GS-YOLO: Lightweight and Highly Efficient Real-time Aerial Image Detector

Rundong Gao[1], Yu Zhang[1,*], Kailai Zhuang[2], Jiajie Fan[3], Zeming Tian[5], Runxiao Gao[4]
[1]Shenyang University of Chemical Technology, China
[2]Tiangong University, China
[3]South China Normal University, China
[4]Shandong University of Technology, China
[5]Wuhan University, China
*Corresponding Author: zhangy@syuct.edu.cn

***Abstract*—Aerial image detection has extensive applications in real-world scenarios. Despite significant progress in object detection, extremely small targets remain difficult to detect due to sparse pixels and complex backgrounds. Existing high-performance models usually contain a large number of parameters, which makes lightweight deployment on edge devices difficult. To address these issues, we propose an innovative detection framework named GS-YOLO, which enhances the representation of small object features through two lightweight modules, the Edge-Aware Gaussian Downsampling Module (EAG-Stem) and the Gaussian Difference Calibration Module (GDCM). The EAG-Stem focuses on the edge information of small objects, while the GDCM models the continuity of local structures to improve the distinction between objects and background. Furthermore, a Scale-Adaptive Weighted Intersection Over Union(SA-WIoU) loss function is designed to handle multi-scale detection disparities by dynamically adjusting the loss weights of targets at different scales, thereby improving localization accuracy for small objects. Extensive experiments conducted on the Visdrone dataset demonstrate that GS-YOLO achieves an effective trade-off between accuracy and efficiency. Compared to the baseline model, GS-YOLO achieves an average improvement of 7.4% in Average Precision (AP) and an average reduction of 14.5% in FLOPs, with a particularly notable average reduction of 71.76% in Params. The proposed model also outperforms many other detectors on the SIMD dataset.**

***Index Terms*—Object Detection, Scale-aware Fusion, Generalization Capability**

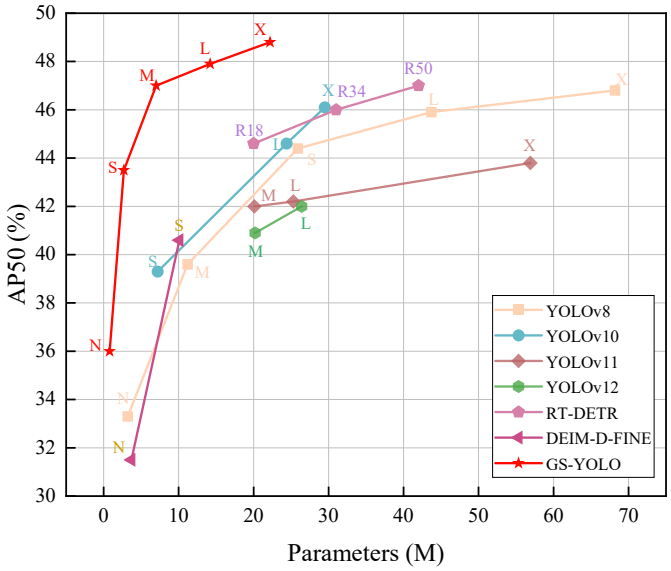

Fig. 1. Comparisons of the real-time object detectors on the Visdrone dataset. The object detection method GS-YOLO achieves the best trade-off between performance and parameters.

I. Introduction

Object detection in aerial imagery is a critical aspect for diverse applications, including urban traffic management, maritime search and rescue, and disaster emergency assessment [1]. Objects in aerial images have substantial scale variation, and small objects are plagued by several challenges. Due to small objects' limited pixel resolution [2], blurred edge textures, susceptibility to background noise, and dense distribution, aerial image object detection is among the most challenging tasks in object detection.

During feature extraction for small objects, an inherent imbalance exists between detailed spatial information and semantic information. An overemphasis on detailed information may introduce extraneous background noise, while an excessive focus on semantic information can compromise localization accuracy. Current research primarily addresses this problem through three main approaches: enhancing detail information, supplementing semantic information, and multi-scale feature fusion. For instance, NAS-FPN [3] utilizes neural architecture search to automatically identify optimal low level feature transmission paths, thereby mitigating excessive compression of critical detail information. FE-YOLOv5 [4] incorporates a feature enhancement module (FEM) with a non-local network to model global relationships, enriching the categorical information of small objects. HRDNet [5] employs a multi-depth image pyramid network (MD-IPN) to align and fuse multi-resolution features, aiming to balance detail and semantic information.

Despite these advances, existing methods still face limitations in detecting very small objects while remaining lightweight. Small objects inherently exhibit a very low signal to noise ratio (SNR). Conventional approaches often perform poorly in noise suppression, and the subsequent feature fusion can further degrade the SNR, ultimately leading to suboptimal

perception of tiny targets.

To address the challenges of small object detection, this paper proposes GS-YOLO, a novel network equipped with two lightweight modules and a customized loss function. Specifically, to mitigate the blurred object contours and excessive background noise generated when small targets undergo downsampling, we introduce the Edge-Aware Gaussian Downsampling Module (EAG-Stem). Replacing the conventional stem structure in the early feature extraction and scale reduction stage, EAG-Stem explicitly enhances edge details and balances spatial compression, thereby providing a more robust early feature representation for small objects. In addition, small scale targets in aerial imagery often suffer from background confusion due to low contrast. To alleviate this issue, we integrate the Gaussian Difference Calibration Module (GDCM) into the backbone. By adaptively modulating detail enhancement and structural smoothing, GDCM produces a target representation that is more distinguishable from the background, thus improving localization accuracy. Finally, to enhance the model's focus on tiny targets, we design the Scale-Adaptive Weighted Intersection Over Union (SA-WIoU) loss function. This loss function increases the relative weight of tiny objects predictions, preventing their gradients from being overwhelmed during optimization.

Experimental validation on the VisDrone 2019 [6] and the Satellite Imagery Multi-Vehicles Dataset (SIMD) [7] demonstrates both the efficiency and effectiveness of GS-YOLO. Compared with the baseline YOLOv8 [8] models, GS-YOLO achieves higher accuracy with substantially fewer parameters. Our approach achieves a better balance between detection performance and model parameters, as illustrated in Fig. 1. The main contributions of this paper are summarized as follows:

(1) We propose the EAG-Stem as the initial layer of the backbone, leveraging a Gaussian prior mechanism to enhance the network's capability for edge-aware feature extraction.

(2) We propose the GDCM, which improves target background separability by jointly coordinating detail enhancement and structural smoothing information.

(3) We design the SA-WIoU loss function, which increases the loss weight of small targets without adding any additional model parameters.

II. Related Work

A. Small Object Detection

Small object detection remains one of the major challenges in object detection due to the limited effective pixels and severe background interference. Recent research has primarily focused on improving detection accuracy through multi-scale feature enhancement and high resolution feature fusion. For instance, HIC-YOLOv5 [9] introduces a high resolution feature fusion module and attention mechanism into the YOLOv5 framework, significantly improving the average precision (AP) for small objects. ScaleKD [10] employs multi-scale feature distillation and cross scale attention to effectively transfer fine granularity small object features from the teacher network to the student network, thereby enhancing the model's multi-scale discriminative capability. Although these methods have achieved remarkable improvements in detection accuracy, they generally suffer from large model size and high computational complexity, which limit their applicability in real-time and edge scenarios. To address this, TinyDet [11] incorporates sparse convolution and an early feature enhancement module in the backbone, increasing the information density of shallow features while reducing redundant computation in deeper layers. UAV-DETR [12], based on the DETR framework, integrates a lightweight feature pyramid and sparse attention mechanism to balance detection performance and inference speed. Nevertheless, existing models still struggle to achieve satisfactory performance for small object detection on low-power devices, underscoring the need to balance accuracy and efficiency.

B. IoU Loss Function

The IoU loss function serves as the core optimization objective for bounding box regression in object detection, and its formulation directly affects both the model's localization accuracy and convergence stability. To address the gradient vanishing and scale sensitivity issues of the original IoU in non-overlapping scenarios, a series of improved variants such as DIoU [13], WIoU [14], and SIoU [15] have been proposed. Although these approaches improve localization precision in general detection tasks, they are insufficiently adapted to the unique characteristics of small objects.

In efforts specifically targeting small objects scenarios, InterpIoU [16] replaces handcrafted geometric penalty terms with interpolated bounding boxes to enhance gradient propagation in non-overlapping small target cases. However, its interpolation coefficients lack dynamic adaptation to local density variations, which can lead to gradient oscillation in dense scenes due to excessive box overlap. Inner-WIoU [17] enhances robustness under occlusion by integrating WIoU's attention mechanism with a strategy that utilizes auxiliary boxes. But it does not specifically optimize sensitivity in low IoU regions, limiting its effectiveness in constraining the localization of extremely small and blurred targets. EWDIoU [18] introduces Gaussian distributions and Wasserstein distance to alleviate over sensitivity to positional deviations in small objects detection. Nevertheless, its handling of scale heterogeneity remains simplistic, and training stability is still influenced by scale fluctuations.

Although existing IoU-based methods have improved small objects localization through enhancements in gradient transmission, sample weighting, and geometric constraints, they still exhibit localization limitations in complex small targets scenes. We balance optimization across IoU ranges to reduce instability from scale heterogeneity and thereby improve small-object localization accuracy.

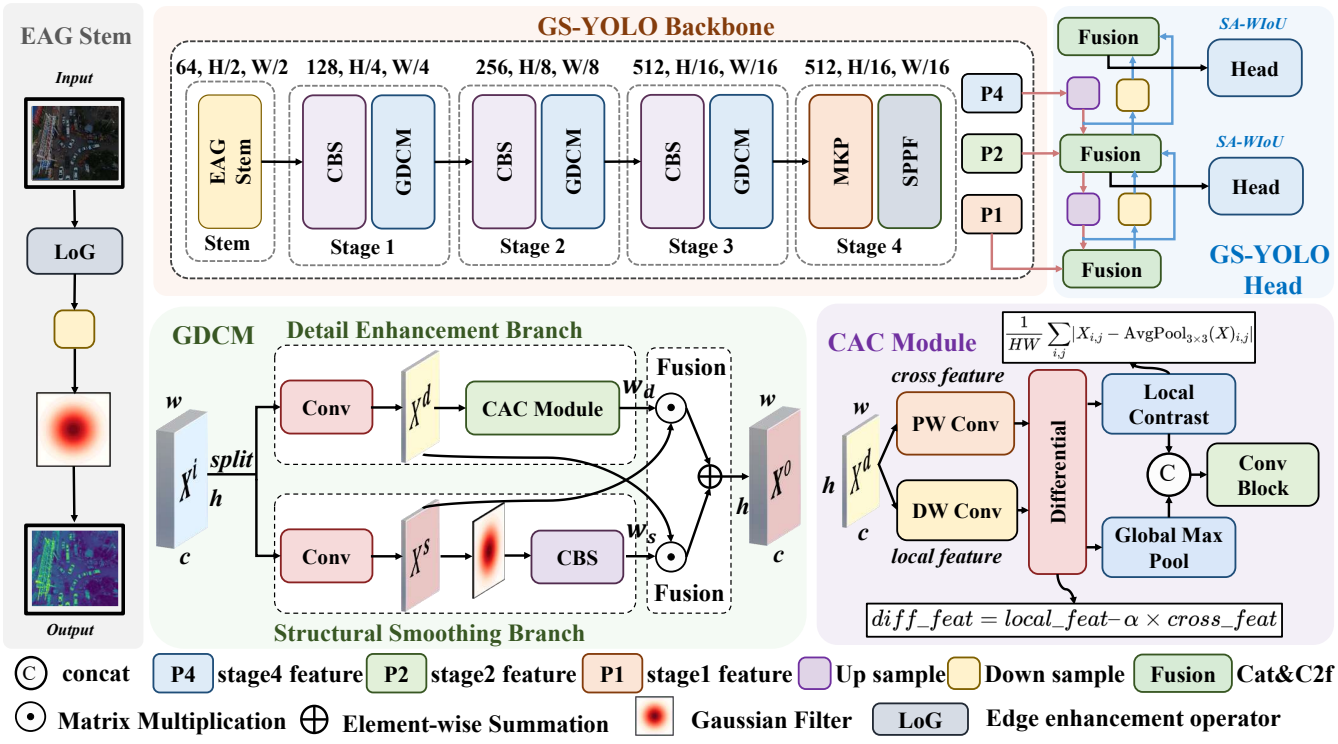

Fig. 2. Framework of GS-YOLO. The EAG-Stem is located in the initial part of the backbone network to enhance the edge details of the target. The GDCM is embedded in each stage of the backbone network, performing target differential calibration through Gaussian-smoothed spatial information and semantic information of salient details.

III. METHODOLOGY

A. GS-YOLO Architecture Overview

This study proposes a GS-YOLO, which is built upon the architecture of YOLOv8, as illustrated in Fig. 2. This incorporates the EAG-Stem for enhancing edge details and the GDCM for refining and highlighting target features. To improve the detection performance of small objects, we also introduce the SA-WIoU. The EAG-Stem is designed to provide the model with edge structure prior features through LoG and Gaussian filters. The GDCM leverages Gaussian-guided spatial information and locally salient semantic information to achieve refinement for small targets. SA-WIoU dynamically distinguishes between large and small targets using a fixed threshold and models their IoU characteristics accordingly. We introduce the MKP module [19] and embed it into the backbone network.

B. Edge-Aware Gaussian Downsampling Stem

Aiming at the problem of edge detail loss after downsampling of small targets, EAG-Stem is proposed as the initial feature extraction and scale compression unit of the network to display the enhanced target edge details. First, the input image is processed with a LoG filter with a kernel size of k and a standard deviation of σ_1 to highlight the small target contour display and strengthen its positioning. The LoG filter consists of a Gaussian filter and a Laplace operator, defined as:

$$LoG(x, y; \sigma_1) = \nabla^2[G(x, y; \sigma_1) * I_0(x, y)] \tag{1}$$

where $*$ denotes the convolution operation. Among them, Gaussian filtering is as follows:

$$G(x, y; \sigma_1) = \frac{1}{2\pi\sigma_1^2} \exp(-\frac{x^2 + y^2}{2\sigma_1^2}) \tag{2}$$

Then it goes through the Gaussian filter of k, σ_2 to smooth the features of the target and suppress the background noise. Between the two filters, the traditional Stem structure is replaced by a downsampling unit to perform scale compression. The whole process can be expressed mathematically as follows:

$$F_{EAG} = G_{\sigma_2} * D_s(LoG(X^i; \sigma_1)) \tag{3}$$

where, $D_s(\cdot)$ is a downsampling operator. Through this process, prior features with clear edge structures are obtained, as shown in the left part of Fig. 2.

C. Gaussian Differential Calibration Module

Due to the inherently low distinguishability of small targets, feature ambiguity often complicates their separation from the background during detection, resulting in decreased localization accuracy and reduced detection robustness. To address this issue, the GDCM is designed to achieve refined recalibration

of small target features by fusing structural smoothing information with detail enhancement information. As illustrated in Fig. 2, the GDCM is constructed based on two core principles: (1) initial feature extraction and (2) cross-branch collaborative calibration.

1) Initial Feature Extraction: Given an input feature map $X^i \in \mathbb{R}^{C\times H\times W}$, it is split the channel dimension into two parallel branches: structural smoothing and detail enhancement. The first branch employs a single-layer convolution for lightweight feature mapping, whereas the second branch applies a three-layer cascaded convolutional transformation to enhance representation capability. The resulting initial features of the two branches are denoted as X^d and X^s, respectively:

$$X^d, X^s = \phi_1(\text{split}(X^i)) \tag{4}$$

where $\phi_1(\cdot)$ represents a convolutional transformation for feature learning on the split branches.

2) Cross-branch Collaborative Calibration: The structural smoothing branch enhances the robustness of small target localization in low SNR environments by suppressing random noise in the feature maps and reinforcing the inherent spatial continuity of the target. For the feature X^s, a learnable two-dimensional Gaussian kernel with a size of $k \times k$ and σ_3 is applied to smooth the feature information, thereby blurring high-frequency noise while preserving the spatial structure of the target. The element-wise modulation weights for structural smoothing are computed as follows:

$$w_s = \text{CBS}(G_{\sigma_3} * X^s) \tag{5}$$

The detail enhancement branch focuses on capturing and emphasizing locally salient regions within the feature map. It performs adaptive denoising and detail enhancement through a local-global difference mechanism. The feature X^d is processed in parallel, where a depthwise separable convolution captures local features (local_{feat}) for each channel, and a grouped pointwise convolution extracts cross-channel global features (cross_{feat}). A learnable difference weight parameter α is then introduced to compute the differential features:

$$\text{diff}_{feat} = \text{local}_{feat} - \alpha \times \text{cross}_{feat} \tag{6}$$

This operation measures the contrast between local and global representations. Subsequently, a position-sensitive local contrast (z_1) feature is constructed in one direction as follows:

$$z_1 = \frac{1}{HW}\sum_{i,j}\left|\text{diff}_{feat,i,j} - \text{AvgPool}_{3\times3}(\text{diff}_{feat})_{i,j}\right| \tag{7}$$

In the other direction, a global context information is extracted and broadcast to obtain global context (z_2) feature. After fusing the z_1 and z_2 and applying a mapping transformation, the detail enhancement modulation weight w_d is obtained:

$$w_d = \phi_2\big(\text{concat}(z_1, z_2)\big) \tag{8}$$

where $\phi_2(\cdot)$ represents a learnable 1×1 convolutional mapping with non-linearity. Following this, cross-branch collaborative calibration is performed: the features of one branch are guided by the modulation weights derived from the other, allowing the structural and detail information to mutually reinforce each other. Specifically, the detail-enhanced feature X^d is modulated by the structural smoothing weight w_s to produce F_D:

$$F_D = w_s \odot X^d \tag{9}$$

where $\odot$ denotes element-wise multiplication. Similarly, the structurally smoothed feature X^s is modulated by the detail enhancement weight w_d to yield F_S:

$$F_S = w_d \odot X^s \tag{10}$$

Finally, the two calibrated features F_D and F_S are fused through element-wise addition to generate the output feature X^o:

$$X^o = F_D \oplus F_S \tag{11}$$

where, the symbol $\oplus$ denotes element-wise summation.

D. Scale-Adaptive Weighted IoU Loss

We propose SA-WIoU, an IoU loss that improves small-object localization and stabilizes training by reducing gradient over-penalization and convergence delays from scale and density variation.

To align the loss statistics among objects of different scales, objects are categorized into small targets and non-small targets according to an area threshold of A_t. The running mean momentum of $L_{\text{IoU}} = 1 - \text{IoU}$ is updated separately for the two groups. A mild exponential scaling is applied to small targets, whereas other targets retain the original scaling:

$$S = \begin{cases} \dfrac{\beta_{\text{small}}}{\delta \cdot \eta \cdot \gamma^{(\beta_{\text{small}}\cdot\zeta-\delta)}}, & A \le A_t, \\ \dfrac{\beta}{\delta \cdot \gamma^{(\beta-\delta)}}, & \text{otherwise.} \end{cases} \tag{12}$$

where η denotes the scale adjustment coefficient, and ζ denotes the offset correction coefficient. The scale coefficient for small targets is defined as

$$\beta_{\text{small}} = \frac{L_{\text{IoU}}}{\mu_{\text{small}}} \tag{13}$$

To improve sensitivity for low IoU small targets while avoiding gradient explosion at high IoU, we introduce a perceptual weighting factor for small targets:

$$W_{\text{small}} = (1 + \mathbf{1}_{A\le\lambda A_t}) \times (1 - \text{IoU}_{\text{detach}} \times \mu) \tag{14}$$

where $\lambda \in (0, 1)$ denotes the proportion threshold for identifying small scale cases, and $\mu \in (0, 1)$ denotes the scaling coefficient for IoU detachment.

The perceptual weighting described above increases the contribution of small objects, encouraging stricter localization. Building upon the distance penalty in WIoU, we further introduce a composite modulation factor that jointly accounts

TABLE I

COMPARISON OF AP(%), AP_{50}(%), PARAMS AND FLOPS ON VISDRONE BY USING DIFFERENT REAL-TIME OBJECT DETECTORS. BEST AND SECOND-BEST RESULTS ARE HIGHLIGHTED FOR EACH METRIC.

Backbone	Model	Publication	Input Size	AP	AP_{50}	FLOPs (G)	Params (M)
CNN-Based	YOLOv8-N [8]	-	640×640	19.6	33.3	8.9	3.2
	YOLOv8-S [8]	-	640×640	23.6	39.6	28.8	11.2
	YOLOv8-M [8]	-	640×640	27.1	44.4	79.3	25.9
	YOLOv8-L [8]	-	640×640	28.4	45.9	165.7	43.7
	YOLOv8-X [8]	-	640×640	28.9	46.8	258.5	68.2
	YOLOv10-S [20]	arXiv2024	640×640	23.8	39.3	21.6	7.2
	YOLOv10-L [20]	arXiv2024	640×640	27.6	44.6	120.3	24.4
	YOLOv10-X [20]	arXiv2024	640×640	28.7	46.1	160.4	29.5
	YOLOv11-M [21]	arXiv2024	640×640	25.0	42.0	68.0	20.1
	YOLOv11-L [21]	arXiv2024	640×640	25.5	42.2	86.9	25.3
	YOLOv11-X [21]	arXiv2024	640×640	26.6	43.8	194.9	56.9
	YOLOv12-M [22]	arXiv2025	640×640	24.4	40.9	67.5	20.2
	YOLOv12-L [22]	arXiv2025	640×640	25.1	42.0	88.9	26.4
	YOLO26-N [23]	-	640×640	13.5	24.9	**5.2**	**2.4**
	YOLO26-S [23]	-	640×640	16.1	29.4	20.5	9.5
Transformer-Based	Deform-DETR [24]	ICLR2020	1300×800	27.1	42.2	173.0	40.0
	Sparse DETR [25]	ICLR2022	1300×800	27.3	42.5	121.0	40.9
	RT-DETR-R18 [26]	CVPR2024	640×640	26.7	44.6	60.0	20.0
	RT-DETR-R34 [26]	CVPR2024	640×640	27.2	46.0	92.0	31.0
	RT-DETR-R50 [26]	CVPR2024	640×640	28.4	47.0	136.0	42.0
	DEIM-D-FINE-N [27]	CVPR2025	640×640	17.8	31.5	**7.1**	3.7
	DEIM-D-FINE-S [27]	CVPR2025	640×640	24.3	40.6	24.9	10.1
Mamba-Based	Mamba-YOLO-T [28]	AAAI2025	640×640	21.0	36.8	13.6	6.0
	Mamba-YOLO-B [28]	AAAI2025	640×640	23.9	40.8	49.6	21.8
GS-YOLO (Ours)	GS-YOLO-N	-	640×640	21.2	36.0	8.4	**0.8**
	GS-YOLO-S	-	640×640	26.4	43.5	26.9	2.7
	GS-YOLO-M	-	640×640	29.1	47.0	65.9	7.0
	GS-YOLO-L	-	640×640	**29.8**	**47.9**	130.4	14.2
	GS-YOLO-X	-	640×640	**30.3**	**48.8**	201.7	22.2

for object scale, overlap density, and IoU perception. The overall modulation factor is expressed as:

$$R_{\text{WIoU}} = \exp\Big\{\frac{\rho^2}{c^2} \times \Big[\kappa_1 + \Big(\kappa_2 - \frac{A}{w_h c_c}\Big) \times \Delta_1\Big] \times \Big(\kappa_3 + \frac{|A \cap B|}{A} \times \Delta_2\Big)\Big\} \quad (15)$$

where κ_1, κ_2, and κ_3 are baseline reference coefficients. Δ_1 controls the influence of the area-sensitive term, while Δ_2 regulates the overlap-density term through the intersection ratio. The final SA-WIoU loss for small targets is formulated as:

$$\mathcal{L}_{\text{SA-WIoU}} = L_{\text{IoU}} \times S \times R_{\text{WIoU}} \times W_{\text{small}} \quad (16)$$

By decoupling scale statistics and applying adaptive modulation to the WIoU loss, SA-WIoU provides more stable gradients and stronger localization supervision for small objects.

IV. EXPERIMENTS

A. Experimental Setups

Comprehensive object detection experiments were conducted on GS-YOLO using the VisDrone2019 dataset and the SIMD dataset. All experiments were performed on an NVIDIA GeForce RTX 4090 GPU. The model was trained using the Stochastic Gradient Descent (SGD) optimizer, with an initial learning rate of 0.01, a momentum coefficient of 0.937, and a weight decay of 0.0005. During training, the batch size was set to 4, and the model was trained for 300 epochs. The hyperparameters are set as shown in Table II. All hyper-parameters are empirically fixed and kept consistent across all experiments.

B. Performance on VisDrone2019 Dataset

1) Quantitative Analysis Based on Tabular Data: We conducted a extensive evaluation of the GS-YOLO series models and various mainstream real-time object detectors on the VisDrone 2019 dataset, with detailed results presented in Table I. GS-YOLO achieves better performance with fewer parameters. Among lightweight models, GS-YOLO-N achieves 21.2% AP using only 0.8M parameters, representing a 75% reduction in parameters compared to the baseline model. Meanwhile, its parameter count is 78.4% fewer than the ultra-lightweight DEIM-D-FINE-N, while achieving 3.4% higher AP. For small models, GS-YOLO-S delivers a 2.8% AP improvement with

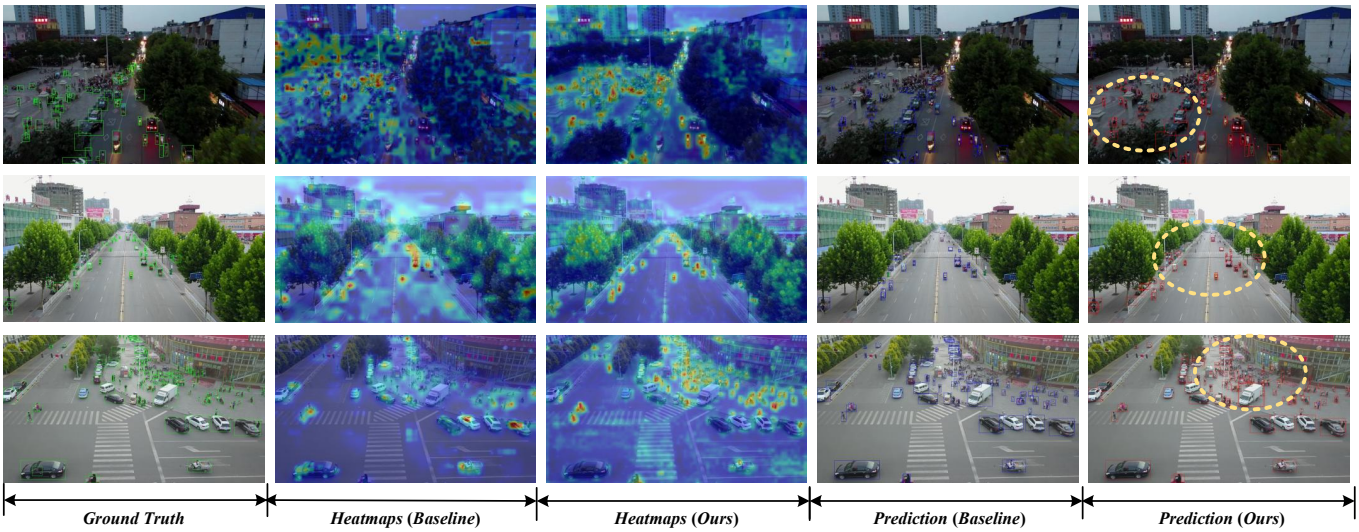

Fig. 3. Visualization results of different Models on the VisDrone2019 Dataset. The yellow boxes highlight areas where our model performs better in detecting objects. As evidenced by the heatmaps, the EAG-Stem suppresses background noise, while the GDCM enhances the contrast of small objects in dense target regions, leading to more discriminative activations for small, closely spaced objects compared to the baseline C2f under identical settings.

TABLE II
HYPER-PARAMETER SETTINGS

Hyper-parameter	Value
σ_1	1.0
σ_2	0.5
σ_3	0.5
k	7
η	0.7
ζ	0.8
λ	0.3
μ	0.5
$\kappa_{1,2,3}$	1.0
Δ_1	0.5
Δ_2	0.3

TABLE III
COMPARISON WITH STATE-OF-THE-ART DETECTORS ON VISDRONE.

Method	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
GFL V1 [29]	23.4	38.2	24.2	14.3	34.7	40.1
CDMNet [30]	29.2	**49.5**	29.8	16.1	35.9	36.5
QueryDet [31]	28.3	48.1	28.7	-	-	-
DINO [32]	26.8	44.2	28.9	17.5	37.3	41.3
D-FINE [33]	23.9	41.6	29.0	13.6	34.6	46.4
FBRT-YOLO [19]	**30.1**	48.4	**31.7**	**18.1**	**40.8**	**47.9**
GS-YOLO-X (ours)	**30.3**	**48.8**	**31.5**	**19.0**	**41.3**	**50.2**

a 75.9% parameter reduction. For larger models, GS-YOLO-X demonstrates the most outstanding performance among all compared models, achieving 30.3% AP and 48.8% AP_{50} with only 22.2M parameters and 201.7G FLOPs, while still improving AP by 1.4% over YOLOv8-X. Even when compared with high-precision Transformer models such as RT-DETR and DEIM, GS-YOLO-X exhibits more favorable AP. These results collectively validate that the GS-YOLO series models, particularly through their Gaussian enhancement and scale-aware mechanisms, not only substantially improve detection accuracy but also significantly reduce model parameter count in the multi-scale and small object challenge scenarios prevalent in VisDrone 2019.

Table III shows that the proposed GS-YOLO-X attains marked improvements in AP and AP_{50} on VisDrone. These results indicate that our method is competitive with other advanced approaches for aerial image object detection.

2) Visualization Result Analysis on Focused Regions of the Network: For a visual comparison of focused areas and detection performance on the VisDrone2019 dataset, we have visualized the heatmaps and prediction results in Fig. 3. As shown in the heatmaps, the EAG-Stem and GDCM jointly enhance the representation of small objects. The yellow box areas clearly illustrate that GS-YOLO achieves fewer missed detections in scenarios with dense vehicles and pedestrians, thereby demonstrating its effectiveness in detecting small and densely packed targets.

C. Performance on SIMD Dataset

As presented in Table IV, GS-YOLO-M exhibits a commendable performance on the SIMD dataset, achieving an AP of 67.0% and an AP_{50} of 82.5%, which are competitive among SOTA methods. Notably, this level of accuracy is attained with a remarkably low parameter count of 7.0 M, indicating superior parameter efficiency. Although GS-YOLO-M requires 65.9 GFLOPs, its detection results indicate a good trade-off between compute and accuracy.

TABLE IV
SOTA COMPARISON ON SIMD DATASET

Method	AP	AP_{50}	Params	FLOPS
RT-DETR-R18 [26]	63.7	78.6	19.8M	**57.0G**
YOLOv8-L [8]	63.1	78.1	43.7M	165.2G
YOLOv9-M [34]	62.2	76.6	20.0M	76.3G
Deform-DETR [24]	59.7	75.6	40.0M	196.0G
EMSD-DETR [35]	64.3	79.4	18.4M	68.3G
HPS-DETR [36]	63.5	79.8	**15.5M**	68.3G
FMC-DETR-B [37]	**65.8**	**80.9**	16.1M	**56.2G**
GS-YOLO-M (ours)	**67.0**	**82.5**	**7.0M**	65.9G

TABLE V
ABLATION STUDY ON MODULE ABLATION FOR THE VISDRONE DATASET

EAG-Stem	GDCM	SA-WIoU	AP	AP_{50}	Params	FLOPS
✗	✗	✗	23.6	39.6	11.2M	28.8 G
✓	✗	✗	24.7	41.5	9.1M	25.8G
✓	✓	✗	25.9	42.8	2.7M	26.9G
✓	✓	✓	**26.4**	**43.5**	**2.7M**	**26.9G**

TABLE VI
EXPERIMENTS ON SA-WIOU LOSS FUNCTION FOR OBJECTS OF DIFFERENT SIZES

A_t	AP	AP_{50}	AP_{75}
32x32	26.1	42.9	27.0
24x24	**26.3**	**43.2**	**27.1**
16x16	**26.4**	**43.5**	**27.2**
12x12	**26.3**	**43.5**	**27.1**

D. Ablation Studies

To verify the individual and synergistic effectiveness of our core modules on object detection models' performance and efficiency, we conducted ablation experiments using YOLOv8-S on the VisDrone dataset. As shown in Table V, the baseline model attains an AP of 23.6, AP_{50} of 39.6, with 11.2M parameters and 28.8G FLOPS. The initial integration of the EAG-Stem improves detection performance, increasing AP to 24.7 and AP_{50} to 41.5, while reducing model complexity to 9.1M parameters and 25.8G FLOPS. Building on this, the subsequent integration of the GDCM further steadily improves model performance, achieving an AP of 25.9 and an AP_{50} of 42.8, with the model complexity reduced to 2.7M parameters. Finally, replacing the standard IoU loss function with SA-WIoU increases the AP to 26.4 and AP_{50} to 43.5, with the model complexity maintained at 2.7 M parameters and 26.9 G FLOPS. These results validate the effectiveness of our design and lay a solid foundation for achieving efficient and high-precision object detection on complex datasets.

Additionally, we evaluate the impact of the area threshold A_t used to distinguish small and non-small targets in the IoU loss. Table VI reports the performance of GS-YOLO-S under different settings of A_t. It can be observed from the experimental results that SA-WIoU enables the model to pay more attention to extremely small targets of 16×16 pixels, and the model performs best at this size.

CONCLUSION

This paper proposes a lightweight detection framework, termed GS-YOLO, specifically tailored for small object detection. The EAG-Stem incorporates edge structure prior information from the input image, while the GDCM enhances the representation of salient target regions. Additionally, the SA-WIoU loss function is introduced to improve localization accuracy for small objects. Experimental evaluations on the VisDrone and SIMD datasets demonstrate that the proposed method achieves a favorable trade-off among detection precision, model efficiency, and real-time inference capability. Future research will concentrate on extending the framework to more complex and practical real-world scenarios.

REFERENCES

[1] G. Cheng, X. Yuan, X. Yao, K. Yan, Q. Zeng, X. Xie, and J. Han, "Towards large-scale small object detection: Survey and benchmarks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 11, pp. 13 467–13 488, 2023.

[2] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proc. Computer Vision – ECCV*. Cham: Springer International Publishing, 2014, pp. 740–755, lecture Notes in Computer Science.

[3] G. Ghiasi, T.-Y. Lin, and Q. V. Le, "Nas-fpn: Learning scalable feature pyramid architecture for object detection," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 7029–7038.

[4] M. Wang, W. Yang, L. Wang, D. Chen, F. Wei, H. KeZiErBieKe, and Y. Liao, "Fe-yolov5: Feature enhancement network based on yolov5 for small object detection," *J. Visual Commun. Image Represent.*, vol. 90, p. 103752, 2023.

[5] Z. Liu, G. Gao, L. Sun, and Z. Fang, "Hrdnet: High-resolution detection network for small objects," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, 2021, pp. 1–6.

[6] P. Zhu, L. Wen, X. Bian, H. Ling, and Q. Hu, "Vision Meets Drones: A Challenge," *arXiv preprint arXiv:1804.07437*, 2018, arXiv preprint.

[7] M. Haroon, M. Shahzad, and M. M. Fraz, "Multisized object detection using spaceborne optical imagery," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 3032–3046, 2020.

[8] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics yolov8," 2023, [Online]. Available: https://github.com/ultralytics/ultralytics. Accessed: Jan. 2, 2026.

[9] S. Tang, S. Zhang, and Y. Fang, "Hic-yolov5: Improved yolov5 for small object detection," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, 2024, pp. 6614–6619.

[10] Y. Zhu, Q. Zhou, N. Liu, Z. Xu, Z. Ou, X. Mou, and J. Tang, "Scalekd: Distilling scale-aware knowledge in small object detector," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 19 723–19 733.

[11] S. Chen, T. Cheng, J. Fang, Q. Zhang, Y. Li, W. Liu, and X. Wang, "Tinydet: Accurate small object detection in lightweight generic detectors," *arXiv preprint arXiv:2304.03428*, 2023, arXiv preprint.

[12] H. Zhang, K. Liu, Z. Gan, and G.-N. Zhu, "Uav-detr: Efficient end-to-end object detection for unmanned aerial vehicle imagery," *arXiv preprint arXiv:2501.01855*, 2025, arXiv preprint.

[13] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-iou loss: Faster and better learning for bounding box regression," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 07, 2020, pp. 12 993–13 000.

[14] Z. Tong, Y. Chen, Z. Xu, and R. Yu, "Wise-iou: Bounding box regression loss with dynamic focusing mechanism," *arXiv preprint arXiv:2301.10051*, 2023, arXiv preprint.

[15] Z. Gevorgyan, "Siou loss: More powerful learning for bounding box regression," *arXiv preprint arXiv:2205.12740*, 2022, arXiv preprint.

[16] H. Liu and H. Watanabe, "Interpiou: Rethinking bounding box regression with interpolation-based iou optimization," *arXiv preprint arXiv:2507.12420*, 2025, arXiv preprint.

[17] K. Hu, J. Z. Lu, C. Q. Zheng, Q. Xiang, and L. Miao, "Mff-yolov8: Small object detection based on multi-scale feature fusion for uav remote sensing images," *IET Image Processing*, vol. 19, no. 1, p. e70066, 2025.

[18] Y. Zhao, Y. Chen, X. Xu, Y. He, H. Gan, N. Wu, Z. Wang, X. Sun, Y. Wang, P. Skobelev, and Y. Mi, "Ta-yolo: Overcoming target blocked challenges in greenhouse tomato detection and counting," *Front. Plant Sci.*, vol. 16, 2025.

[19] Y. Xiao, T. Xu, Y. Xin, and J. Li, "Fbrt-yolo: Faster and better for real-time aerial image detection," *in Proc. AAAI Conf. Artif. Intell.*, vol. 39, no. 8, pp. 8673–8681, 2025.

[20] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, and G. Ding, "Yolov10: Real-time end-to-end object detection," 2024, arXiv preprint arXiv:2405.14458.

[21] R. Khanam and M. Hussain, "Yolov11: An overview of the key architectural enhancements," 2024, arXiv preprint arXiv:2410.17725.

[22] Y. Tian, Q. Ye, and D. Doermann, "Yolov12: Attention-centric real-time object detectors," *arXiv preprint arXiv:2502.12524*, 2025, arXiv preprint.

[23] G. Jocher and J. Qiu, "Ultralytics yolo26," 2026. [Online]. Available: https://github.com/ultralytics/ultralytics

[24] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020, conference paper.

[25] B. Roh, J. Shin, W. Shin, and S. Kim, "Sparse detr: Efficient end-to-end object detection with learnable sparsity," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022, conference paper.

[26] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen, "Detrs beat yolos on real-time object detection," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 16 965–16 974.

[27] S. Huang, Z. Lu, X. Cun, Y. Yu, X. Zhou, and X. Shen, "Deim: Detr with improved matching for fast convergence," 2025, arXiv preprint arXiv:2412.04234.

[28] Z. Wang, C. Li, H. Xu, X. Zhu, and H. Li, "Mamba yolo: A simple baseline for object detection with state space model," *in Proc. AAAI Conf. Artif. Intell.*, vol. 39, no. 8, pp. 8205–8213, 2025.

[29] X. Li, W. Wang, L. Wu, S. Chen, X. Hu, J. Li, J. Tang, and J. Yang, "Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection," 2020, arXiv preprint arXiv:2006.04388.

[30] C. Duan, Z. Wei, C. Zhang, S. Qu, and H. Wang, "Coarse-grained density map guided object detection in aerial images," in *Proc. IEEE/CVF Int. Conf. Computer Vision Workshops (ICCVW)*, 2021, pp. 2789–2798.

[31] C. Yang, Z. Huang, and N. Wang, "Querydet: Cascaded sparse query for accelerating high-resolution small object detection," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 13 668–13 677.

[32] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, and H.-Y. Shum, "Dino: Detr with improved denoising anchor boxes for end-to-end object detection," 2022, arXiv preprint arXiv:2203.03605.

[33] Y. Peng, H. Li, P. Wu, Y. Zhang, X. Sun, and F. Wu, "D-fine: Redefine regression task in detrs as fine-grained distribution refinement," 2024, arXiv preprint arXiv:2410.13842.

[34] C.-Y. Wang and H.-Y. M. Liao, "Yolov9: Learning what you want to learn using programmable gradient information," 2024, arXiv preprint arXiv:2402.13616.

[35] C. Zhang and J. Yang, "Emsd-detr: Efficient small object detection for uav aerial images based on enhanced rt-detr model," *J. Supercomput.*, vol. 81, p. 1052, 2025.

[36] X. Wang and H. Chen, "Hps-detr: Enhancing small object detection with lightweight feature extraction and transformer integration," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, pp. 1–20, 2025.

[37] B. Liang, Y. Liu, B. Qiu, Y. Wang, X. Sui, and Q. Chen, "Fmc-detr: Frequency-decoupled multi-domain coordination for aerial-view object detection," 2025, arXiv preprint arXiv:2509.23056.