv10: Real-Time End-to-End Object Detection. arXiv:2405.14458 (2024)
24. Ma, J., Huang, S.L., Jin, D.Y., Wang, X.Z., Li, L.C., Guo, Y.: LA-YOLO: An Effective Detection Model for Multi-UAV under Low Altitude Background. Measurement Science and Technology 35(5), 055401 (2024)
25. Tang, Q., et al.: YOLO-SS: Optimizing YOLO for Enhanced Small Object Detection in Remote Sensing Imagery. Journal of Supercomputing 81(1), 303 (2025)
26. Zeng, Y.L., Guo, D.J., He, W.K., Zhang, T., Liu, Z.T.: ARF-YOLOv8: A Novel Real-Time Object Detection Model for UAV-Captured Images Detection. Journal of Real-Time Image Processing 21(4), 107 (2024)
27. Chen, H., Wu, G., Li, J., Wang, J., Tao, H.: Research Advances on Deep Learning Recommendation Based on Attention Mechanism. Computer Engineering and Science 43(2), 370–380 (2021)
28. Ye, N.F., Zhang, L., Xiong, D., Wu, H., Song, A.G.: Accelerating Activity Inference on Edge Devices Through Spatial Redundancy in Coarse-Grained Dynamic Networks. IEEE Internet of Things Journal 11(24), 41273–41285 (2024)
29. Fu, J., Liu, J., Wang, Y., Zhou, J., Wang, C., Lu, H.: Stacked Deconvolutional Network for Semantic Segmentation. IEEE Transactions on Image Processing (2019)
30. Chen, H., Wu, G., Li, J., Wang, J., Tao, H.: Research Advances on Deep Learning Recommendation Based on Attention Mechanism. Computer Engineering and Science 43(2), 370–380 (2021)
31. Hu, J., Shen, L., Albanie, S., Sun, G., Wu, E.H.: Squeeze-and-Excitation Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence 42(8), 2011–2023 (2020)
32. Lau, K.W., Po, L.M., Rehman, Y.A.U.: Large Separable Kernel Attention: Rethinking the Large Kernel Attention Design in CNN. Expert Systems with Applications 236, 121352 (2024)
33. Wang, H., Zhao, B., Zhang, W.J., Liu, G.H.: LGANet: Local and Global Attention Are Both You Need for Action Recognition. IET Image Processing 17(12), 3453–3463 (2023)