Scientific Reports

https://doi.org/10.1038/s41598-026-43560-2

Article in Press

A lightweight feature fusion network for weak and small target detection in remote sensing

Received: 30 November 2025

Accepted: 5 March 2026

Published online: 12 March 2026

Cite this article as: Wu Z., Li N., Tian Z. *et al.* A lightweight feature fusion network for weak and small target detection in remote sensing. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-43560-2

Zhenyuan Wu, Ning Li, Zhengyu Tian & Di Wu

A Lightweight Feature Fusion Network for Weak and Small Target Detection in Remote Sensing

Zhenyuan Wu [1,2], **Ning Li** [1] , **Zhengyu Tian** [3],**and Di Wu** [1,*]

[1] Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China; wuzhenyuan22@mails.ucas.ac.cn (Z.W.); lining@ciomp.ac.cn(N.L.);
[2] University of Chinese Academy of Sciences, Beijing 100049, China
[3] Beijing Jiaotong University, Beijing 100044, China; 25531606@bjtu.edu.cn(Z.T.)
[*] Correspondence: ccgjs3888@163.com (D.W.);

ABSTRACT

Remote sensing imagery presents unique challenges for object detection due to wide fields of view, complex backgrounds, and the dense distribution of small targets, often rendering traditional methods ineffective. To address these limitations, we introduce GSS-YOLO, a lightweight network tailored for remote sensing environments. Our architecture integrates a Spatial Information Aggregation (SIA) module within a Cross-Stage Partial Network (C3) to optimize both detection accuracy and processing efficiency. Furthermore, we incorporate Spatial Pyramid Dilated Convolution (SPD-Conv) to enhance adaptability to low-resolution inputs, and embed a Global Context-Aware Module (GCAM) prior to the detection head to refine multi-scale feature representation. Evaluations on the USOD, VisDrone2019 and DIOR datasets demonstrate that GSS-YOLO achieves superior precision, recall, and robustness across both color and grayscale imagery, all while maintaining a lightweight architecture. Validated by ablation studies, this approach provides an efficient and robust solution for small target detection in complex remote sensing scenarios.

Keywords: dim target detection; lightweight network; remote sensing

Introduction

Optical remote sensing target detection technology relies on sensors operating within the visible light spectrum (0.38 to 0.76 μm) to acquire high-precision aerial and satellite imagery, enabling the precise identification and localization of targets such as aircraft, ships, and buildings. With the advancement of deep learning and Earth observation technologies, this field holds immense potential in applications including urban planning[1], land-use planning[2], traffic management[3], military reconnaissance[4], and disaster emergency response[5]. However, the detection of small targets remains a significant challenge. In high-altitude remote sensing images, small targets typically occupy only 10×10 to 20×20 pixels[6] and exhibit limited features, making it difficult for conventional detection networks to process them effectively, thereby reducing detection accuracy. Consequently, exploring deep learning techniques to overcome these difficulties and improve detection accuracy and efficiency has become a key research focus. While weak and small target detection in unmanned aerial vehicle (UAV) remote sensing images is still in its infancy—having progressed from manually designed features to modern approaches—it continues to face numerous challenges. The most prominent issue is that remote sensing images often encompass complex terrestrial landscapes, such as urban structures, dense forests, and vast water bodies, where backgrounds may possess textures or colors similar to the targets, causing significant interference. Furthermore, imaging conditions are influenced by variable factors such as illumination, season, and weather, which lead to fluctuations in image quality. Therefore, accurately and efficiently identifying small targets against complex backgrounds remains an urgent problem to be resolved in the advancement of remote sensing technology.

The primary challenges confronting remote sensing target detection are mainly concentrated in the following two aspects:

1. Complex background interference: The backgrounds of remote sensing images are intricate, encompassing cloud layers, shadows, and a variety of terrains and landforms, which

significantly interfere with target detection and increase its difficulty.

2. High computational complexity: Existing remote sensing target detection models are characterized by a large number of parameters, intensive convolutional computations, and the necessity for multi-scale processing, resulting in slow speeds and difficulty in meeting real-time requirements.

To address the challenges of weak and small target detection in remote sensing images, this paper optimizes the YOLOv5 framework, selected for its superior balance of accuracy and parameter efficiency relative to newer versions. This approach enhances recognition accuracy and robustness while introducing a lightweight architecture to reduce computational complexity and satisfy real-time processing requirements. The main contributions are summarized as follows:

I. A Global Context-Aware Module (GCAM) is proposed, which integrates multi-scale features through parallel convolutional layers and directional pooling. Under conditions of interference, it prioritizes the capture of positional information of weak and small targets, achieving high-precision detection.

II. Shallow-Deep Information Aggregation Module (SIA) is proposed, which fuses 3×3 and 5×5 convolutional layers to expand the receptive field. By incorporating dropout and MLP layers, it enhances feature fusion, compensates for the limitations of local convolutional operations, accelerates convergence, reduces the number of parameters, and lowers model costs while accelerating model training.

III. The SPD-Conv architecture is introduced, which maintains feature details of small targets through slicing and recombination along with 1×1 convolutions, improving detection performance on low-resolution images and reducing the number of parameters.

The subsequent sections of this paper are arranged as follows: Section II reviews the evolution of target detection technologies, particularly the development of small target detection and attention mechanisms. In Section III, we construct a lightweight detection network based on YOLOv5 and elaborate on its innovative improvements. Section IV comprehensively validates the effectiveness of the proposed improvements through comparative experiments and ablation studies. Finally, we summarize the work presented in this paper and outline future research directions.

Related work

Development of YOLO Model in Object Detection for Remote Sensing

In the field of object detection, there are commonly two types of algorithms: single-stage and two-stage. The single-stage approach directly predicts objects, offering high speed and the ability to process massive amounts of images, making it suitable for high real-time scenarios, albeit with slightly lower accuracy. The two-stage approach first generates candidate regions and then localizes objects, achieving high accuracy and precise identification of complex targets, but it is slower and struggles to meet stringent time-sensitive tasks. Among single-stage detectors, the YOLO model is widely applied in small object detection, with notable YOLO-based small object detection models including Sunflower-YOLO[7], CBGS-YOLO[8], and DSAA-YOLO[9].

LS-YOLO[10] effectively enhances the accuracy of multi-scale landslide detection by constructing a multi-scale landslide dataset and designing a multi-scale feature extraction module. CM-YOLO[11], addressing the issue of cloud and fog interference, proposes a component-decoupled background suppression module and a local-global semantic joint mining module, significantly improving target detection capabilities under complex backgrounds. SEB-YOLO[12] incorporates an adaptive visible-infrared fusion mechanism that dynamically adjusts feature weights in response to variations in environmental illumination and thermal infrared characteristics. This approach significantly enhances detection robustness under complex meteorological or low-light conditions. Furthermore, it establishes a multi-scale feature optimization framework to effectively integrate heterogeneous sensor data, achieving a profound balance between detection accuracy and inference speed. In summary, due to its strong scalability, the YOLO model has been widely applied in small object detection.

The Application of Feature Fusion in Object Detection

In remote sensing image object detection tasks, multi-layer stacking of convolutional and bottleneck structures is commonly employed for single-feature extraction. However, due to the inherent local operation nature of convolutional operations, models struggle to fully capture global information when processing isolated image information, resulting in suboptimal performance in

tasks such as small object detection under complex scenarios. Meanwhile, existing architectures suffer from issues of redundant computations and information loss during information propagation, leading to slow model convergence and insufficient robustness. Consequently, feature fusion is now widely adopted for relevant detection tasks.

In the field of feature fusion, numerous studies have proposed innovative techniques: Bian et al.[13] introduced a multi-scale feature extraction module (MSFE) combined with a cross-stage feature pyramid network (CSFPN) to enhance small object detection capabilities; Xia et al.[14] designed a multi-scale detailed feature fusion module and an object relationship reasoning module, leveraging detailed features and object correlations to optimize fusion performance; Zhang et al.[15] proposed a spatial-frequency fusion framework to detect dim, weak-signal targets by capturing subtle textures and structural features. Leveraging a lightweight attention module, it achieves precise multi-scale reconstruction with minimal parameters, significantly boosting resolution and robustness for real-time embedded remote sensing; Li et al.[16] proposed VDG, a novel driving simulation method that for the first time integrates self-supervised visual odometry into a pose-free dynamic 3D Gaussian splatting framework, eliminating the dependency on expensive sensors or pre-computed poses. By employing a specially designed motion supervision mechanism for precise static-dynamic scene decomposition, VDG outperforms state-of-the-art techniques in reconstruction accuracy; meanwhile, Tang et al.[17] built a hierarchical semantic representation from global scenes to local geometries, enhancing cross-modal matching robustness. By introducing a region-to-point localization architecture, it overcomes traditional point-matching limitations and handles significant viewpoint shifts or occlusions, boosting UAV positioning accuracy and reliability. In summary, feature fusion has been widely adopted in object detection due to its significant performance improvements.

Recent advancements in remote sensing imagery processing have witnessed the emergence of various advanced architectures aimed at addressing feature aggregation and long-range dependency modeling in complex spatial contexts. Regarding spectral-spatial feature fusion, DBMLLA[18] proposed a double-branch Mamba-like linear attention network to achieve global dependency modeling with efficient linear complexity. Subsequently, LKMA[19] further enhanced edge feature representation and spatial-spectral attention fusion by integrating learnable kernels with the Mamba architecture. Moreover, Graph Neural Networks (GNNs) have demonstrated significant advantages in handling non-linear spatial relationships in remote sensing data. The MARP framework[20] achieves automatic adjustment of the feature extraction process in complex noise environments through multi-receptive field adaptive path aggregation. Similarly, the Deep Hybrid network[21] leverages a multi-graph neural network collaborative mechanism to improve feature discriminability with limited samples. These studies demonstrate the pivotal role of multi-path aggregation and global context awareness in enhancing remote sensing image representation.

Multimodal fusion-based target detection in remote sensing images

In the field of remote sensing object detection, multimodal-based approaches are widely adopted. To address the inaccurate fusion caused by imbalanced multimodal information distribution and modal conflicts, researchers proposed CCSFuse[22], a progressive detection framework. This framework achieves complementary enhancement via a cross-modal feature compensation module, and employs an adaptive feature-selection fusion module to dynamically assess modal importance for adaptive fusion. To tackle the computational cost of small object detection in UAV imagery, TAF-YOLO[23] introduces a lightweight and efficient multimodal detection framework. It utilizes a two-branch adaptive fusion network for early information integration, and combines a large adaptive selective kernel with a dual-stream attention bridge, which improves localization accuracy while preserving the fine details of small objects. Furthermore, to overcome the significant discrepancies between RGB and infrared (IR) modalities along with the high computational overhead of complex architectures, LMDENet[24] is designed as a lightweight detection network tailored for low-light remote sensing images. This network employs a heterogeneous backbone to separately model the features of different modalities, and explicitly amplifies complementary information while suppressing redundant noise through a difference-complement enhancement module. Additionally, addressing the spatial misalignment of RGB-IR image pairs caused by sensor discrepancies, ADCNet[25] proposes an adaptive dual-discrepancy calibration network. This method achieves spatial affine alignment of features via a spatial discrepancy calibration module, and utilizes a domain-discrepancy calibration module to separately align object and background features, thereby effectively enhancing the localization performance of multimodal object detection.

Materials and Methods

The Framework of GSS-YOLO Network

To address the limitations of high parameter counts and suboptimal accuracy in small-object detection for drone imagery, we present GSS-YOLO, a lightweight model derived from the YOLOv5 framework. We introduce a Shallow-deep Information Aggregation (SIA) module within the C3 block to reduce model complexity while preserving accuracy. Furthermore, we incorporate Spatial Pyramid Dilated Convolution (SPD-Conv) into the neck's concatenation structure to refine the extraction of subtle features. Finally, a Global Context Awareness Module (GCAM) is embedded prior to the detection layer to enhance feature discrimination and highlight critical contextual information.

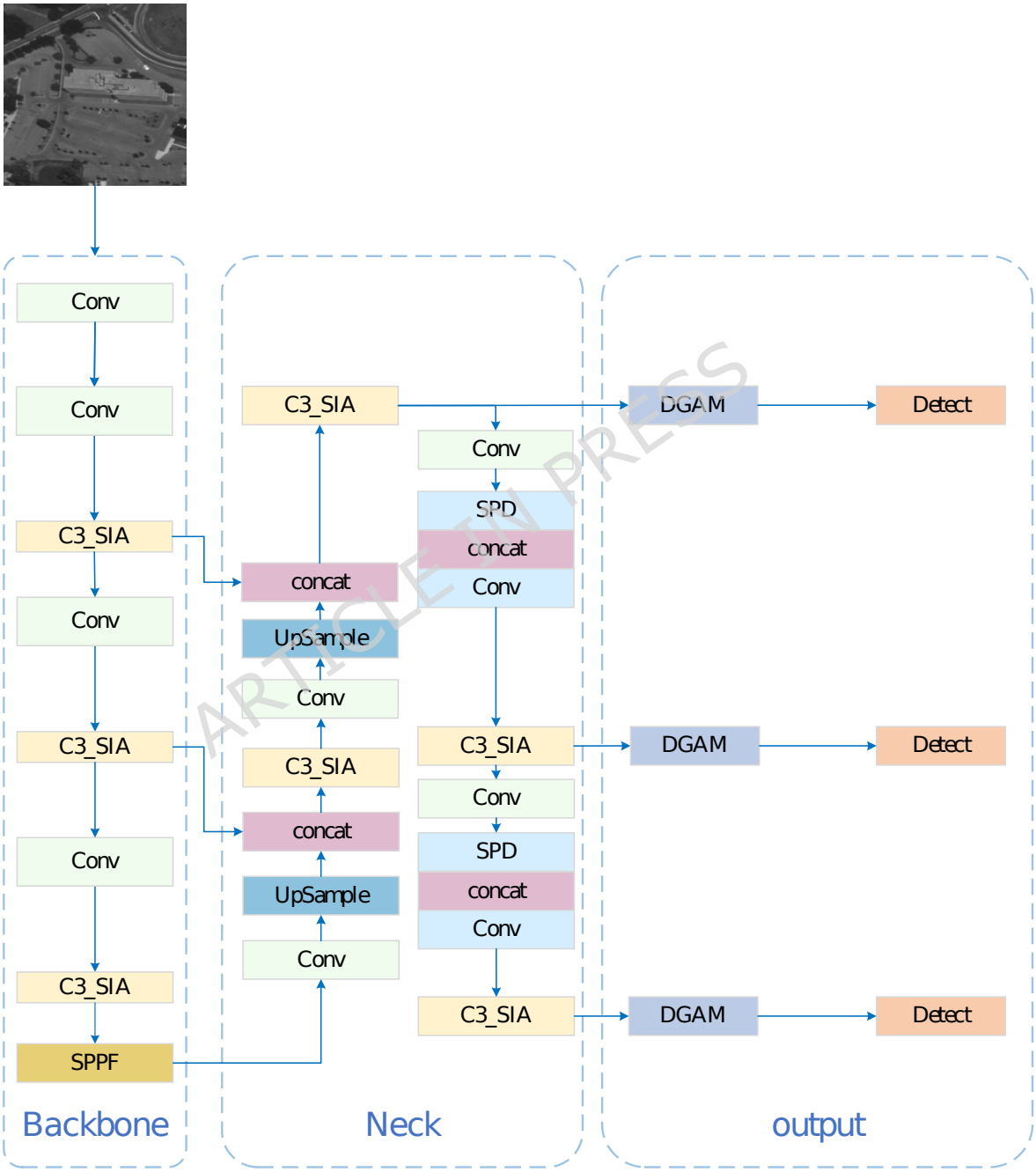


Figure 1. The overall architecture of the GSS-YOLO model. Built upon the YOLOv5 model, we incorporate a newly designed SIA module within the C3 module and utilize the SPD-Conv module for downsampling operations. Furthermore, a GCAM module is integrated prior to the detection stage to strengthen the representation of multi-scale receptive field features.

Global Context-Aware Module

The contextual information from a global perspective in images can facilitate object detection by maximizing the utilization of contextual cues to distinguish the importance of different dimensions in feature maps, emphasizing object-related features, and suppressing background information to

extract critical clues. To better adapt to small object detection, this paper proposes an improved Global Context-Aware Module (GCAM). By integrating different convolutional layers in a parallel manner, the resulting feature matrices processed by these layers are concatenated along the depth dimension to form a unified matrix. This design not only maintains model lightweightness but also significantly enhances multi-scale feature extraction capabilities, particularly improving small object detection accuracy, making it more suitable for deployment on mobile devices.

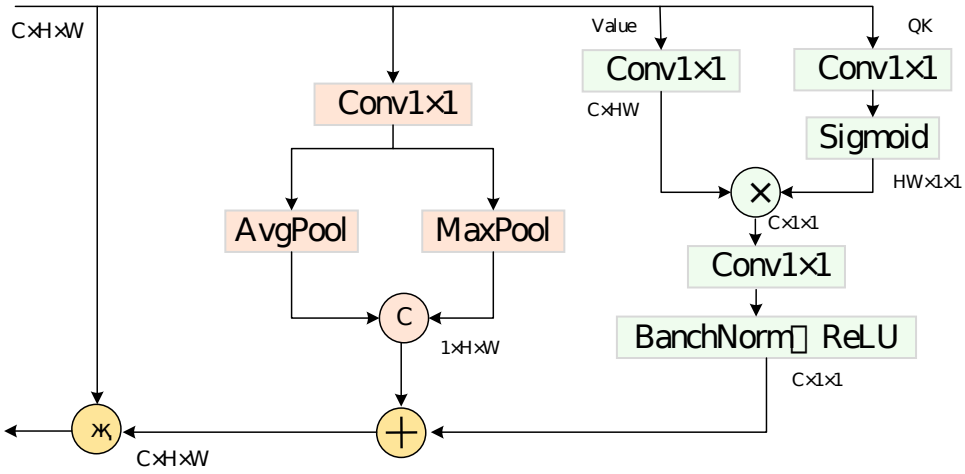


Figure 2. The GCAM structure contains spatial and channel contexts, aiming to capture key features in multiple dimensions and enhance the model's perception and understanding capabilities.

The structure of the GCAM module is illustrated in Figure 2. It comprises two main components. The first component captures directional channel-wise context by integrating global information through Average Pooling (AvgPool) and Max Pooling (MaxPool). By employing directional pooling during channel context acquisition, this module prioritizes positional information of weak and small objects, thereby enhancing detection accuracy. In the figure, C, H, and W denote the number of channels, height, and width of the input feature map, respectively. X-Average Pool and Y-Average Pool represent one-dimensional average pooling operations performed along the horizontal and vertical directions of the input feature map, denoted as M. The operational workflow is described as follows:

$$f = f_{conv}^{1\times1}\left(\left[\frac{1}{W}\sum_{0\le j\le W} m_c(h,j), \frac{1}{H}\sum_{0\le j\le W} m_c(i,w)\right]\right) (1)$$

Specifically, the output feature map f generated in Equation (1) is a direction-aware feature map that explicitly encodes spatial information in both the horizontal and vertical directions. The primary purpose of this directional pooling operation is to decouple global spatial context into two orthogonal one-dimensional feature vectors. Unlike standard 2D global average pooling, which collapses all spatial dimensions and completely loses positional data, this specific design retains the exact coordinate distribution of targets along the X and Y axes. Consequently, this allows the module to generate precise spatial saliency weights, effectively "locking onto" the exact row and column coordinates of weak and small objects, thereby prioritizing the capture of critical positional information amidst complex backgrounds.

The second part of the GCAM module incorporates a parallel structure. Initially, a 1×1 convolution is directly applied to obtain linearly transformed results, referred to as value. Subsequently, another 1×1 convolution combined with a sigmoid function is employed to simplify the scaling of query and key components, denoted as QK. The generated matrices are then multiplied together. The resulting output undergoes further processing through a 1×1 convolution, followed by a Batch Normalization (BN) layer and a ReLU activation function. Finally, the output feature map is derived by applying a broadcast Hadamard product to the channel context, spatial context, and the input feature map. The formula for the parallel structure is expressed as follows:

$$z_i = x_i + W_{v2}\mathrm{ReLU}\left(\mathrm{LN}\left(W_{v1}\sum_{j=1}^{N_p}\frac{e^{W_k x_j}}{\sum_{m=1}^{N_p} e^{W_k x_m}}\right)\right) (2)$$

$$f = \mathrm{Relu}(\mathrm{BN}(\mathrm{Conv}_{1\times1}(X))) (3)$$

The GCAM module captures global context through a decoupled spatial and channel parallel architecture. As illustrated in Figure 2, the module processes the input feature map by splitting it into distinct flows: First, the spatial enhancement branch (middle path) integrates information via a 1×1 convolution, followed by parallel Average and Maximum pooling to extract salient spatial features, resulting in a 1×H×W spatial weight map after concatenation. Concurrently, the channel context branch (right path) employs simplified self-attention logic, splitting the input into Value and QK paths. Through multiplication and subsequent convolution and normalization, it extracts a 1×1 channel importance descriptor. These two types of weights are integrated via additive fusion to form a multi-dimensional attention map, which is then applied to the original feature map through a broadcast Hadamard product. This design significantly enhances the feature representation of weak and small objects by suppressing complex remote sensing background noise.

Since the Global Context-Aware Module (GCAM) can enhance the precise localization of small objects, capture spatial pixel relationships to better distinguish objects from the background, and thereby significantly improve the accuracy of vehicle detection, it is placed before the detection stage rather than integrated into the backbone network. Additionally, GCAM enhances the perception of object position information, enabling more accurate localization, particularly for small objects that typically occupy few pixels, lack distinct features, and are densely distributed. Consequently, GCAM significantly improves both the accuracy and robustness of small object detection. Comparison with Global Context (GC) and Non-Local (NL) Blocks: While GC and NL blocks excel at capturing long-range dependencies, they are often computationally heavy and insensitive to the precise spatial coordinates of tiny objects. GCAM introduces directional pooling (horizontal and vertical) to generate spatial-aware weights. This allows the model to "lock onto" the precise location of weak targets within dense environments, a capability that standard global attention mechanisms often lack.

Shallow-Deep Information Aggregation

In the overall model architecture, the backbone and neck rely on stacked convolutional and bottleneck structures for feature extraction. However, the local operation characteristics of convolution make it difficult to fully capture global information. To address this, the SIA module is designed, The structure is shown in Figure 3. which can effectively fuse deep semantic information, efficiently extract contextual information, and integrate original features to better detect small objects. Meanwhile, by integrating the SIA module with the C3 module, the integrated output is directly fed into the next layer, avoiding redundant computations and information loss. This enables the network to balance local and global features, thereby improving task performance, accelerating convergence speed, and enhancing robustness.

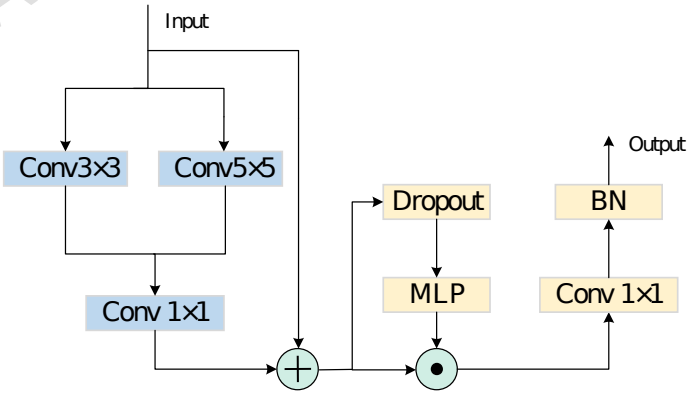


Figure 3. The SIA module, integrating deep semantics, context, and features for better small target detection, is combined with C3 to directly output combined results, avoiding redundancy, balancing features, and boosting performance and convergence.

For this module, given an input feature map $X \in R^{c_1 \times H \times W}$, the module first feeds X into separate 3×3 and 5×5 convolutional layers, yielding two intermediate outputs. Subsequently, 1×1 convolutions are applied to adjust the channels of these intermediate outputs. The adjusted results are then summed with the original feature X to preliminarily expand the receptive field, producing the feature map X_1. Next, X_1 undergoes feature transformation and regularization sequentially through a dropout layer and an MLP layer. Finally, the channel dimensions are adjusted via 1×1 convolution and batch normalization, resulting in the module's final output. The mathematical formulation for this stage is expressed as:

$$Y_1 = f_{conv}^{1\times1}\left(f_{conv}^{3\times3}(X), f_{conv}^{5\times5}(X)\right) \oplus X \quad (4)$$

$$Y_2 = \mathrm{MLP}(\mathrm{Dropout}(Y_1)) \odot X \quad (5)$$

$$Y_3 = \mathrm{BN}\left(f_{conv}^{1\times1}(Y_2)\right) \quad (6)$$

The integration of the SIA module into the C3 network follows a "plug-and-play" strategy within the Cross-Stage Partial (CSP) framework. Specifically, as shown in the figure 4, the input feature map is partitioned into two distinct branches: a direct transition branch and a primary feature extraction branch where the conventional convolution in the bottleneck is replaced by the SIA module as a holistic operator. This design enables the network to effectively compensate for the limitations of local convolutional operations in capturing global information by efficiently extracting and fusing contextual cues with original features. By outputting integrated features directly to the subsequent layer, this architecture avoids redundant computations and information loss, allowing the network to balance local details with global context. Consequently, this integration not only significantly enhances the representation capability and task performance but also accelerates model convergence while bolstering robustness across diverse remote sensing scenarios.

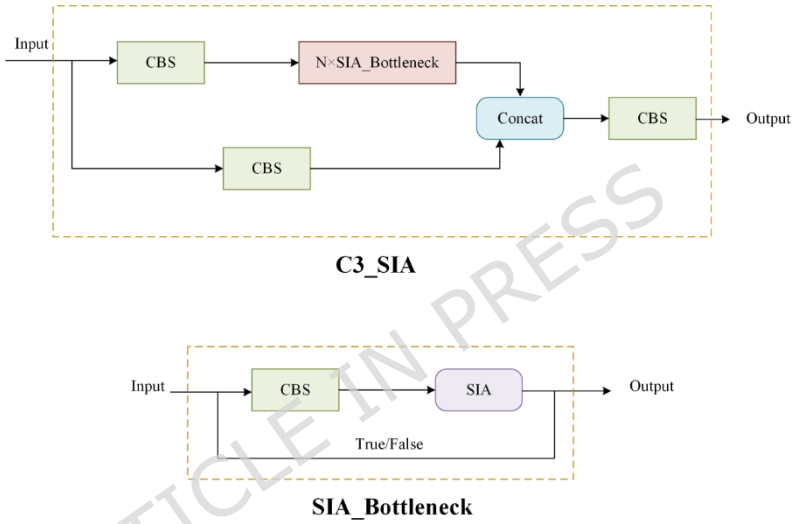


Figure 4. Schematic diagram illustrating the integration of the SIA module and the C3 module.

Unlike the standard C3 module or Inception blocks that primarily increase network width, the SIA module is specifically tailored for remote sensing by integrating 3x3 and 5x5 kernels with an MLP-Dropout architecture. This design not only expands the receptive field but also mitigates the risk of overfitting in complex backgrounds, providing a more robust feature aggregation for weak targets compared to traditional fusion methods.

SPD-Conv

Traditional object detection networks struggle to accurately identify small objects due to inherent architectural limitations. During downsampling via stride convolution and pooling, spatial resolution drops sharply, causing significant feature loss. Small objects' limited size and spatial occupancy make it hard to align with fixed receptive fields, constraining contextual information capture and weakening semantic parsing and spatial localization. Additionally, the network poorly processes spatial context, and combined with reduced spatial resolution, this severely impedes the learning of distinctive small object features, complicating target-distractor distinction in complex scenes and degrading detection performance.

SPD-Conv (Spatial-to-Depth Convolution) is a design (shown in Figure 5) addressing traditional CNN limitations for small objects and low-resolution images. Its core strategies include: (1) replacing stride convolution and pooling layers (which lose fine details) with SPD-Conv; (2) incorporating a spatial-to-depth (SPD) transformation layer to compress channel dimensions via data rearrangement while minimizing information loss; (3) following the SPD layer with a stride-1 convolution to further reduce channels and refine feature extraction. SPD-Conv effectively mitigates information loss, preserves image details, and enhances feature representation, showing significant advantages for small object and low-resolution image tasks.

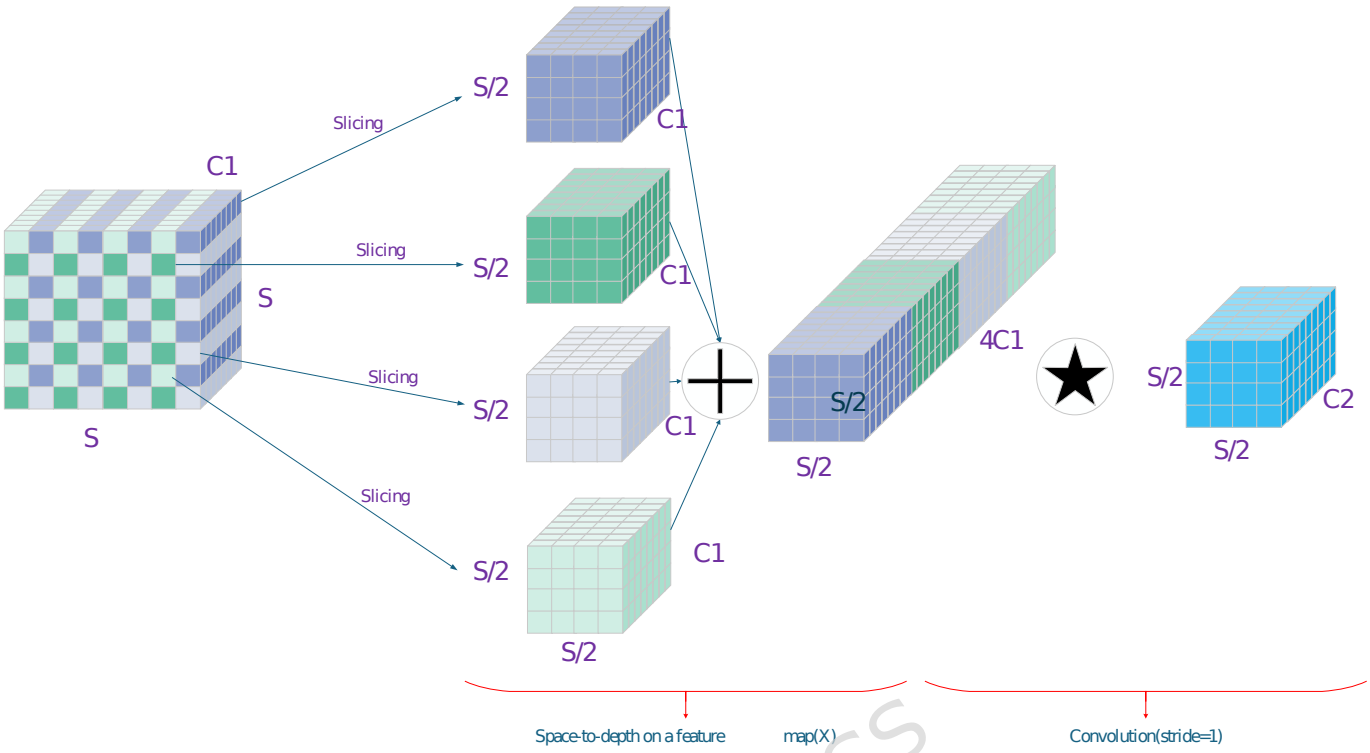


Figure 5. The procedure for downsampling the feature map through SPD-Conv when the scale parameter is set to 2.

As illustrated in Figure 5, SPD-Conv replaces traditional strided convolutions with a Space-to-Depth transformation. For an input feature map of size $S \times S \times C_1$, the process slices it into four sub-feature maps, each with dimensions of $S/2 \times S/2 \times C_1$. These sub-maps are then concatenated along the channel dimension to form an expanded feature map of $S/2 \times S/2 \times 4C_1$. This transformation preserves all spatial pixel information, effectively avoiding the loss of fine-grained details inherent in strided convolutions, followed by a 1×1 convolution to adjust the channel depth.

$$f_{0,0} = X[0:S:scale,0:S:scale], f_{1,0} = X[1:S:scale,0:S:scale], \ldots,$$

$$f_{scale-1,0} = X[scale - 1:S:scale,0:S:scale];$$

$$f_{0,1} = X[0:S:scale,1:S:scale], f_{1,1}, \ldots,$$

$$f_{scale-1,1} = X[scale - 1:S:scale,1:S:scale](7)$$

$$\ldots\ldots$$

$$f_{0,scale-1} = X[0:S:scale,scale - 1:S:scale], f_{1,scale-1}, \ldots,$$

$$f_{scale-1,scale-1} = X[scale - 1:S:scale,scale - 1:S:scale].$$

Here, the scale acts as a scaling factor that precisely determines the downsampling level of the input feature map X. The symbol f(i, j) denotes the sub-feature maps obtained via stepwise slicing operations, where indices i and j start at 0 and increment up to (scale - 1). Each sub-feature map f(i, j) has a spatial size of S/scale × S/scale and retains the same number of channels as the input feature map, which is C1 (where S represents the spatial dimension of the input feature map, and C1 represents the number of input channels). For instance, when scale = 2, the input feature map is evenly partitioned into four sub-feature maps, each with a spatial size of S/2 × S/2 and the same channel count, C1. These sub-feature maps are then concatenated along the channel dimension to produce a new feature map X', whose spatial size becomes S/2 × S/2 while the number of channels increases to 4C1. Finally, a 1×1 convolution kernel with an output dimension of C2 is applied to transform the concatenated feature map X' into the desired size for the target feature map.

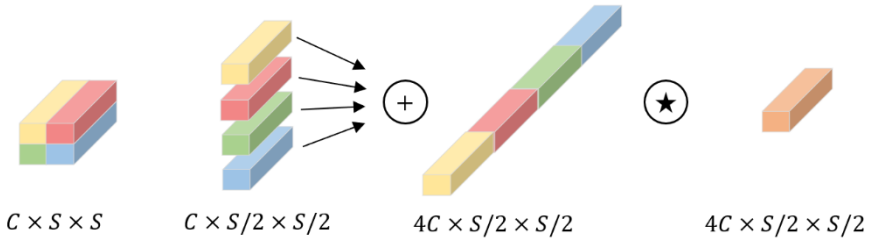


Figure 6. Schematic diagram of SPD-Conv convolution

Comparison with Standard Downsampling, conventional strided convolutions or pooling layers lead to significant loss of fine-grained information, often causing "feature disappearance" for targets smaller than 16x16 pixels. Our SPD-Conv utilizes a slicing and recombination strategy to preserve all pixel details in the channel dimension. Unlike generic SPD implementations, our version is optimized for low-resolution inputs, ensuring that the structural integrity of small targets is maintained throughout the downsampling process.

Experiments

Datasets

In this study, to improve the environmental adaptability and robustness of small object detection models, two datasets are employed to assess the accuracy of the experimental results. The first dataset is USOD, derived from FFCA-YOLO[26]. The second is VisDrone-DET2019[27], introduced by Zhu et al. The third is DIOR[28], proposed by Li et al.

USOD

Another benchmark employed in this study is the publicly available USOD dataset, which is accessible through the research on FFCA-YOLO. This dataset comprises 3,000 images spanning a variety of complex aerial scenarios characterized by low illumination and low resolution, such as parks, streets, forests, and snowy terrains. It features 43,378 meticulously annotated instances of dim and small objects, with all labels professionally calibrated to ensure pixel-level bounding box precision. The dataset is partitioned into training, validation, and test sets according to a 7:2:1 ratio. The most prominent characteristic of USOD is the extreme miniaturization of its targets: quantitative statistics indicate that ultra-small objects (smaller than 16×16 pixels) account for 94.4%, while those smaller than 32×32 pixels reach 99.9%, posing a significant detection challenge within cluttered backgrounds.

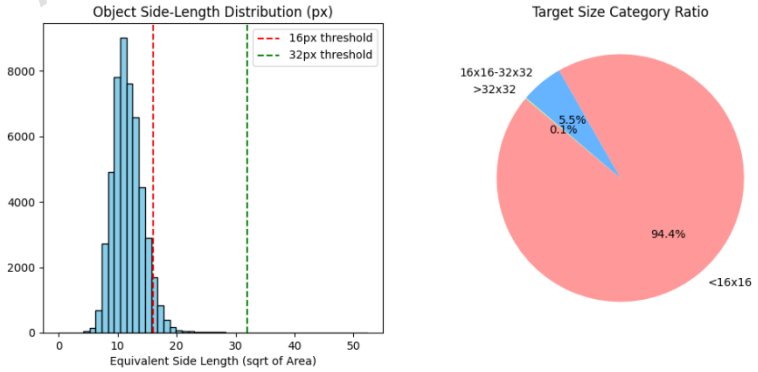


Figure 7. Histogram of the number and distribution of target pixel sizes in the USOD dataset(a)

VisDrone

The VisDrone2019 dataset, released by the AISKYEYE team at Tianjin University, is a large-scale visual benchmark library captured from a drone's perspective. It encompasses a rich diversity of materials collected from dozens of different cities in China, covering various scenarios (e.g., streets, commercial districts, and residential areas) under diverse weather and lighting conditions. The dataset consists of a vast number of high-resolution static images and video sequences, featuring over 2.6 million manually annotated bounding boxes across ten primary object categories, including pedestrians, vehicles, and bicycles. Its core challenges stem from the drone-captured viewpoint, which leads to extreme small object distribution, dense spatial arrangements, and significant perspective variations. Consequently, it has become a critical standard for

evaluating the performance of computer vision algorithms in tasks such as object detection, single-object tracking (SOT), and multi-object tracking (MOT). Compared to traditional street-view datasets, VisDrone2019 is recognized as one of the most authoritative reference benchmarks in the fields of UAV-based perception and smart city surveillance due to its high scene complexity and unique perspective.

DIOR

The DIOR dataset is a comprehensive benchmark widely utilized for object detection in optical remote sensing images. It comprises 23,463 images collected via the Google Earth platform, with spatial resolutions ranging from 0.5 m to 30 m, all of which have been resized to a uniform resolution of 800 × 800 pixels. The dataset contains 192,472 object instances categorized into 20 distinct classes. For model training and performance evaluation, the dataset is partitioned into training, validation, and test sets with a ratio of 7:2:1. The 20 categories are: airplane (AL), airport (AT), baseball field (BF), basketball court (BC), bridge (BD), chimney (CH), dam (DM), expressway service area (ESA), expressway toll station (ETS), golf course (GF), ground track field (GTF), harbor (H), overpass (O), ship (SH), stadium (ST), storage tank (STO), tennis court (TC), train station (TS), vehicle (V), and wind mill (WM).

Experimental Parameter Configuration

This experiment was conducted using the Windows 10 operating system, PyTorch1.13.1 as the deep learning framework, Python version 3.11, and CUDA version 12.1 to accelerate network computations. The hardware configuration included an Intel Core i9-14900KF CPU and an NVIDIA GeForce RTX 4090 GPU with 24GB of memory. Furthermore, during the model training process, the configuration of hyperparameters played a critical role in achieving optimal performance. The detailed hyperparameter settings are presented in Table 1.

Table 1. Hyperparameter configuration for the model training process.

System Configuration	Configuration Parameters
Epochs	150
Batch Size	16
Optimizer	Adam
Learning Rate	0.001
Decay Coefficient	0.0005
Image Input Size	416×416

To ensure the reproducibility of the study, a systematic data augmentation scheme and optimization strategy were employed. For data augmentation, Mosaic augmentation (1.0 probability), random horizontal flipping (0.5 probability), and color jittering in the HSV space (hue 0.015, saturation 0.7, and brightness 0.4) were utilized. Additionally, translation (factor 0.1) and scaling (factor 0.5) were applied to enhance the model's robustness against variations in target position and size. During the model training phase, Stochastic Gradient Descent (SGD) was used as the optimizer, with momentum set to 0.937 and weight decay to 0.0005. The learning rate was managed via a Cosine Annealing scheduler, starting at an initial rate of 0.01 and gradually decaying to 20% of its original value. Prior to formal training, a warmup period of 3 epochs was implemented to facilitate stable weight convergence.

Evaluation Metrics

In the field of object detection, the selection of appropriate evaluation metrics plays a critical role in assessing model performance. In this experiment, we adopted three key metrics: Precision, Recall, and mean Average Precision (mAP). Below is a detailed explanation of these metrics:

Precision evaluates the proportion of true positive (TP) samples among all instances predicted as positive, directly reflecting the model's accuracy and the "purity" of its classifications. Calculated based on the number of correctly identified targets (TP) relative to non-targets incorrectly classified as positive (False Positives, FP), this metric indicates the model's reliability in detecting target objects. The calculation formula for Precision is:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (8)$$

Recall is another critical metric for evaluating the performance of a classifier. It measures the proportion of actual positive samples that are correctly identified by the classifier. Specifically, recall is calculated as the ratio of true positives (TP) to the total number of actual positive samples

(TP + FN). A high recall value suggests that the classifier can effectively identify most positive samples, though this may come at the cost of reduced precision. The calculation formula is:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (9)$$

In complex object detection tasks, mean Average Precision (mAP) serves as a robust metric for assessing overall performance by calculating the mean of Average Precision (AP) values across all categories, where a higher value indicates superior detection capability. This study specifically employs two mAP-derived metrics: mAP50, which measures precision at an Intersection over Union (IoU) threshold of 0.5, and the more stringent mAP50-95, which averages precision across IoU thresholds from 0.5 to 0.9. Since IoU quantifies the overlap between predicted and ground-truth bounding boxes, mAP50-95 provides a comprehensive assessment of the model's detection stability and robustness under varying overlap conditions. The calculation formula for mAP is as follows:

$$\text{mAP} = \frac{\sum_{q=1}^{Q} \text{AP}(q)}{Q} \quad (10)$$

The F1-score is a widely adopted metric in classification tasks that integrates the strengths of precision and recall, offering a more balanced and comprehensive evaluation of model performance. A higher F1-score signifies the model's ability to not only accurately identify positive samples but also effectively reduce the misclassification of negative samples as positive, thereby reflecting an optimal trade-off between precision and recall. Its calculation formula is:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (11)$$

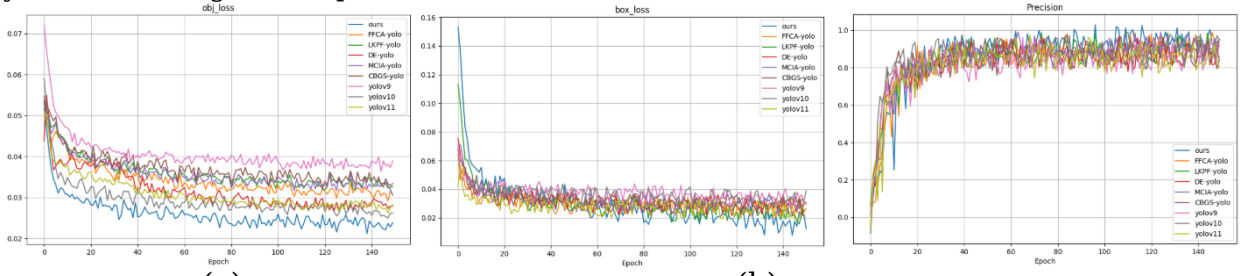
Results

In this experiment, we employed the three aforementioned datasets to evaluate our algorithm. Our model was benchmarked against state-of-the-art approaches, including FFCA-YOLO[26], LKPF-YOLO[29], DE-YOLO[30], MCIA-YOLO[31], CBGS-YOLO[8], YOLOv9[32], YOLOv10[33], and YOLOv11[34], across all three datasets. Furthermore, comprehensive ablation studies were conducted to validate the effectiveness and reliability of the proposed modules. The detailed experimental procedure is described below.

Comparative Experiments on the USOD Dataset

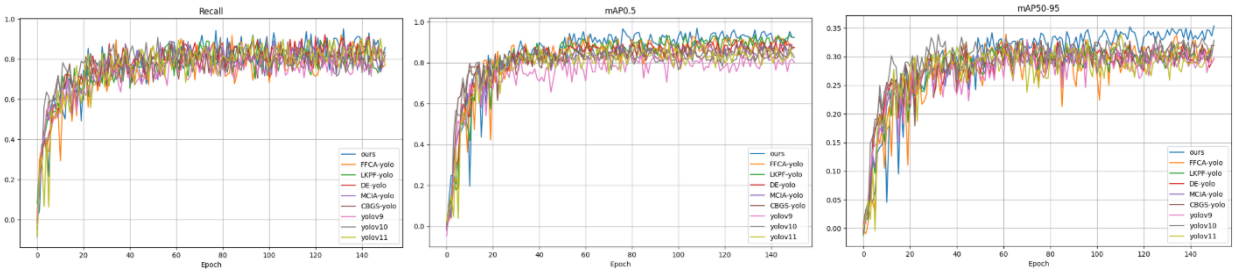
To more comprehensively evaluate the effectiveness and robustness of the GSS-YOLO model, we specifically selected the USOD dataset for in-depth comparative analysis. A notable feature of this new dataset is its relatively low image clarity and grayscale format, which introduces additional challenges for the detection task. We performed comparative experiments using the same model configuration as in the first dataset to objectively demonstrate the superior detection capabilities and optimization performance of the GSS-YOLO model in complex and low-quality image environments. This study not only validates the universality of GSS-YOLO but also underscores its significant advantages in handling challenging grayscale and low-resolution images.

Figure 8 clearly illustrates the downward trends of box_loss and obj_loss. It is evident that the loss curve of the GSS-YOLO model converges toward zero more rapidly, demonstrating the model's fast training convergence, its ability to efficiently learn data features, and its superior learning efficiency and performance optimization capabilities. Furthermore, by analyzing the performance change curves of the other four metrics, it can be observed that as training progresses, GSS-YOLO consistently outperforms the other eight comparison models across all four evaluation metrics. This result not only fully validates the exceptional detection capability of the GSS-YOLO model on complex datasets but also underscores its significant contributions to improving detection accuracy and enhancing overall performance.



(a) (b)

(c)



(d) (e)

(f)

Figure 8. Comparison of training result metrics between GSS-YOLO and FFCA-YOLO, LKPF-YOLO, DE-YOLO, MCIA-YOLO, CBGS-YOLO, yolov9, yolov10, and yolov11 on the USOD dataset. (a) obj_loss (b) box_loss (c) Precision (d) Recall (e) mAP50 (f) mAP50-95

The experimental results presented in Table 2 indicate that the GSS-YOLO model surpasses other models across all five evaluation metrics, achieving improvements of 1.7%, 2.5%, 4.2%, 1.5%, and 1.1%, respectively, compared to the best-performing metrics among the comparison models. This demonstrates the effectiveness of the enhancements incorporated into our proposed model. Furthermore, the model not only effectively detects weak and small targets in color images but also achieves high accuracy in grayscale images. Regarding detection speed, the GSS-YOLO model ranks among the top performers across all evaluated models, ensuring both high-speed detection and superior performance in other key metrics. In summary, the GSS-YOLO model introduced in this paper achieves significant improvements in the accuracy of detecting weak and small targets while maintaining a lightweight design.

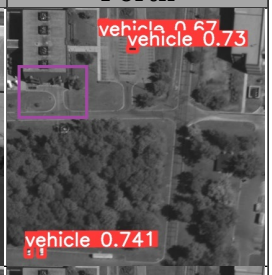
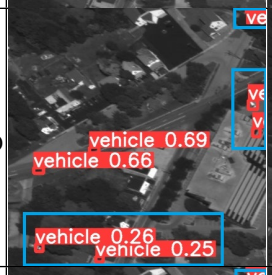
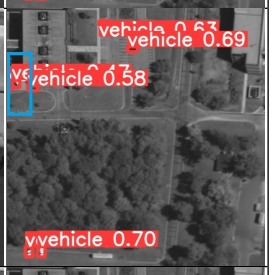
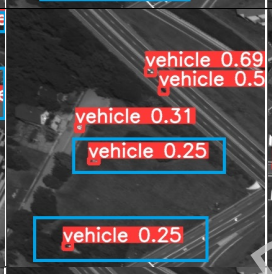
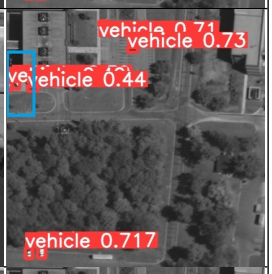
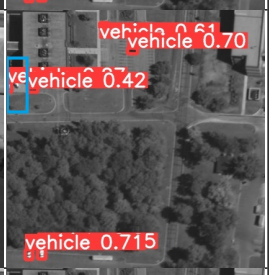
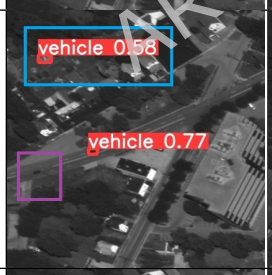
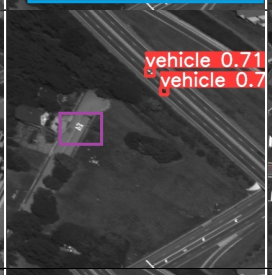
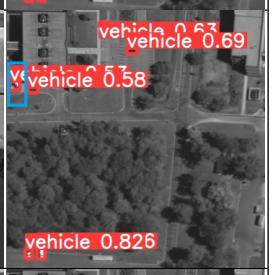
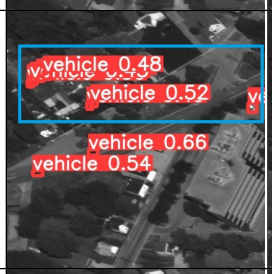
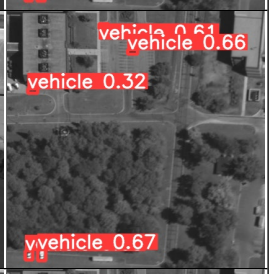
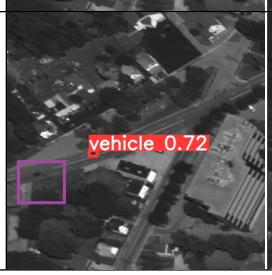
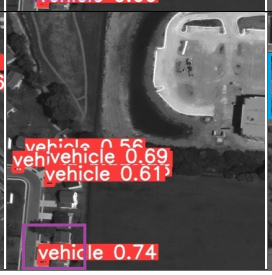
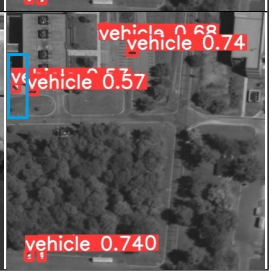
Crucially, GSS-YOLO achieves state-of-the-art accuracy while maintaining a highly lightweight architecture. As indicated in the table, our model possesses the fewest parameters (5.04M) among all compared methods, yet it delivers the highest inference speed of 482 FPS. Compared to the second-best model in terms of accuracy, CBGS-YOLO (5.12M parameters and 457 FPS), GSS-YOLO further reduces the model size while improving processing efficiency. This optimal trade-off between model compactness and high-speed throughput underscores the effectiveness of our lightweight design, making it particularly suitable for deployment on hardware-constrained remote sensing platforms.

Table 2. A comparison of the other eight models and GSS-YOLO on the test data of the USOD dataset. (The bolded items indicate the performance indicators with the highest values across all options, and all subsequent tables adhere to this formatting convention.)

Method	Precision	Recall	F1	mAP50	mAP50-95	Parameters	FPS
Yolov9	0.801	0.697	0.745	0.742	0.323	6.31	255
Yolov10	0.813	0.724	0.766	0.799	0.318	6.09	240
Yolov11	0.821	0.808	0.814	0.803	0.295	6.88	261
FFCA-YOLO	0.893	0.859	0.876	0.834	0.347	5.33	444
LKPF-YOLO	0.873	0.864	0.868	0.862	0.296	5.55	452
DE-YOLO	0.883	0.877	0.880	0.881	0.281	7.02	460
MCIA-YOLO	0.909	0.818	0.861	0.853	0.301	6.15	415
CBGS-YOLO	0.927	0.883	0.904	0.916	0.339	5.12	457
Ours	**0.938**	**0.894**	**0.915**	**0.921**	**0.343**	**5.04**	**482**

Figure 9 provides a comprehensive and detailed comparison of the weak target detection capabilities of nine models, including GSS-YOLO, across four complex scenarios: multi-target dispersion, moving targets, occluded targets, and multi-target stacking. Analysis reveals that other models exhibit false detections or missed detections in various scenarios, whereas GSS-YOLO demonstrates superior performance with no false or missed detections and achieves the highest detection confidence level. This clearly indicates that GSS-YOLO exhibits exceptional detection performance and robustness in complex environments. Not only does it achieve precise detection, but it also attains an excellent level in confidence assessment, further underscoring the advanced

nature and effectiveness of this model in weak target detection tasks.

	First	Second	Third	Forth
FFCA-YOLO				
LKPF-YOLO				
DE-YOLO				
MCIA-YOLO				
CBGS-YOLO				
Yolov9				
Yolov10				

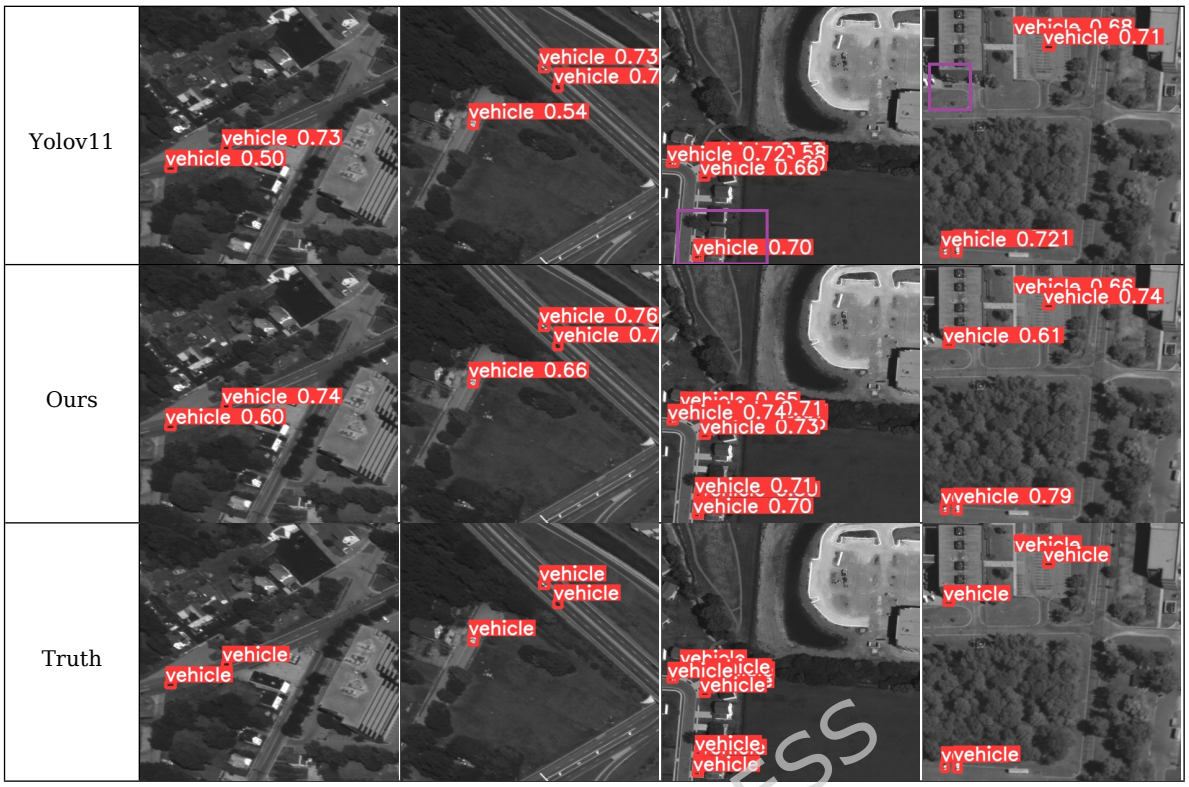


Figure 9. In four representative real-world scenarios, a comparative analysis between the GSS-YOLO model proposed in this paper and several other detection models reveals that the visual results of the detection images clearly highlight its superiority and accuracy. The blue boxes indicate regions that were incorrectly detected, whereas the purple boxes highlight regions that were missed during detection.

The superior performance of GSS-YOLO on grayscale imagery stems from a synergistic mechanism across its core modules. In grayscale scenarios where chromatic cues are absent, the SIA module enhances the discriminability of subtle edges and structural features through multi-scale receptive field aggregation during the early stages of feature extraction. Subsequently, SPD-Conv ensures that these fine-grained intensity gradients are fully preserved and transferred to the channel dimension via a space-to-depth transformation, effectively avoiding the information loss inherent in traditional convolutions. Finally, the GCAM leverages directional spatial attention to precisely pinpoint targets based on their spatial distribution and geometric signatures without the aid of color. This synergy allows the architecture to prioritize spatial-structural integrity, rendering it inherently robust against the loss of spectral information.

Comparative Experiments on the VisDrone2019 Dataset

The experimental results in the table 3 indicate that the proposed GSS-YOLO model significantly outperforms all baseline models across nearly all core evaluation metrics, demonstrating superior detection performance. In terms of comprehensive performance, GSS-YOLO achieves an mAP50 of 0.509 and an mAP50-95 of 0.296, representing improvements of 5.5% and 3.9% respectively compared to the best-performing baseline (0.424 and 0.227). This strongly validates the effectiveness of the architectural enhancements designed in our model.

Regarding category-specific performance, GSS-YOLO demonstrates a comprehensive advantage across all ten categories of the VisDrone2019 dataset. Our model achieves the top-ranking results in five key categories: Pedestrian (0.551), People (0.499), Truck (0.502), Tricycle (0.555), and Awning-tricycle (0.336), with the latter two showing significant absolute gains of 5.6% and 3.4% over the closest competitors. In the remaining categories, including Bicycle (0.302), Car (0.708), Van (0.547), Bus (0.617), and Motor (0.580), our approach consistently maintains high-level competitiveness as the second-best performer, the gap relative to the state-of-the-art technological achievements remains narrow. These consistent category-wise strengths lead to a superior overall performance, with mAP50 (0.509) and mAP50-95 (0.296) markedly outperforming all baseline models. Notably, these precision gains are achieved alongside an impressive inference speed of 257 FPS, which significantly outpaces high-precision counterparts such as YOLOv10 (214 FPS) and FFCA-YOLO (208 FPS). This optimal trade-off between detection efficacy and architectural lightness confirms that the integration of SIA, SPD-Conv, and GCAM modules effectively

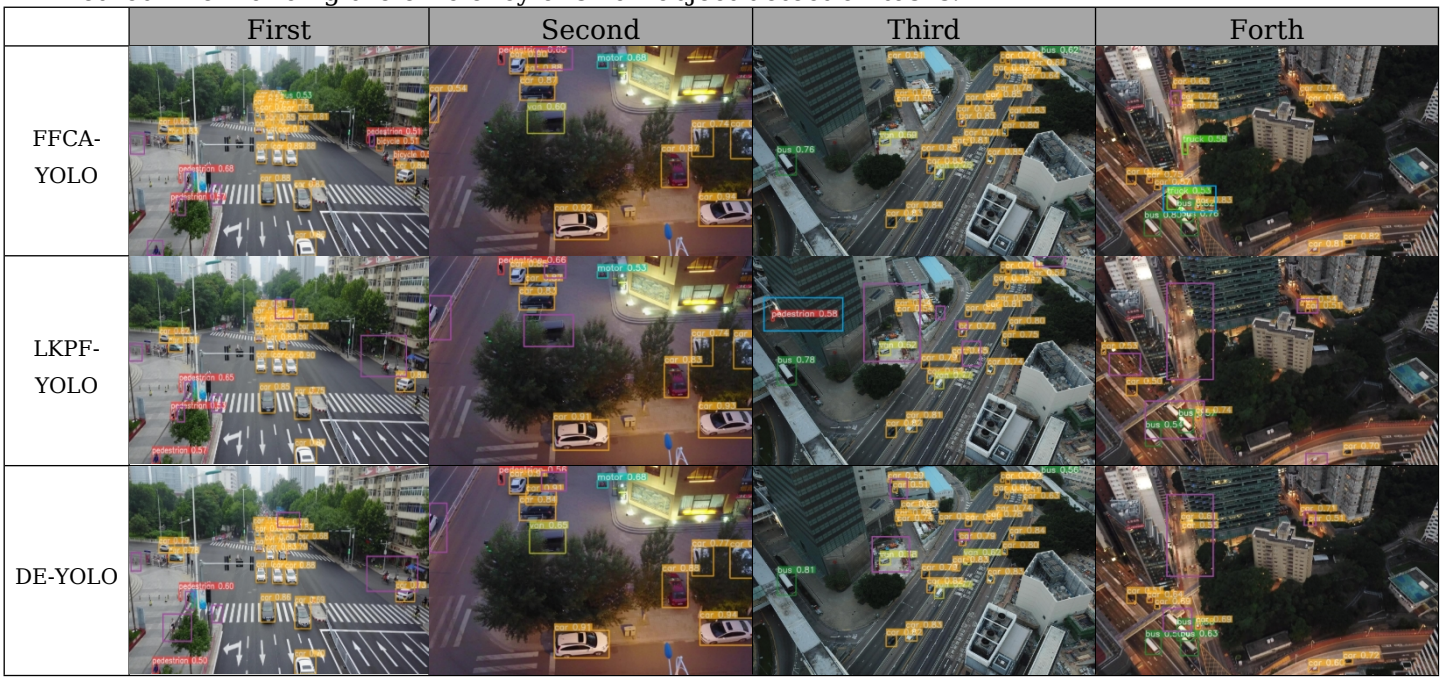
suppresses computational redundancy while capturing fine-grained features of weak and small targets.

In summary, the proposed model not only achieves a breakthrough in overall detection accuracy but also exhibits exceptional balance and robustness in recognizing various traffic participants, effectively addressing diverse detection challenges in complex traffic scenarios.

Table 3. A comparison of the other eight models and GSS-YOLO on the test data of the VisDrone2019 dataset.

Method	Pedestrain	People	Bicycle	Car	Van	Truck	Tricycle	awning-tricycle	Bus	Motor	mAP50	mAP50-95	F
Yolov9	0.303	0.426	0.181	0.431	0.292	0.218	0.394	0.089	0.349	0.418	0.252	0.123	2
Yolov10	0.346	0.425	0.193	0.489	0.329	0.232	0.444	0.236	0.423	0.441	0.289	0.144	2
Yolov11	0.335	0.437	0.179	0.470	0.341	0.238	0.413	0.210	0.423	0.433	0.291	0.147	2
FFCA-YOLO	0.529	0.479	0.297	0.661	**0.552**	0.453	0.489	0.302	**0.618**	0.569	0.424	0.224	2
LKPF-YOLO	0.498	0.478	0.298	**0.737**	0.508	0.423	0.491	0.287	0.567	0.549	0.416	0.211	2
DE-YOLO	0.519	0.492	0.275	0.654	0.519	0.438	0.478	0.289	0.605	0.563	0.456	0.217	2
MCIA-YOLO	0.528	0.477	0.288	0.652	0.493	0.471	0.465	0.231	0.596	0.542	0.411	0.225	2
CBGS-YOLO	0.545	0.489	**0.309**	0.668	0.531	0.431	0.499	0.291	0.601	**0.581**	0.401	0.227	2
Ours	**0.551**	**0.499**	0.302	0.708	0.547	**0.502**	**0.555**	**0.336**	0.617	0.580	**0.509**	**0.296**	**2**

Figure 10 systematically evaluates the detection performance of GSS-YOLO and eight other mainstream models for dim and small objects under complex backgrounds, characterized by dispersed targets, dynamic motion, and severe occlusion. The experimental results indicate that, in contrast to the common defects of missed detections and false alarms prevalent in the baseline models, GSS-YOLO demonstrates superior robustness and detection precision. Notably, our model achieves zero missed or false detections across all tested scenarios and consistently yields significantly higher confidence scores. These findings strongly validate the precise capture capability and reliable confidence estimation mechanism of GSS-YOLO for dim and small targets in complex environments, fully underscoring the advancement and effectiveness of the proposed method in enhancing the efficiency of small object detection tasks.



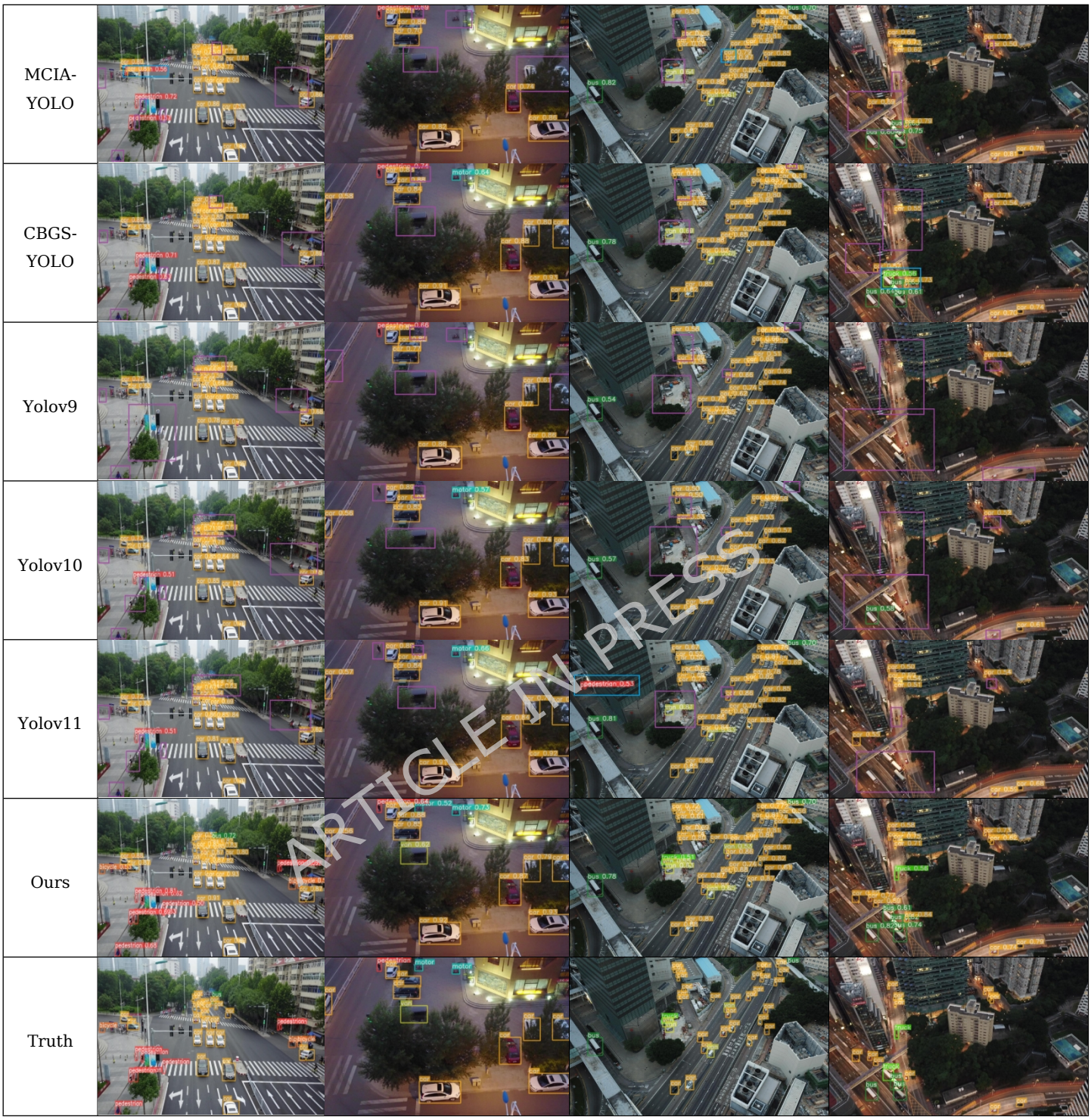


Figure 10. In four representative real-world scenarios, the visual comparisons between the proposed GSS-YOLO and several mainstream models significantly demonstrate its superior detection precision. In the figure, blue boxes mark the false positives, while purple boxes identify the missed detections, intuitively validating the exceptional performance of GSS-YOLO in complex environments.

Comparative Experiments on the DIOR Dataset

In the comprehensive evaluation of the DIOR dataset, as shown in Table 4, the proposed GSS-YOLO exhibits outstanding performance advantages, achieving a core metric mAP50 of 0.822. This represents a significant absolute gain of 3.7% and 4.6% over the state-of-the-art baseline models, LKPF-YOLO (0.785) and FFCA-YOLO (0.776), respectively. This achievement not only marks a successful breakthrough of the 0.80 performance bottleneck but also validates the model's robustness in multi-scale feature extraction within complex remote sensing backgrounds. Through

the deep optimization of the feature fusion mechanism, the model demonstrates enhanced semantic capture capabilities and global consistency when processing densely distributed targets in wide-field-of-view scenarios.

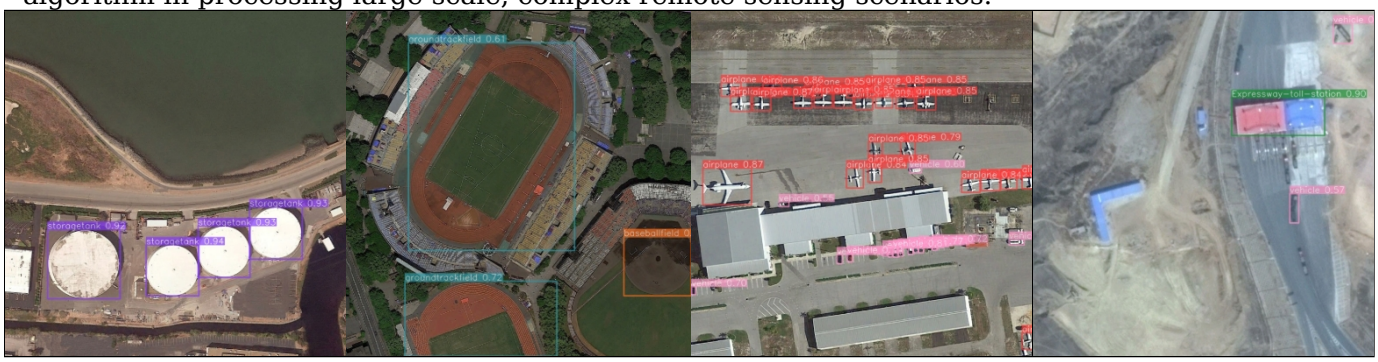
A granular analysis of category-wise performance reveals that GSS-YOLO maintains exceptional adaptability across various complex scenes. In the infrastructure and public facility sectors, the model achieves peak precision in airplane (AL, 0.943), baseball field (BF, 0.921), ground track field (GTF, 0.726), and train station (TS, 0.702), while securing top-tier second-best rankings in categories such as airport (AT), dam (D), overpass (O), and storage tank (ST). For industrial and specific traffic scenarios, the model ranks first in chimney (C, 0.931), expressway service area (ESA, 0.806), ship (SP, 0.876), vehicle (V, 0.598), and wind mill (WM, 0.820), while maintaining highly competitive runner-up performance in expressway toll station (ETS) and stadium (SD). This balanced superiority across categories underscores the comprehensive lead of the architecture in handling targets of varying scales and complex spatial arrangements.

In terms of computational efficiency and real-time performance, this study has achieved a substantial breakthrough, with an inference speed reaching 84.1 FPS. This rate significantly outpaces high-precision algorithms such as DE-YOLO (76.4 FPS) and CBGS-YOLO (77.7 FPS). The seamless synergy between precision and speed fully demonstrates the efficacy of the lightweight architecture in suppressing computational redundancy. This ensures that GSS-YOLO is perfectly suited for resource-constrained UAV embedded platforms, meeting the engineering requirements for real-time and high-efficiency monitoring of complex ground environments.

Table 4. Performance comparison between the proposed GSS-YOLO and eight other mainstream models on the VisDrone2019 test set.

	AL	AT	BF	BC	BD	C	D	ESA	ETS	GF	GTF	H	O	SP	SD	ST	TC	TS	V
Yolov9	0.721	0.684	0.761	0.725	0.582	0.751	0.506	0.630	0.546	0.588	0.578	0.614	0.665	0.692	0.642	0.725	0.838	0.477	0.414
Yolov10	0.711	0.679	0.807	0.716	0.674	0.835	0.507	0.652	0.615	0.564	0.551	0.592	0.765	0.67	0.609	0.654	0.843	0.451	0.465
Yolov11	0.736	0.716	0.720	0.769	0.807	0.820	0.515	0.697	0.669	0.598	0.557	0.794	0.417	0.734	0.718	0.714	0.820	0.505	0.528
FFCA-YOLO	0.898	0.788	0.844	0.825	0.557	0.890	0.672	0.713	0.770	0.679	0.663	0.622	0.588	0.833	0.777	0.781	0.858	0.618	0.501
LKPF-YOLO	0.905	0.779	0.867	**0.819**	**0.586**	0.901	0.682	0.743	0.813	0.717	0.690	0.638	0.617	0.841	0.809	0.797	**0.872**	0.650	0.533
DE-YOLO	0.871	0.759	0.815	0.809	0.538	0.878	**0.734**	0.647	**0.843**	0.702	0.652	0.572	0.574	0.793	0.691	0.720	0.840	0.571	0.432
MCIA-YOLO	0.921	0.775	0.822	0.798	0.541	0.884	0.701	0.702	0.752	**0.724**	0.685	**0.659**	0.614	0.844	**0.812**	0.785	0.802	0.599	0.457
CBGS-YOLO	0.881	**0.802**	0.824	0.807	0.564	0.893	0.696	0.668	0.778	0.669	0.662	0.588	0.576	0.798	0.718	**0.837**	0.845	0.559	0.443
Ours	**0.943**	0.798	**0.921**	0.813	0.570	**0.931**	0.725	**0.806**	0.801	0.717	**0.726**	0.658	**0.687**	**0.876**	0.806	0.828	0.824	**0.702**	**0.598**

As shown in Figure 11, the visualization results of the proposed GSS-YOLO algorithm on the DIOR remote sensing dataset are presented. It can be intuitively observed from the figure that the model accurately identifies and localizes various geographical objects with significant scale variations. Particularly for categories with distinct geometric features, such as stadium, wind mill, and tennis court, the generated bounding boxes closely align with the object boundaries, demonstrating high localization precision. Leveraging its enhanced feature fusion mechanism, GSS-YOLO effectively distinguishes adjacent entities and captures key semantic information precisely. It maintains a low false alarm rate and stable detection performance even for dim and small targets. This superior visual performance aligns with the quantitative metrics in Table 5, providing strong empirical evidence for the robustness and practical utility of the proposed algorithm in processing large-scale, complex remote sensing scenarios.



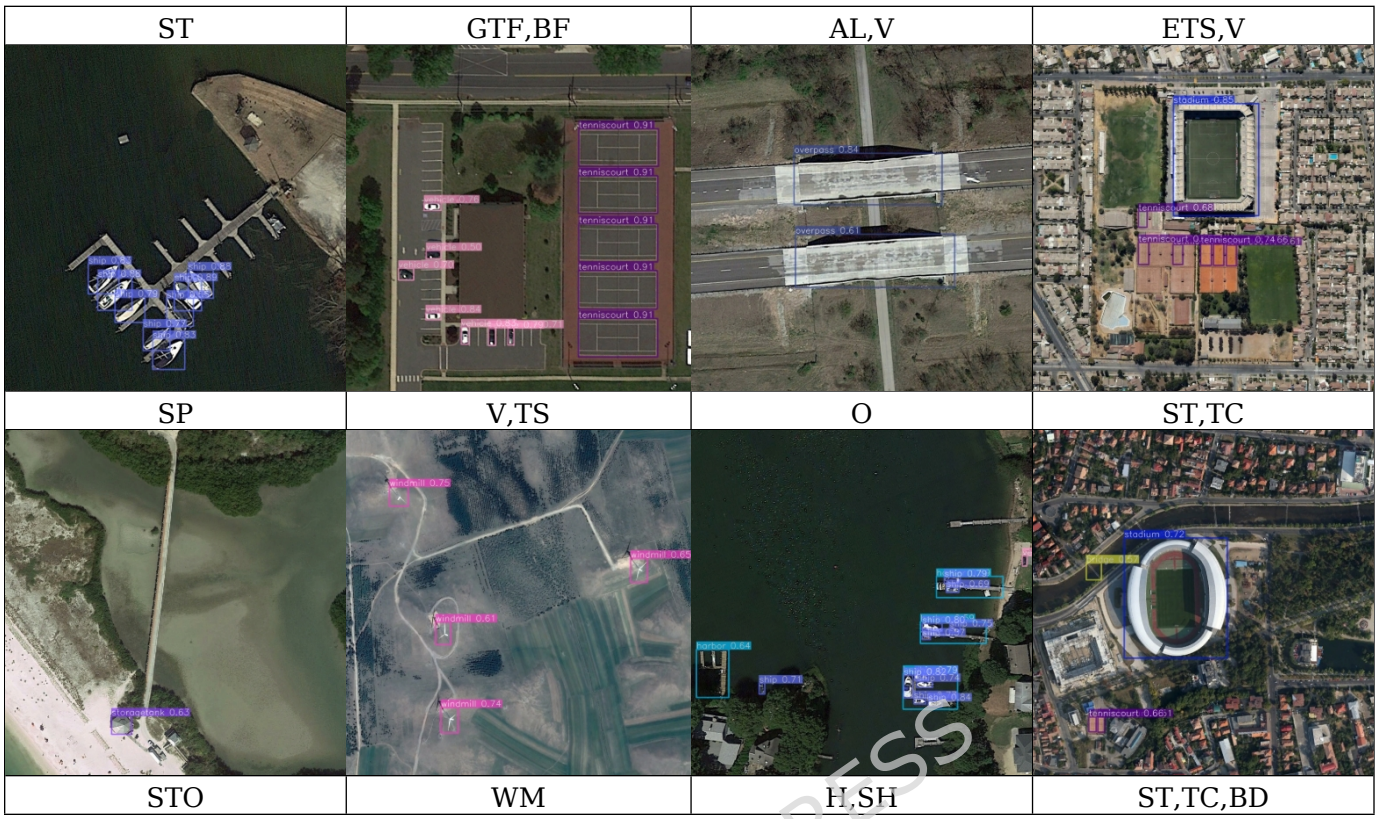


Figure 11. The visualization of GSS-YOLO's object detection performance on the DIOR dataset

Ablation Experiment

To comprehensively evaluate the effectiveness of each improvement strategy in object detection, we conducted a series of ablation experiments. These experiments were categorized into two groups. The first group demonstrated that the performance of object detection could be enhanced through the rational integration of modules, achieved by strategically adding, removing, or fine-tuning key components. The second group involved substituting the modules proposed in this paper with analogous ones from other models, thereby validating the superior performance of our designed modules.

The baseline configuration for the ablation study is the standard YOLOv5-s model without any modifications. To ensure the reliability of the comparative results under small metric margins, all experiments were performed using fixed random seeds and identical training protocols. The incremental improvements observed with the integration of SIA, GCAM, and SPD-Conv demonstrate their specific roles in enhancing feature fusion and localization, particularly for small objects where pixel-level accuracy is critical.

Module Adjustment Experiment

We performed a progressive ablation study on the USOD validation set using the YOLOV5 benchmark model as the baseline. This study systematically evaluated the individual contributions and synergistic effects of each module in the GSS-YOLO model, thereby illustrating the practical impacts of these measures on enhancing object detection performance. The results of the ablation experiments are presented in Table 5.

Table 5. Ablation experiment results of different improved module on the USOD dataset. The symbol "✓" indicates the inclusion of the corresponding module.

GCAM	SIA	SPD-Conv	Precision	Recall	mAP50	mAP50:95
-	-	-	0.821	0.766	0.846	0.297
	✓	✓	0.843	0.783	0.859	0.313
✓	✓	-	0.875	0.814	0.863	0.329
✓	-	✓	0.909	0.833	0.884	0.336
✓	✓	✓	0.938	0.894	0.921	0.343

The ablation experiments demonstrate that selectively introducing two of the GCAM, SIA, and SPD-Conv modules can significantly enhance model performance. Compared to the baseline model without any additional modules, adding SIA and SPD-Conv improves Precision and Recall by 2.2%

and 1.7%, respectively, increases mAP@50 by 1.3%, and boosts mAP@50:95 by 1.6%. When GCAM and SIA are incorporated, Precision and Recall reach 0.875 and 0.814, respectively, while mAP@50 rises to 86.3% and mAP@50:95 climbs to 32.9%. Introducing GCAM and SPD-Conv results in Precision and Recall improvements of 8.8% and 1.7% over the baseline model, with corresponding increases in mAP@50 and mAP@50:95 of 3.8% and 3.9%, respectively. Finally, when all three modules—GCAM, SIA, and SPD-Conv—are integrated, Precision and Recall increase to 0.938 and 0.894, respectively, mAP@50 reaches 92.1%, and mAP@50:95 climbs to 34.3%. These findings indicate that the three modules exhibit strong structural complementarity and can collaboratively enhance the model's ability to detect multi-scale targets. Moreover, they achieve significant improvements in detection accuracy without substantially increasing computational complexity.

To ensure the statistical significance of the results, especially given the incremental nature of some improvements, all ablation experiments were conducted under identical random seeds and hardware environments. Furthermore, we performed three independent training runs for the final GSS-YOLO configuration. Experimental results indicate that the fluctuations across all four evaluation metrics are consistently around 0.5%, confirming that the observed performance gains are consistent and statistically stable rather than the result of stochastic fluctuations.

Comparative Experiment of SIA Module

To compare the impact of different improved C3 structures on model performance, three modules, namely C3_DSA, C3_SDSA, and C3_SIA, were designed for experiments. The results are shown in Table 6. It shows that C3_SIA performs best in terms of overall performance, with a Precision of 0.938, a Recall of 0.894, and mAP@50 and mAP@50:95 increasing to 0.921 and 0.343 respectively. It has 5.04M parameters, demonstrating an excellent balance between detection accuracy and model efficiency. C3_DSA has a slightly better Recall (0.904) and the smallest number of parameters (4.94M), with mAP@50:95 reaching 0.334, but its overall accuracy is slightly lower. While C3_SDSA has the highest Precision (0.941), its mAP performance is poor, especially with mAP@50:95 being only 0.282. In a comprehensive comparison, C3_SIA achieves a better balance between accuracy, recall rate, and detection effect, making it a more advantageous structural design.

Table 6. Performance comparison of different modules combined with the C3 module

Method	Precision	Recall	mAP50	mAP50:95	Parameters
C3_DSA	0.891	0.904	0.853	0.334	4.94
C3_SDSA	0.941	0.882	0.902	0.282	5.98
C3_SIA	**0.938**	**0.894**	**0.921**	**0.343**	**5.04**

In summary, the SIA module not only maintains a compact model structure while effectively reducing the parameter count but also exhibits strong stability and robustness in detection performance. Compared with other alternative modules, SIA achieves a notable improvement in precision and recall while minimizing the decrease in mAP@50:95, thereby achieving an optimal balance between performance and efficiency.

GCAM Comparative Experiment

This study delves deeply into the effectiveness of the GCAM module in enhancing model performance and conducts a series of comparative experiments to comprehensively verify its superiority over other functionally similar modules. We selected four other representative and functionally similar modules as benchmarks, including the NL block and Simplified NL block proposed by Wang et al., the GC block proposed by Cao et al., and the DG module designed by Zhang et al., to establish a comprehensive and fair evaluation framework. We replaced the position of the GCAM module designed in this paper with the comparison modules and compared the results on the USOD dataset, as shown in Table 7.

Table 7. Comparative Analysis of Different Context Acquisition Modules

Method	Precision	Recall	mAP50	mAP50:95	Parameters
NL block	0.825	0.864	0.884	0.299	5.32
Simplified NL block	0.888	0.833	0.913	0.303	5.07
GC block	0.904	0.874	0.908	0.286	4.99
DG block	0.929	0.860	0.922	0.336	6.04
GCAM	**0.938**	**0.894**	**0.921**	**0.343**	**5.04**

The experimental results indicate that not all introduced modules contribute positively to model performance. GCAM (Global Context Attention Module) achieves the best performance in key

metrics such as Precision (0.938), Recall (0.894), and mAP@50:95 (0.343), while maintaining a compact parameter size of 5.04M, thus demonstrating an excellent balance between performance and efficiency. In comparison, GCAM also shows a notable improvement in mAP@50:95 (0.336), but with a slightly higher parameter count of 6.04M. The NL block and Simplified NL block exhibit certain limitations in either accuracy or recall. Although the GC block has a smaller parameter size (4.99M), its mAP@50:95 is the lowest at 0.286. Overall, GCAM significantly enhances the detection accuracy and robustness of the model while ensuring lightweight design.

SPD-Conv Comparative Experiment

To verify the rationality of the SPD-Conv adopted in this paper, two sets of ablation experiments were designed for this convolution. Firstly, to test the rationality of its application in small object detection, we replaced the SPD-Conv in the model with Depthwise Separable Convolution and Group Convolution for comparative experiments. The experimental results are shown in Table 8.

Table 8. Comparison experiments of different convolutional modules

Method	Precision	Recall	mAP50	mAP50:95	Parameters
GroupConv	0.931	0.920	0.893	0.344	5.29
DSConv	0.781	0.812	0.793	0.299	3.98
SPD-Conv	**0.938**	**0.894**	**0.921**	**0.343**	**5.04**

As shown in Table 7, SPD-Conv achieves the best performance in Precision (0.938) and mAP@50 (0.921), with a relatively high mAP@50:95 of 0.343 and a parameter count of 5.04M, demonstrating an excellent balance between detection accuracy and computational efficiency. GroupConv exhibits slightly better Recall (0.920) and a marginally higher mAP@50:95 (0.344), but its overall accuracy is slightly lower, and it has a larger parameter count of 5.29M. In contrast, DSConv, despite having the smallest parameter count (3.98M), shows significantly lower Precision and mAP metrics compared to the other two methods. Overall, SPD-Conv achieves superior detection accuracy while maintaining low complexity, making it a more practical choice for convolutional structure design.

Conclusion

Addressing the persistent challenge of detecting small targets in remote sensing, this study introduces GSS-YOLO, a lightweight architecture designed to optimize feature representation. By embedding a Spatial Information Aggregation (SIA) module within the network and utilizing Spatial Pyramid Dilated Convolution (SPD-Conv), the framework effectively preserves features in low-resolution imagery while intelligently assessing candidate regions. Additionally, the integration of a Global Context Awareness Module (GCAM) refines multi-scale receptive fields, enabling precise identification against complex backgrounds without incurring excessive computational costs.

Empirical validation demonstrates that GSS-YOLO outperforms state-of-the-art detectors, such as YOLOv9 and FFCA-YOLO, across diverse testing scenarios. Experiments on the USOD, VisDrone2019 and DIOR datasets confirm the model's robust stability in challenging conditions, including nighttime, occlusion, and target dispersion in both color and grayscale images. Ablation studies verify the efficiency of the optimized structure, which achieves a Precision of 0.938 and mAP@50 of 0.921 with only 5.04M parameters, striking a superior balance between detection accuracy and model lightness compared to alternative designs.

Despite its strong performance, the practical deployment of GSS-YOLO is subject to specific constraints and boundary conditions. In terms of its scope of applicability, the framework is primarily optimized for edge-computing platforms, such as UAV-based surveillance and real-time Earth observation, where computational resources are limited but high-frequency small targets are prevalent. However, certain limitations persist. Possible failure scenarios include cases of extreme target occlusion (where over 80% of the object is obscured) or severe motion blur induced by high-speed platform instability. In such instances, the reliance on spatial and context-aware features may not suffice to distinguish a faint target from complex background noise. Acknowledging these limitations is essential for ensuring the reliability of the system in unpredictable real-world environments.

These findings establish GSS-YOLO as a highly adaptable solution that effectively reconciles the trade-off between computational efficiency and accuracy in Earth observation tasks. The model’s strong generalization capability marks a significant advancement for practical remote sensing applications. Future research will extend this work by leveraging temporal information for moving target detection in video streams and integrating super-resolution techniques to further enhance the perceptibility of faint targets in low-quality data.

Author Contributions: Conceptualization, methodology, software, validation, Z.W.; formal

analysis, investigation, resources, N.L.; writing—original draft preparation, Z.W.; writing—review and editing, D.W.; supervision, Z.W.; project administration, Z.T..; funding acquisition, D.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Jilin Province Youth Growth Science and Technology Program Project (No. 20220508041RC).

Data Availability Statement: The data used in this study are available upon request from the corresponding author via email.

Conflicts of Interest: The authors declare no conflict of interest.

References:

[1]. Zhang, M., et al., IRSAM: Advancing segment anything model for infrared small target detection. 2024, Springer. p. 233--249.
[2]. Yang, R., et al., KPE-YOLOv5: An improved small target detection algorithm based on YOLOv5. Electronics, 2023. 12(4): p. 817.
[3]. Wang, H., H. Qian and S. Feng, GAN-STD: small target detection based on generative adversarial network. Journal of Real-Time Image Processing, 2024. 21(3): p. 65.
[4]. Tong, X., et al., MSAFFNet: A multiscale label-supervised attention feature fusion network for infrared small target detection. IEEE Transactions on Geoscience and Remote Sensing, 2023. 61: p. 1--16.
[5]. Dai, Y., et al., One-stage cascade refinement networks for infrared small target detection. IEEE transactions on geoscience and remote sensing, 2023. 61: p. 1--17.
[6]. Dai, Y., et al., Attentional local contrast networks for infrared small target detection. IEEE transactions on geoscience and remote sensing, 2021. 59(11): p. 9813--9824.
[7]. Jing, R., et al., Sunflower-YOLO: Detection of sunflower capitula in UAV remote sensing images. European Journal of Agronomy, 2024. 160: p. 127332.
[8]. Wu, Z., et al., Cbgs-yolo: A lightweight network for detecting small targets in remote sensing images based on a double attention mechanism. Remote Sensing, 2024. 17(1): p. 109.
[9]. Hui, Y., J. Wang and B. Li, DSAA-YOLO: UAV remote sensing small target recognition algorithm for YOLOV7 based on dense residual super-resolution and anchor frame adaptive regression strategy. Journal of King Saud University-Computer and Information Sciences, 2024. 36(1): p. 101863.
[10]. Zhang, W., et al., LS-YOLO: A novel model for detecting multiscale landslides with remote sensing images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2024. 17: p. 4952--4965.
[11]. Hu, J., et al., CM-YOLO: Typical object detection method in remote sensing cloud and mist scene images. Remote Sensing, 2025. 17(1): p. 125.
[12]. Hui, Y., et al., SEB-YOLO: An improved YOLOv5 model for remote sensing small target detection. Sensors, 2024. 24(7): p. 2193.
[13]. Bian, D., et al., A refined methodology for small object detection: Multi-scale feature extraction and cross-stage feature fusion network. Digital Signal Processing, 2025: p. 105297.
[14]. Xia, Y., et al., Behavior detection Algorithm of Caged White-feather broiler based on multi-scale detail feature fusion and object relation inference. 2023, IEEE. p. 1002--1006.
[15]. Zhang, D., et al., Unsupervised Pre-training with Language-Vision Prompts for Low-Data Instance Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025.
[16]. Li, H., et al., VDG: vision-only dynamic gaussian for driving simulation. IEEE Robotics and Automation Letters, 2025.
[17]. Tang, B., et al., R2PLoc: A region-to-point UAV visual geo-localization framework leveraging hierarchical semantic representation. IEEE Transactions on Geoscience and Remote Sensing, 2025.
[18]. Liang, L., et al., DBMLLA: Double-branch Mamba-like linear attention network for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing, 2025.
[19]. Liang, L., et al., LKMA: Learnable Kernel and Mamba with Spatial-Spectral Attention Fusion for Hyperspectral Image Classification. IEEE Transactions on Geoscience and Remote Sensing, 2025.
[20]. Zhang, Z., et al., Multireceptive field: An adaptive path aggregation graph neural framework for hyperspectral image classification. Expert Systems with Applications, 2023. 217: p. 119508.
[21]. Yao, D., et al., Deep hybrid: multi-graph neural network collaboration for hyperspectral image classification. Defence Technology, 2023. 23: p. 164--176.
[22]. Zhang, T., et al., CCSFuse: Collaborative Compensation and Selective Fusion for UAV-Based RGB-IR Object Detection. IEEE Transactions on Geoscience and Remote Sensing, 2025. 64: p. 1--14.
[23]. Zhuo, Z., et al., TAF-YOLO: A Small-Object Detection Network for UAV Aerial Imagery via

Visible and Infrared Adaptive Fusion. Remote Sensing, 2025. 17(24): p. 3936.
[24]. Weng, T. and X. Niu, LMDENet: A Lightweight RGB-IR Object Detection Network for Low-Light Remote Sensing Images. Sensors, 2026. 26(4): p. 1130.
[25]. He, M., et al., Misaligned RGB-infrared object detection via adaptive dual-discrepancy calibration. Remote Sensing, 2023. 15(19): p. 4887.
[26]. Zhang, Y., et al., FFCA-YOLO for Small Object Detection in Remote Sensing Images. IEEE Transactions on Geoscience and Remote Sensing, 2024. 62({}): p. 1-15.
[27]. Du, D., et al., VisDrone-DET2019: The vision meets drone object detection in image challenge results. 2019. p. 0--0.
[28]. Li, K., et al., Object detection in optical remote sensing images: A survey and a new benchmark. ISPRS journal of photogrammetry and remote sensing, 2020. 159: p. 296--307.
[29]. Chen, J., et al., LKPF-YOLO: A Small Target Ship Detection Method for Marine Wide-Area Remote Sensing Images. IEEE Transactions on Aerospace and Electronic Systems, 2024.
[30]. Li, Y. and Y. Zhang, Robust infrared small target detection using local steering kernel reconstruction. Pattern Recognition, 2018. 77: p. 113--125.
[31]. Wang, J., et al., Remote sensing small object detection based on multi-contextual information aggregation. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2025.
[32]. Wang, C., I. Yeh and H. Mark Liao, Yolov9: Learning what you want to learn using programmable gradient information. 2024, Springer. p. 1--21.
[33]. Wang, A., et al., Yolov10: Real-time end-to-end object detection. Advances in Neural Information Processing Systems, 2024. 37: p. 107984--108011.
[34]. Khanam, R. and M. Hussain, Yolov11: An overview of the key architectural enhancements. arXiv preprint arXiv:2410.17725, 2024.