Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

SGL-YOLO improves nighttime vehicle detection in unmanned aerial imagery

Zeen Song
Beijing University of Technology
Tieneng Guo
guotn@163.com

Beijing University of Technology

Article

Keywords: UAV, nighttime vehicle detection, SGL-YOLO, multiscale feature fusion, global context, lightweight object detection

Posted Date: November 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11335317/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

SGL-YOLO improves nighttime vehicle detection in unmanned aerial imagery

Zeen Song[1] and Tieneng Guo[2,3,*]

[1]School of Mathematics, Statistics and Mechanics,
Beijing University of Technology, Beijing 100124, People's Republic of China

[2]Present address: College of Mechanical and Energy Engineering, Beijing University of Technology, Beijing 100124, People's Republic of China

[3]Key Laboratory of Green and Intelligent Development and Efficient Utilization of Strategic Mineral Resources of Xinjiang Production and Construction Corps, School of Mechatronics and Automation Engineering, Xinjiang Polytechnic Institute, Hotan Xinjiang 848000 P.R. China

Email (Zeen Song): zeensong17@gmail.com

[*]Corresponding author: Tieneng Guo (guotn@163.com)

Abstract

Nighttime vehicle detection using unmanned aerial vehicles (UAVs) is important for traffic monitoring, road safety management, and emergency response. However, substantial variations in object scale, low target-to-background contrast, complex backgrounds, and limited onboard computing resources pose considerable challenges to accurate and efficient detection. To address these challenges, this study proposes SGL-YOLO (Similarity-guided Global-Lightweight YOLO), a lightweight detector based on YOLO26. We propose a similarity-based attention bidirectional feature pyramid network (SBA-FPN) that integrates bidirectional feature propagation, adaptive weighted fusion, and parameter-free attention to enhance multiscale feature representation and preserve informative vehicle features across different scales. We propose a global context enhancement module (GCEM) that jointly models channel and spatial importance to strengthen vehicle-related responses while suppressing interference from complex backgrounds. A Ghost-DFC Fusion Block (GDFB) is proposed to combine lightweight local feature extraction with long-range spatial modeling, enabling effective integration of vehicle structural features and contextual information with limited computational overhead. Experiments show that SGL-YOLO achieves mean average precision (mAP) of 57.3% on DroneVehicle_Night and 43.6% on VisDrone_Night, outperforming all selected comparison models on both datasets. These results demonstrate the effectiveness of the architecture for vehicle detection in complex nighttime UAV imagery and highlight its potential for resource-constrained UAV traffic monitoring.

Keywords: UAV; nighttime vehicle detection; SGL-YOLO; multiscale feature fusion; global context; lightweight object detection

Introduction

Vehicle detection in unmanned aerial vehicle (UAV) imagery offers a flexible approach to traffic monitoring [1, 2]. Studies of fatal pedestrian crashes in darkness further highlight the road-safety concerns associated with nighttime traffic [3]. However, underexposure and low target-to-background contrast [4–6], combined with poorly defined vehicle edges and textures [7–9], complicate detection in nighttime UAV images. Accurate, lightweight detectors are therefore needed for UAV monitoring under challenging nighttime illumination [10, 11].

Vision-based nighttime vehicle detection has evolved from methods based on predefined visual rules to approaches that learn discriminative representations [12]. Conventional methods primarily use vehicle lights, shape, and motion. Zhang et al. [13] combined reflection-aware headlight extraction with bidirectional reasoning for tracking and pairing, enabling nighttime vehicle surveillance and speed estimation. Parvin et al. [14] integrated taillight and headlight features with centroid-based tracking to detect and track vehicles at night. Wang et al. [15] combined taillight pairing with time-series prediction of vehicle regions in subsequent frames, narrowing the search area and reducing incorrect pairings. Nevertheless, changes in illumination and ambiguous bright regions can disrupt these visual cues [15, 16]. Dependence on predefined rules thus motivates learning-based methods that use richer vehicle representations [12].

Learning-based detectors range from classifiers using handcrafted descriptors to deep networks that learn hierarchical image features [17–19]. For UAV imagery, Xu et al. [20] combined Viola–Jones detection with histogram of oriented gradients (HOG) features and support vector machine (SVM) classification. Road-orientation adjustment and adaptive switching between the detectors improved detection performance and efficiency in low-altitude aerial videos. For nighttime road scenes, Zhang et al. [16] combined illumination removal and salient-light extraction with superpixel-enhanced HOG descriptors and SVM classification, improving vehicle recognition when contours were poorly defined. These pipelines, however, still require explicit feature design and scene-dependent preprocessing. Bridging handcrafted and learned representations, Kuang et al. [21] combined bio-inspired image enhancement with SVM classifiers using convolutional neural network (CNN), HOG, and local binary pattern features. Weighted fusion of the classifier scores improved nighttime vehicle detection by integrating complementary representations. Deep detectors extend this approach by learning hierarchical features for localization and recognition [18], addressing the need for expressive representations under the scale, viewpoint, and background variations encountered in UAV imagery [17].

Deep learning detectors commonly use Transformer or CNN architectures [18]. Transformers have attracted considerable interest in computer vision, including object detection [22]. Carion et al. [23] introduced the Detection Transformer (DETR), which uses global attention and set prediction to eliminate anchor boxes and non-maximum

suppression (NMS). To address slow convergence and limited small-object detection performance, Zhu et al. [24] developed Deformable DETR, which restricts attention to a small set of sampling locations around reference points. Li et al. [25] introduced query denoising in DN-DETR to stabilize matching and accelerate convergence. Zhao et al. [26] developed RT-DETR, combining an efficient hybrid encoder for multiscale features with uncertainty-minimal query selection to support real-time, end-to-end detection. Although these designs improve training or inference efficiency, standard global self-attention has quadratic complexity with respect to the number of tokens [22]. Computational and memory demands therefore remain important considerations for resource-constrained UAV applications [17].

CNN-based detectors are generally classified as two-stage or one-stage methods [18]. Two-stage detectors, including R-CNN [27], Fast R-CNN [28], Faster R-CNN [29], and Mask R-CNN, identify candidate regions before performing region-level recognition [18]. For nighttime vehicle detection, Xu et al. [30] incorporated deformable convolution and side-aware boundary localization into Faster R-CNN to improve feature extraction and bounding-box localization under challenging illumination. He et al. [31] developed Mask R-CNN by adding an instance-mask prediction branch to Faster R-CNN alongside bounding-box prediction. Cai and Vasconcelos [32] introduced Cascade R-CNN, using successive detectors trained at increasing intersection-over-union (IoU) thresholds to improve detection quality. These developments strengthen region-level recognition, but proposal generation and subsequent region processing incur additional computation [18]. Comparisons using aerial imagery also show that performance depends on object size and differences between training and testing conditions [33]. The need to balance detection quality with onboard computational constraints motivates the use of efficient one-stage detectors for UAV monitoring.

One-stage detectors, such as SSD [34] and the YOLO family, directly predict object locations and classes, supporting efficient detection [18]. Recent studies have focused on feature fusion and inference efficiency. Zhao et al. [35] proposed FAWDet, a dual-stream detector based on YOLOv8 that combines visible and infrared features through cross-modal cross-attention. For small-object detection in UAV imagery, Zhuo et al. [10] developed TAF-YOLO, which adaptively fuses visible and infrared information to exploit complementary cues. Deployment of these multimodal approaches depends on the availability of suitable paired inputs. To improve computational efficiency, Wang et al. [36] introduced YOLOv10 with consistent dual assignments for NMS-free training and an efficiency-oriented architecture. Weng and Niu [5] developed ELS-YOLO from YOLO11 for low-light UAV detection. Their design combines lightweight multiscale feature selection, a shared-convolution detection head, and model pruning to reduce resource demands. Tian et al. [37] introduced the attention-centric YOLOv12 architecture to balance accuracy and speed in general object detection. Jocher et al. [38] developed Ultralytics YOLO26 with a dual-head design supporting NMS-free, end-to-end inference and a regression head without distribution focal loss (DFL). For nighttime

UAV applications, these efficiency advances must be considered alongside the need to preserve weak features of small vehicles and suppress background interference [4, 5, 17].

Three interrelated challenges remain in nighttime UAV detection [4, 7, 17]. First, changes in flight altitude and vehicle size create substantial scale variation. Feature pyramids may not adequately account for the differing contributions of feature levels, weakening small-object information during fusion. Second, low illumination, road textures, shadows, and uneven lighting reduce target-to-background contrast and limit the discriminative value of local features. Third, increasing network depth or complexity to strengthen feature representations adds parameters and computational cost, hindering deployment on UAV edge devices. To address these challenges, this study proposes SGL-YOLO (Similarity-guided Global-Lightweight YOLO), a lightweight detector based on YOLO26 that integrates multiscale feature fusion, global context enhancement, and local–global feature modeling. Its main contributions are as follows:

1. SBA-FPN improves multiscale feature fusion by combining bidirectional propagation, adaptive weighting, and energy-based parameter-free attention. It adjusts the contribution of each feature scale and emphasizes informative vehicle responses, reducing feature attenuation during fusion.

2. GCEM jointly models channel and spatial importance to incorporate broader context into vehicle representations. It strengthens target-related responses while suppressing interference from road textures, shadows, and local illumination changes.

3. GDFB combines a Ghost branch for low-cost extraction of local vehicle structure with a DFC branch for long-range spatial modeling. Fusion of these complementary representations strengthens feature expressiveness while limiting computational cost.

Methods

Overall architecture of SGL-YOLO

SGL-YOLO uses YOLO26n [38] as its baseline and retains four main components: input processing, feature extraction and enhancement, multiscale feature fusion, and detection output. SBA-FPN improves feature interactions across scales, GCEM enhances contextual information, and GDFB combines lightweight local feature extraction with long-range spatial modeling. Together, these modifications address scale variation, background interference, and onboard resource constraints (Figure 1).

Input images are resized to 640×640 using LetterBox. Mosaic augmentation, random flipping, and color jitter increase training-sample diversity and support generalization.

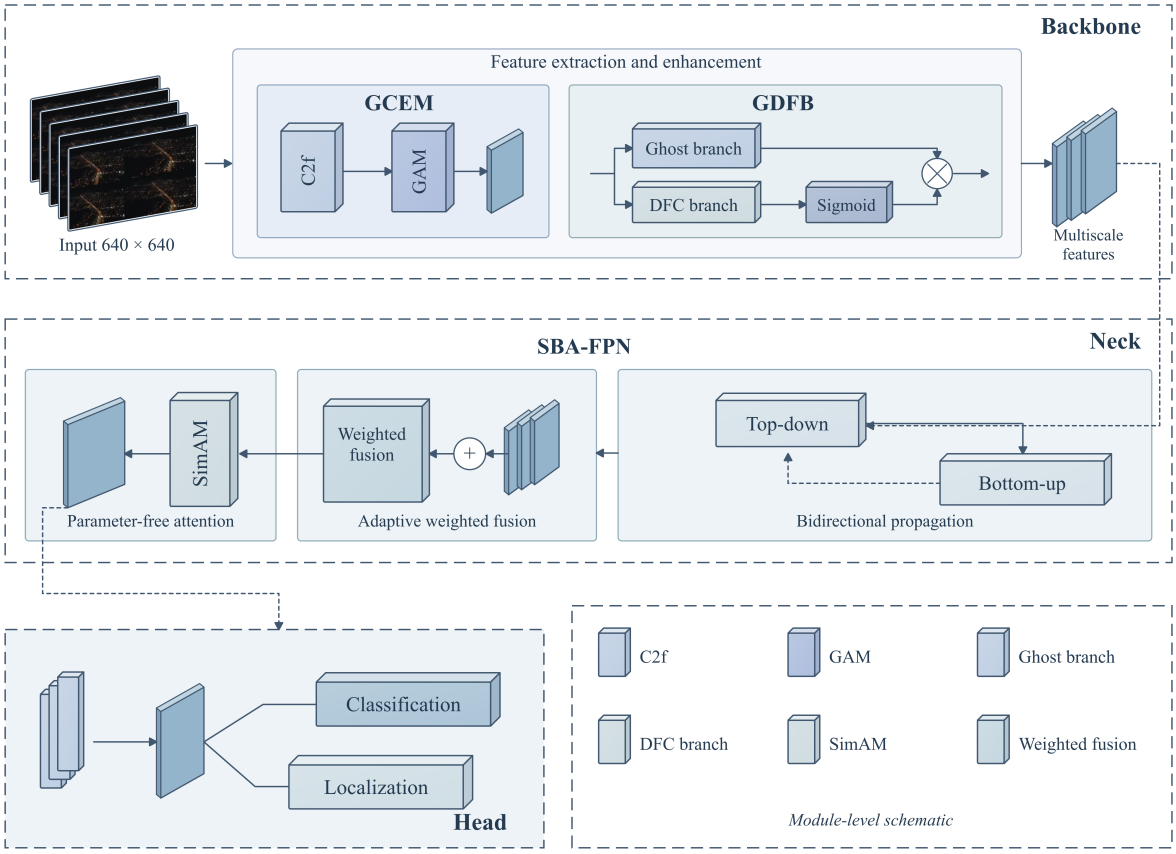


Figure 1. Module-level overview of SGL-YOLO in the Backbone–Neck–Head layout. GCEM and GDFB are grouped by function rather than exact layer placement; SBA-FPN provides multiscale fusion, followed by classification and localization.

During feature extraction, GCEM models channel and spatial importance to strengthen vehicle responses and suppress background interference. GDFB also combines local structural features from its Ghost branch with broader spatial dependencies modeled by its DFC branch.

SBA-FPN then fuses multiscale features through top-down and bottom-up paths. Learnable weights with fast normalization adjust the contributions of different feature scales, while energy-based parameter-free attention emphasizes informative responses and preserves small-vehicle features.

The detection output component uses the enhanced features to predict vehicle classes and bounding boxes. The three modules jointly improve multiscale representation while limiting model complexity and computational cost.

SBA-FPN

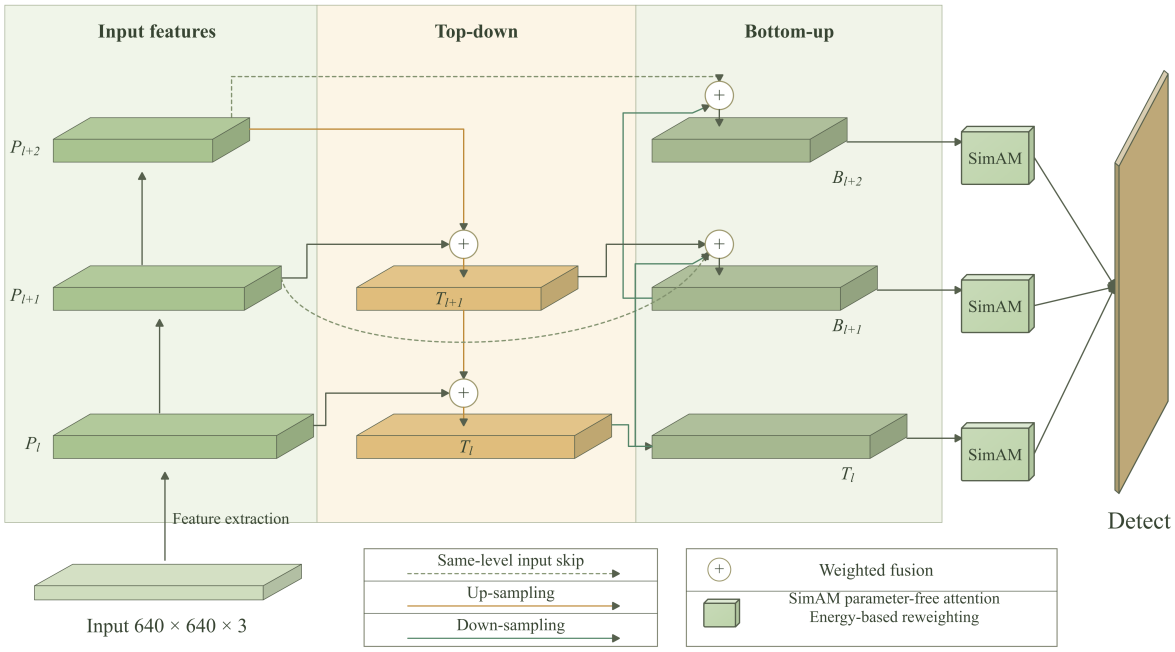
Changes in UAV altitude and vehicle size produce substantial scale variation in aerial imagery. The PAN-FPN in the YOLO26 baseline propagates information across feature levels but may inadequately account for their differing contributions during fusion. This

can weaken features associated with small vehicles. SBA-FPN is designed to address this limitation.

SBA-FPN combines the bidirectional propagation strategy of the bidirectional feature pyramid network (BiFPN) [39] with similarity-based parameter-free attention (SimAM) [40]. Top-down and bottom-up paths exchange information across feature resolutions [41, 42]. Learnable normalized weights adjust each input's contribution, and energy-based attention emphasizes informative target responses.

SBA-FPN comprises bidirectional feature propagation, adaptive weighted fusion, and parameter-free attention. Its fusion topology and attention structure are illustrated in Figure 2a,b.

a



b

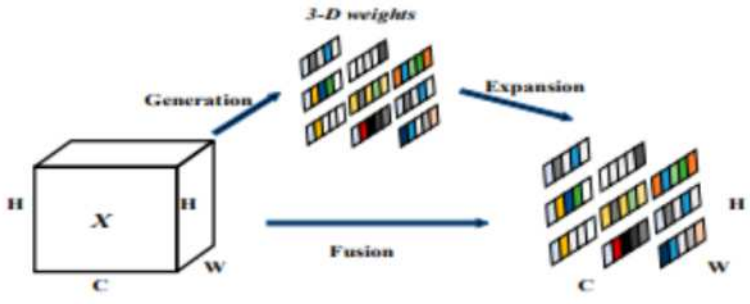


Figure 2. Bidirectional weighted feature fusion and SimAM in SBA-FPN. (a) Three representative feature levels illustrate top-down and bottom-up propagation; the indices do not specify the implemented scale count. (b) SimAM parameter-free attention in SBA-FPN, following Yang et al. [40].

SBA-FPN removes intermediate nodes that receive only one input and therefore perform no multisource fusion. It also connects original input features to output nodes at the

same scale, reducing redundant paths while preserving input information.

The top-down and bottom-up paths exchange high-level semantic information and low-level spatial details. Stacking these bidirectional fusion layers further strengthens interactions across scales.

Because input features contribute unequally to the fused representation, SBA-FPN assigns them learnable weights using fast normalization:

Fast-normalized feature fusion:

$$O = \sum_i \frac{w_i}{\varepsilon + \sum_j w_j} \cdot I_i \tag{1}$$

Here, O denotes the output features, w_i the learnable node weights, and I_i the input features; $\varepsilon = 0.0001$ ensures numerical stability.

After weighted fusion, SBA-FPN applies energy-based parameter-free attention. Neuron importance is estimated from its difference from other neurons in the same channel. This produces three-dimensional attention weights without an additional parameterized attention branch.

The energy function for a target neuron is defined as:

$$\begin{aligned} e_t(w_t, b_t, y, x_i) = \frac{1}{M-1} \sum_{i=1}^{M-1} [-1 - (w_i x_i + b_t)]^2 \\ + [1 - (w_t t + b_t)]^2 + \lambda w_i^2 \end{aligned} \tag{2}$$

Here, t, i, and x denote the target neuron, spatial position index, and other neurons in the same channel of the input tensor X, respectively. M is the number of neurons in the channel, and y measures neuron importance. The weight w and bias b are given by:

$$w_t = -\frac{2(t - \mu_t)}{(t - \mu_t)^2 + 2\delta_t^2 + 2\lambda} \tag{3}$$

$$b_t = -\frac{1}{2}(t + \mu_i) w_t \tag{4}$$

Here, μ_t and δ_t^2 are the mean and variance of the other neurons in the same channel, excluding the target neuron:

$$\mu_t = \frac{1}{M-1} \sum_{i=1}^{M-1} x_i \tag{5}$$

$$\delta_t^2 = \frac{1}{M-1}\sum_{i=1}^{M-1}(x_i - \mu_t)^2 \tag{6}$$

By combining bidirectional propagation, adaptive fusion, and parameter-free attention, SBA-FPN improves the selection and integration of multiscale vehicle features with limited additional complexity.

Global context enhancement module (GCEM)

Low illumination, uneven lighting, reflections, shadows, and background clutter reduce vehicle contrast and obscure edges and textures in nighttime UAV imagery. Strong background responses can further interfere with feature extraction, particularly for small vehicles. Broader contextual information is therefore needed to distinguish vehicle cues from background patterns.

GCEM augments C2f-based feature extraction with global context (Figure 3). Although C2f aggregates local features through cross-stage fusion, its convolutional operations primarily capture local neighborhoods. GCEM supplements these representations to strengthen vehicle-related responses and suppress distracting background structures.

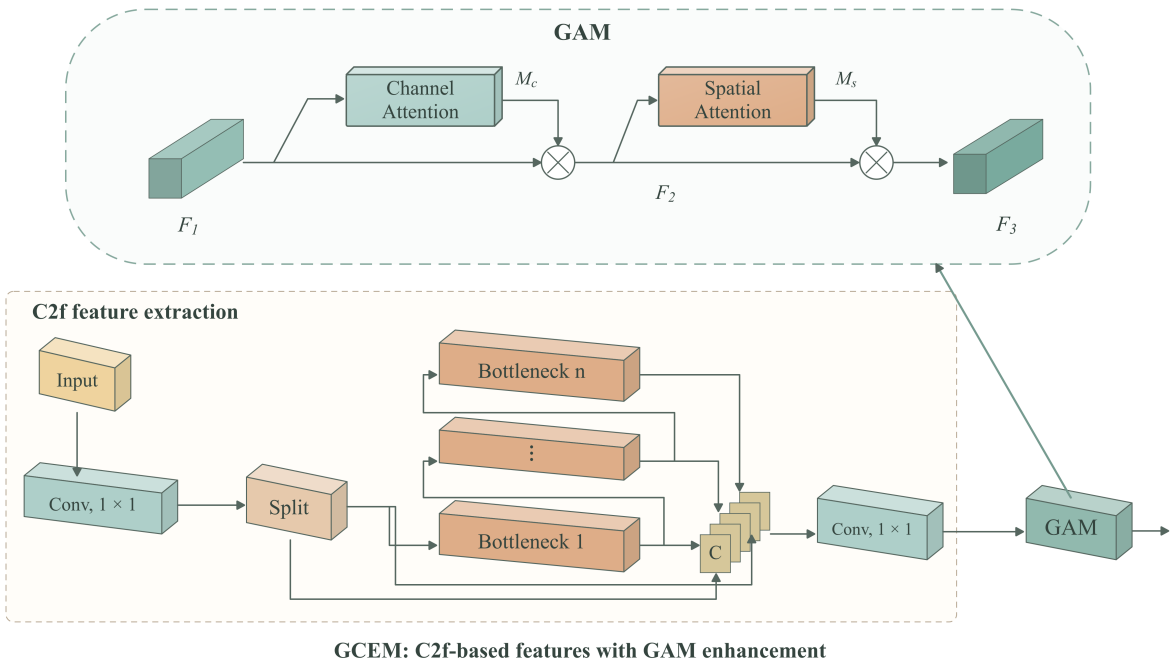


Figure 3. Conceptual GCEM schematic combining C2f-based feature extraction with GAM enhancement. The upper panel expands channel and spatial reweighting. The Bottleneck sequence is symbolic: n does not specify the implemented depth.

GCEM uses a global attention mechanism (GAM) [43] to model channel and spatial importance. The channel branch evaluates dependencies among channels, which respond

to patterns such as vehicle contours, illumination cues, and road textures. It emphasizes channels containing discriminative vehicle information and suppresses redundant responses.

The spatial branch then highlights informative target locations and suppresses irrelevant background regions. Together, the two branches incorporate broader contextual relationships to improve target–background discrimination.

As shown in Figure 4, GAM applies channel attention followed by spatial attention. The channel branch first permutes the dimensions of the C2f feature map. A two-layer multilayer perceptron (MLP) then models cross-channel dependencies and reweights channel responses, producing the intermediate feature F2.

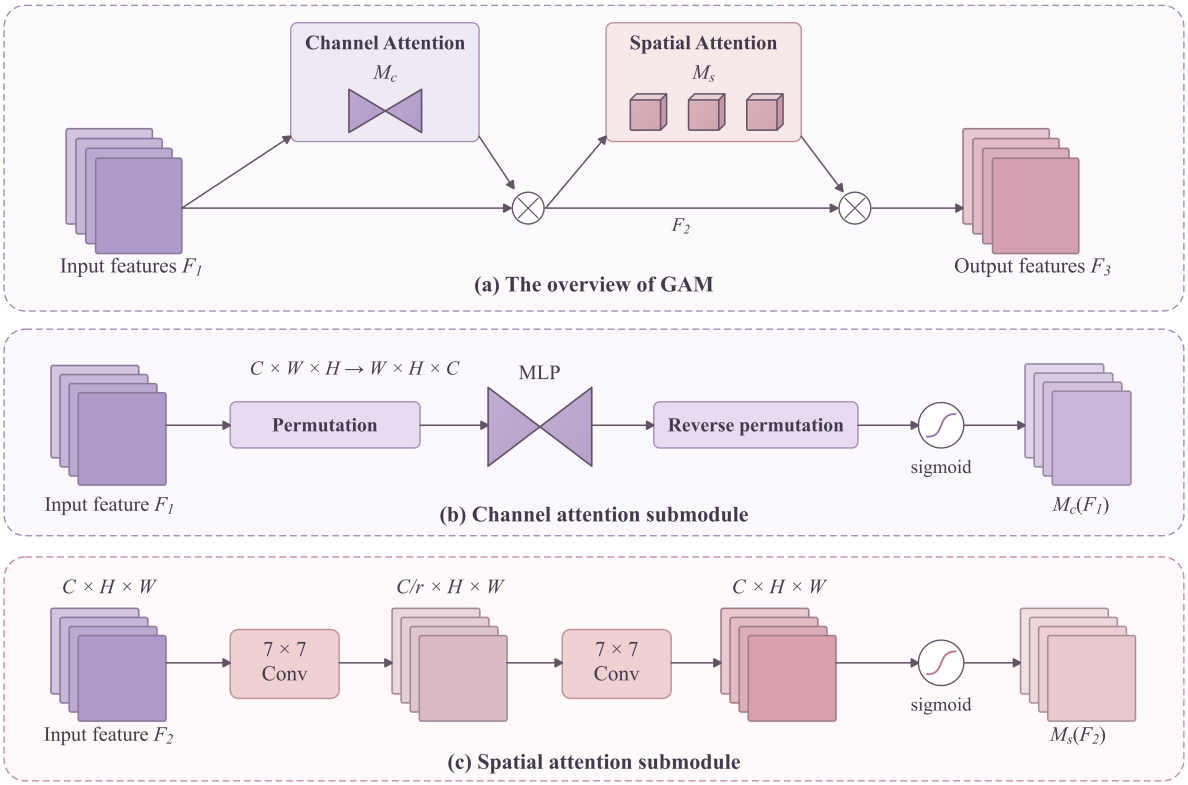


Figure 4. Global attention mechanism (GAM) [43] used in GCEM.

The spatial branch processes F2 through two consecutive 7×7 convolutions to generate a spatial attention map. This map emphasizes informative vehicle locations and suppresses responses associated with shadows, textures, reflections, and uneven lighting. Feature enhancement is expressed as:

$$F_2 = M_c(F_1) \otimes F_1 \tag{7}$$

$$F_3 = M_s(F_2) \otimes F_2 \tag{8}$$

Here, $M_c(\cdot)$ and $M_s(\cdot)$ denote the channel and spatial attention mappings, respectively,

and $\otimes$ denotes element-wise multiplication.

Ghost-DFC fusion block (GDFB)

Vehicles viewed from high altitudes occupy few pixels, while UAV edge devices have limited computing and storage resources. Increasing convolutional depth or channel count raises computational cost, whereas lightweight local convolutions have limited capacity to model long-range dependencies. GDFB addresses this trade-off by combining low-cost Ghost feature generation [44] with the long-range spatial modeling of DFC attention [45]. The resulting representation integrates local vehicle structure with broader context.

For input features $X \in \mathbb{R}^{H\times W\times C}$, where H, W, C denote feature-map height, width, and channel count, respectively, the Ghost branch first generates a small set of intrinsic feature maps using standard convolution. Inexpensive linear transformations, implemented as 3×3 or 5×5 depthwise convolutions, generate additional features. The two sets are concatenated along the channel dimension to obtain the local representation $v(X)$. Exploiting feature redundancy reduces expensive convolutional operations while preserving vehicle edges and contours [44].

Although the Ghost branch reduces extraction cost, 3×3 and 5×5 depthwise convolutions and 1×1 pointwise convolutions provide limited spatial interaction. Local responses can therefore remain sensitive to road textures, shadows, and uneven lighting. GDFB supplements them with decoupled fully connected attention (DFC attention) [45], which aggregates information separately along the height and width dimensions. This captures long-range dependencies with modest additional computation:

$$a'_{hw} = \sum_{h'=1}^{H} f_H \odot x_{h'w}, \quad h' = 1, 2, \ldots, H \tag{9}$$

$$a_{hw} = \sum_{w=1}^{W} f_W \odot a'_{hw}, \quad w = 1, 2, \ldots, W \tag{10}$$

Here, $x_{h'w}$ is the element of the input feature X at row h' and column w. The weights f_H and f_W govern aggregation along the height and width dimensions, respectively, and $\odot$ denotes an element-wise operation. Information is first aggregated along one spatial direction and then the other to obtain the DFC response a_{hw}. This decoupled aggregation captures dependencies across regions without explicitly constructing global self-attention [45].

The input X is processed by two complementary branches. The Ghost branch produces the local representation $v(X)$, while the DFC branch generates the spatial attention map A. This attention map adaptively modulates the local features to produce the output:

$$O = Sigmoid(A) \odot v(X) \tag{11}$$

Here, $\odot$ denotes element-wise multiplication, $Sigmoid$ maps the attention values A to $(0, 1)$, and O is the GDFB output. Long-range spatial attention reweights the local vehicle features, combining structural detail with broader context. This design strengthens local–global representation while limiting parameter count and computational cost.

Datasets

Nighttime images were selected from the public DroneVehicle [46] and VisDrone [47] datasets to construct DroneVehicle_Night and VisDrone_Night. These subsets were used to assess nighttime vehicle detection across different UAV scenes.

DroneVehicle_Night retains the five vehicle classes in DroneVehicle [46]: bus, freight car, truck, car, and van. It contains 3,200 training images, 400 validation images, and 400 test images, each with a resolution of 640 × 512. The subset includes varied vehicle types and object scales and was used for training and primary performance evaluation.

VisDrone_Night contains 1,103 nighttime images selected from VisDrone [47], with a resolution of 2000 × 1500. It retains the original 10 detection classes: pedestrian, people, bicycle, car, van, truck, tricycle, awning-tricycle, bus, and motor. Its broader class coverage and more complex scenes provide an additional setting for evaluating model performance across nighttime UAV environments.

Experimental environment and evaluation metrics

All models were trained and tested under the same hardware and software conditions. The experiments were conducted on a Linux platform equipped with an NVIDIA GeForce RTX 4090 GPU (24 GB VRAM), using PyTorch 2.1.0 and Python 3.11. Training was performed for 500 epochs with a batch size of 32 and an initial learning rate of 0.01.

Detection performance was evaluated using precision, recall, F1-score, and mean average precision (mAP). Precision is the proportion of predicted positives that are correct; recall is the proportion of actual positives detected. F1-score is their harmonic mean. The mAP aggregates average precision across classes. These metrics are defined in Equations (12)–(16).

$$Precision = \frac{TP}{TP + FP} \tag{12}$$

$$Recall = \frac{TP}{TP + FN} \tag{13}$$

$$AP = \int_0^1 P(R)dR \tag{14}$$

$$mAP = \frac{\sum_{q=1}^{Q} AP(q)}{N} \tag{15}$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \tag{16}$$

Here, TP, FP, and FN denote the numbers of true positives, false positives, and false negatives, respectively.

Model complexity was additionally characterized by the number of trainable parameters (Params), floating-point operations per forward pass (FLOPs), and model weight file size. These metrics describe parameter count, computational demand, and storage requirements, respectively. Together with detection metrics, they support assessment of suitability for resource-constrained UAV devices.

Results

Detection performance on DroneVehicle_Night

To verify its superiority, SGL-YOLO is benchmarked against mainstream detectors (EfficientDet, DETR, YOLOv10, YOLO11, Swin-T [48], RT-DETR, and YOLOv12) on the DroneVehicle_Night dataset. As reported in Table 1, our method ranks first across all metrics, attaining 65.2% precision, 55.3% recall, 59.8% F1-score, and 57.3% mAP.

Table 1. Detection performance on DroneVehicle_Night.

Model	Precision (%)	Recall (%)	F1 (%)	mAP (%)
EfficientDet	58.6	48.4	53.0	50.4
DETR	59.4	49.1	53.8	51.2
YOLOv10	60.1	50.0	54.6	52.0
YOLO11	61.2	50.8	55.5	52.9
Swin-T	61.7	51.4	56.1	53.7
RT-DETR	62.1	51.9	56.5	54.4
YOLOv12	62.7	52.8	57.3	54.8
Ours	65.2	55.3	59.8	57.3

Specifically, SGL-YOLO outperforms EfficientDet by substantial margins of 6.6, 6.9, 6.8, and 6.9 percentage points in precision, recall, F1-score, and mAP. Against Transformer-based architectures, it yields mAP gains of 6.1, 3.6, and 2.9 percentage points over DETR, Swin-T, and RT-DETR, respectively.

Furthermore, SGL-YOLO consistently surpasses recent YOLO variants, outperforming YOLOv10, YOLO11, and YOLOv12 by 5.3, 4.4, and 2.5 percentage points in mAP. Even against the strongest baseline, YOLOv12, SGL-YOLO boosts precision from 62.7% to 65.2%, recall from 52.8% to 55.3%, F1-score from 57.3% to 59.8%, and mAP from 54.8% to 57.3%.

Detection performance on VisDrone_Night

To assess generalization in a different nighttime UAV scenario, the same models are evaluated on the VisDrone_Night dataset. As shown in Table 2, SGL-YOLO again ranks first across all four metrics, achieving 53.7% precision, 40.8% recall, 46.4% F1-score, and 43.6% mAP.

Table 2. Detection performance on VisDrone_Night.

Model	Precision (%)	Recall (%)	F1 (%)	mAP (%)
EfficientDet	47.3	34.1	39.6	36.8
DETR	47.9	34.8	40.3	37.5
YOLOv10	48.7	35.6	41.1	38.6
YOLO11	49.4	36.4	41.9	39.4
Swin-T	49.8	36.8	42.3	39.9
RT-DETR	50.6	37.5	43.1	40.5
YOLOv12	51.1	38.2	43.7	41.1
Ours	53.7	40.8	46.4	43.6

Relative to EfficientDet, SGL-YOLO improves precision, recall, F1-score, and mAP by 6.4, 6.7, 6.8, and 6.8 percentage points, respectively. Among Transformer-based detectors, it surpasses DETR, Swin-T, and RT-DETR by 6.1, 3.7, and 3.1 percentage points in mAP.

Compared with YOLO-series models, SGL-YOLO achieves mAP gains of 5.0, 4.2, and 2.5 percentage points over YOLOv10, YOLO11, and YOLOv12, respectively. Even against the strongest competitor, YOLOv12 (51.1% precision, 38.2% recall, 43.7% F1-score, and 41.1% mAP), our model maintains a clear lead of 2.6, 2.6, 2.7, and 2.5 percentage points across the four metrics.

Overall, all evaluated detectors achieve lower scores on VisDrone_Night than on DroneVehicle_Night, likely due to its higher image resolution, broader class diversity, and

greater scene complexity. Nevertheless, SGL-YOLO retains a consistent performance advantage, demonstrating robust generalization across diverse nighttime datasets.

Ablation experiments

The proposed modules were progressively added to YOLO26 to assess their incremental contributions (Table 3). Method (1) adds SBA-FPN; Method (2) additionally incorporates GCEM; and Method (3) adds GDFB to form the complete SGL-YOLO model.

Table 3. Ablation study of the proposed modules.

Model	SBA-FPN	GCEM	GDFB	Params (M)	FLOPs (G)	Precision (%)	Recall (%)	F1 (%)	mAP (%)
YOLO26				5.2	2.4	60.3	52.1	55.9	53.5
Method (1)	✓			4.8	2.1	62.1	53.5	57.5	54.3
Method (2)	✓	✓		4.9	2.1	63.6	54.2	58.5	56.7
Method (3)	✓	✓	✓	4.9	2.1	65.2	55.3	59.8	57.3

The YOLO26 baseline requires 5.2M parameters and 2.4G FLOPs, achieving precision, recall, F1-score, and mAP of 60.3%, 52.1%, 55.9%, and 53.5%, respectively. Adding SBA-FPN reduces parameters to 4.8M and FLOPs to 2.1G while raising these metrics to 62.1%, 53.5%, 57.5%, and 54.3% (+0.8% mAP).

This simultaneous efficiency and accuracy gain aligns with SBA-FPN's lightweight multiscale fusion design. Bidirectional propagation exchanges cross-level features, learnable weights adjust contributions, and energy-based attention highlights informative responses, preserving vehicle features across scales with lower overhead.

Adding GCEM in Method (2) slightly increases parameters to 4.9M while maintaining 2.1G FLOPs, boosting precision, recall, F1-score, and mAP to 63.6%, 54.2%, 58.5%, and 56.7%. This 2.4-percentage-point mAP jump is the largest gain, confirming the value of contextual modeling.

Broader context helps distinguish dim vehicles from road textures, shadows, and lighting artifacts when local cues are weak. GCEM thus effectively complements SBA-FPN with negligible parameter cost.

Integrating GDFB in Method (3) yields the complete SGL-YOLO model without extra parameters (4.9M) or FLOPs (2.1G), achieving top scores: 65.2% precision, 55.3% recall, 59.8% F1-score, and 57.3% mAP. These exceed the baseline by 4.9, 3.2, 3.9, and 3.8 percentage points, while cutting parameters by 0.3M and FLOPs by 0.3G.

GDFB combines local contours from the Ghost branch with spatial context from the DFC branch, aiding small-vehicle localization and reducing background interference while preserving efficiency.

All accuracy metrics steadily improved as modules were added, confirming each component's incremental value and overall lightweight effectiveness.

Comparison of attention mechanisms

To further validate the effectiveness of GCEM, we compare it against classic attention mechanisms—namely SE [49] and CBAM [50]—by integrating each separately into the YOLO26 baseline (Table 4).

Table 4. Comparison of attention mechanisms.

Model	Precision (%)	Recall (%)	F1 (%)	mAP (%)
YOLO26	60.3	52.1	55.9	53.5
+SE	61.2	52.8	56.7	54.1
+CBAM	62.7	53.5	57.7	55.7
+GCEM	63.1	53.8	58.1	56.3

All three modules yield consistent improvements over the baseline. Adding SE elevates precision, recall, F1-score, and mAP to 61.2%, 52.8%, 56.7%, and 54.1%, respectively, corresponding to a 0.6-percentage-point gain in mAP.

CBAM further boosts precision, recall, F1-score, and mAP to 62.7%, 53.5%, 57.7%, and 55.7%, surpassing the YOLO26 baseline and SE in mAP by 2.2 and 1.6 percentage points, respectively. This improvement reflects the advantage of jointly modeling channel and spatial attention.

Notably, GCEM achieves the best performance among all tested configurations, reaching 63.1% precision, 53.8% recall, 58.1% F1-score, and 56.3% mAP. These metrics exceed the baseline by 2.8, 1.7, 2.2, and 2.8 percentage points, while outperforming CBAM by 0.6 percentage points in mAP.

The superiority of GCEM indicates that capturing broader contextual information is critical when low illumination and cluttered backgrounds degrade local vehicle features. By synergizing global context with channel and spatial cues, GCEM significantly enhances target–background discrimination. Overall, these comparative results justify selecting GCEM for SGL-YOLO, as it consistently delivers the highest gains across all four evaluation metrics.

Discussion

SGL-YOLO effectively addresses three core challenges in nighttime UAV detection: extreme scale variation, severe background interference, and constrained onboard resources. Specifically, SBA-FPN enhances multiscale feature fusion, GCEM enriches

global contextual representations, and GDFB couples local structural cues with long-range spatial dependencies via lightweight operations. Their synergistic integration yields consistent performance gains across both evaluated benchmarks.

The proposed architecture balances the inherent trade-offs among existing detection paradigms. Handcrafted features remain vulnerable to illumination shifts and background clutter, whereas Transformer-based models offer broad contextual modeling at the cost of substantial computational and memory overhead. Meanwhile, two-stage detectors introduce latency through region proposal generation. Adopting a one-stage YOLO framework provides an efficient foundation, while our tailored modules overcome the feature representation bottlenecks typical of nighttime UAV imagery. Experimental results confirm that SGL-YOLO surpasses all state-of-the-art baselines across both datasets. Progressive ablation studies validate the incremental contribution of each component, and comparative attention analyses further justify the design of GCEM. Together, these findings demonstrate a clear performance advantage under complex nighttime conditions.

Nevertheless, certain limitations warrant further investigation. Extremely low illumination can obliterate fine textures, edges, and object details prior to feature extraction. Moreover, tiny vehicles captured from high altitudes occupy very few informative pixels, restricting the discriminative cues available for precise localization and classification.

Future work will explore joint low-light image enhancement and high-resolution feature preservation strategies for tiny objects. We also plan to extend evaluations to diverse weather conditions, complex road topologies, and broader UAV scenarios to thoroughly assess generalization. Finally, deploying SGL-YOLO on embedded UAV edge devices will be investigated to benchmark its real-world efficiency in nighttime traffic monitoring.

Conclusion

This study proposed SGL-YOLO, a lightweight YOLO26-based model for nighttime UAV vehicle detection. The model integrates SBA-FPN for multiscale feature fusion, GCEM for contextual enhancement, and GDFB for efficient local–global feature modeling, jointly addressing scale variation, background interference, and computational constraints in nighttime UAV imagery. SGL-YOLO achieved precision, recall, F1-score, and mAP values of 65.2%, 55.3%, 59.8%, and 57.3% on DroneVehicle_Night, respectively, and 53.7%, 40.8%, 46.4%, and 43.6% on VisDrone_Night. It outperformed the selected comparison models on both datasets, while the ablation and attention experiments further demonstrated the effectiveness of the proposed modules. Despite these improvements, detection under extremely low illumination, for tiny objects, and across more diverse scenes remains challenging. Future work will therefore investigate low-light enhancement and improved preservation of small-object features, while

extending evaluation to more diverse environments and exploring deployment on UAV edge devices.

Data Availability

The data used in this study can be obtained from the corresponding author.

References

[1] Yan, Z. & Li, Y. AMSRDet: An adaptive multi-scale UAV infrared-visible remote sensing vehicle detection network. *Sensors* **26,** 817 (2026).

[2] Wang, H. et al. RSW-YOLO: A vehicle detection model for urban UAV remote sensing images. *Sensors* **25,** 4335 (2025).

[3] Ferenchak, N. N. & Abadi, M. G. Nighttime pedestrian fatalities: A comprehensive examination of infrastructure, user, vehicle, and situational factors. *J. Saf. Res.* **79,** 14–25 (2021).

[4] Zhao, Z., Zhang, W., Xiao, Y., Li, C. & Tang, J. Reflectance-guided progressive feature alignment network for all-day UAV object detection. *IEEE Trans. Geosci. Remote Sens.* **63,** 1–15 (2025).

[5] Weng, T. & Niu, X. Enhancing UAV object detection in low-light conditions with ELS-YOLO: A lightweight model based on improved YOLOv11. *Sensors* **25,** 4463 (2025).

[6] Li, C. et al. Low-light image and video enhancement using deep learning: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **44,** 9396–9416 (2022).

[7] Di, X., Cui, K. & Wang, R.-F. Toward efficient UAV-based small object detection: A lightweight network with enhanced feature fusion. *Remote Sens.* **17,** 2235 (2025).

[8] Yang, H., Xiong, Y., Zhang, D., Yan, X. & Hu, X. Lightweight attention-enhanced multi-scale detector for robust small object detection in UAV. *IEEE Signal Process. Lett.* **33,** 271–275 (2026).

[9] Li, G. et al. Cross-modal edge-enhanced detector for UAV-based multispectral object detection. *Sci. Rep.* **16,** 1252 (2026).

[10] Zhuo, Z. et al. TAF-YOLO: A small-object detection network for UAV aerial imagery via visible and infrared adaptive fusion. *Remote Sens.* **17,** 3936 (2025).

[11] Dong, Y. et al. RTUAV-YOLO: A lightweight network for multiscale object detection in UAV imagery. *Digit. Signal Process.* **169,** 105701 (2026).

[12] Adam, M. A. A. & Tapamo, J. R. Survey on image-based vehicle detection methods. *World Electr. Veh. J.* **16,** 303 (2025).

[13] Zhang, W., Wu, Q. M. J., Wang, G. & You, X. Tracking and pairing vehicle headlight in night scenes. *IEEE Trans. Intell. Transp. Syst.* **13,** 140–153 (2012).

[14] Parvin, S., Rozario, L. J. & Islam, M. E. Vision-based on-road nighttime vehicle detection and tracking using taillight and headlight features. *J. Comput. Commun.* **9,** 29–53 (2021).

[15] Wang, J., Sun, X. & Guo, J. A region tracking-based vehicle detection algorithm in nighttime traffic scenes. *Sensors* **13,** 16474–16493 (2013).

[16] Zhang, L., Xu, W., Shen, C. & Huang, Y. Vision-based on-road nighttime vehicle detection and tracking using improved HOG features. *Sensors* **24,** 1590 (2024).

[17] Wu, X., Li, W., Hong, D., Tao, R. & Du, Q. Deep learning for unmanned aerial vehicle-based object detection and tracking: A survey. *IEEE Geosci. Remote Sens. Mag.* **10,** 91–124 (2022).

[18] Zou, Z., Chen, K., Shi, Z., Guo, Y. & Ye, J. Object detection in 20 years: A survey. *Proc. IEEE* **111,** 257–276 (2023).

[19] Cheng, G. & Han, J. A survey on object detection in optical remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **117,** 11–28 (2016).

[20] Xu, Y., Yu, G., Wang, Y., Wu, X. & Ma, Y. A hybrid vehicle detection method based on Viola-Jones and HOG + SVM from UAV images. *Sensors* **16,** 1325 (2016).

[21] Kuang, H., Zhang, X., Li, Y.-J., Chan, L. L. H. & Yan, H. Nighttime vehicle detection based on bio-inspired image enhancement and weighted score-level feature fusion. *IEEE Trans. Intell. Transp. Syst.* **18,** 927–936 (2017).

[22] Khan, S. et al. Transformers in vision: A survey. *ACM Comput. Surv.* **54,** 1–41 (2022).

[23] Carion, N. et al. End-to-end object detection with transformers. In *Computer Vision – ECCV 2020*, 213–229 (2020).

[24] Zhu, X. et al. Deformable DETR: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations (ICLR)* (2021).

[25] Li, F. et al. DN-DETR: Accelerate DETR training by introducing query denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13619–13627 (2022).

[26] Zhao, Y. et al. DETRs beat YOLOs on real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16965–16974 (2024).

[27] Girshick, R., Donahue, J., Darrell, T. & Malik, J. Region-based convolutional networks for accurate object detection and segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **38,** 142–158 (2016).

[28] Girshick, R. Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 1440–1448 (2015).

[29] Ren, S., He, K., Girshick, R. & Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **39,** 1137–1149 (2017).

[30] Xu, Y., Chu, K. & Zhang, J. Nighttime vehicle detection algorithm based on improved Faster-RCNN. *IEEE Access* **12,** 19299–19306 (2024).

[31] He, K., Gkioxari, G., Dollár, P. & Girshick, R. Mask R-CNN. *IEEE Trans. Pattern Anal. Mach. Intell.* **42,** 386–397 (2020).

[32] Cai, Z. & Vasconcelos, N. Cascade R-CNN: High quality object detection and instance segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **43,** 1483–1498 (2021).

[33] Ammar, A., Koubaa, A., Ahmed, M., Saad, A. & Benjdira, B. Vehicle detection from aerial images using deep learning: A comparative study. *Electronics* **10,** 820 (2021).

[34] Liu, W. et al. SSD: Single shot MultiBox detector. In *Computer Vision – ECCV 2016*, 21–37 (2016).

[35] Zhao, P., Ye, X. & Du, Z. Object detection in multispectral remote sensing images based on cross-modal cross-attention. *Sensors* **24,** 4098 (2024).

[36] Wang, A. et al. YOLOv10: Real-time end-to-end object detection. *Adv. Neural Inf. Process. Syst.* **37,** 107984–108011 (2024).

[37] Tian, Y., Ye, Q. & Doermann, D. YOLOv12: Attention-centric real-time object detectors. Preprint at `https://arxiv.org/abs/2502.12524` (2025).

[38] Jocher, G. et al. Ultralytics YOLO26: Unified real-time end-to-end vision models. Preprint at `https://arxiv.org/abs/2606.03748` (2026).

[39] Tan, M., Pang, R. & Le, Q. V. EfficientDet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10781–10790 (2020).

[40] Yang, L., Zhang, R.-Y., Li, L. & Xie, X. SimAM: A simple, parameter-free attention module for convolutional neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, **139,** 11863–11874 (2021).

[41] Lin, T.-Y. et al. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2117–2125 (2017).

[42] Liu, S., Qi, L., Qin, H., Shi, J. & Jia, J. Path aggregation network for instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 8759–8768 (2018).

[43] Liu, Y., Shao, Z. & Hoffmann, N. Global attention mechanism: Retain information to enhance channel-spatial interactions. Preprint at `https://arxiv.org/abs/2112.05561` (2021).

[44] Han, K. et al. GhostNet: More features from cheap operations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1580–1589 (2020).

[45] Tang, Y. et al. GhostNetV2: Enhance cheap operation with long-range attention. *Adv. Neural Inf. Process. Syst.* **35,** 9969–9982 (2022).

[46] Sun, Y., Cao, B., Zhu, P. & Hu, Q. Drone-based RGB-infrared cross-modality vehicle detection via uncertainty-aware learning. *IEEE Trans. Circuits Syst. Video Technol.* **32,** 6700–6713 (2022).

[47] Zhu, P. et al. Detection and tracking meet drones challenge. *IEEE Trans. Pattern Anal. Mach. Intell.* **44,** 7380–7399 (2022).

[48] Liu, Z. et al. Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 10012–10022 (2021).

[49] Hu, J., Shen, L., Albanie, S., Sun, G. & Wu, E. Squeeze-and-excitation networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **42,** 2011–2023 (2020).

[50] Woo, S., Park, J., Lee, J.-Y. & Kweon, I. S. CBAM: Convolutional block attention module. In *Computer Vision – ECCV 2018*, 3–19 (2018).

Funding

This work was supported by the National Natural Science Foundation of China (Grant Nos. 51875009 and 52475502) and the Key Project of the 2025 Educational Science Planning Project of Xinjiang Uygur Autonomous Region (Grant No. HEK2025013).

Author contributions

Zeen Song carried out the research and wrote the manuscript. Tieneng Guo supervised the research.

Additional Information

Competing interests: The authors declare no conflicts of interest.

Figure legends

Figure 1. Module-level overview of SGL-YOLO in the Backbone–Neck–Head layout. GCEM and GDFB are grouped by function rather than exact layer placement; SBA-FPN provides multiscale fusion, followed by classification and localization.

Figure 2. Bidirectional weighted feature fusion and SimAM in SBA-FPN. (a) Three representative feature levels illustrate top-down and bottom-up propagation; the indices do not specify the implemented scale count. (b) SimAM parameter-free attention in SBA-FPN, following Yang et al. [40].

Figure 3. Conceptual GCEM schematic combining C2f-based feature extraction with GAM enhancement. The upper panel expands channel and spatial reweighting. The Bottleneck sequence is symbolic: n does not specify the implemented depth.

Figure 4. Global attention mechanism (GAM) [43] used in GCEM.

Table legends

Table 1. Detection performance on DroneVehicle_Night.

Table 2. Detection performance on VisDrone_Night.

Table 3. Ablation study of the proposed modules.

Table 4. Comparison of attention mechanisms.