remote sensing

MDPI

Article

SCM-YOLO for Lightweight Small Object Detection in Remote Sensing Images

Hao Qiang [1,2], Wei Hao [1,2,*], Meilin Xie [1,2], Qiang Tang [1,2], Heng Shi [1,2], Yixin Zhao [1,2] and Xiaoteng Han [1,2]

[1] Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, China; qianghao23@mails.ucas.ac.cn (H.Q.); xiemeilin@opt.ac.cn (M.X.); tangqiang22@mails.ucas.ac.cn (Q.T.); shiheng@opt.ac.cn (H.S.); zhaoyixin20@mails.ucas.ac.cn (Y.Z.); hanxiaoteng23@mails.ucas.ac.cn (X.H.)
[2] University of Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: haowei@opt.ac.cn

Abstract: Currently, small object detection in complex remote sensing environments faces significant challenges. The detectors designed for this scenario have limitations, such as insufficient extraction of spatial local information, inflexible feature fusion, and limited global feature acquisition capability. In addition, there is a need to balance performance and complexity when improving the model. To address these issues, this paper proposes an efficient and lightweight SCM-YOLO detector improved from YOLOv5 with spatial local information enhancement, multi-scale feature adaptive fusion, and global sensing capabilities. The SCM-YOLO detector consists of three innovative and lightweight modules: the Space Interleaving in Depth (SPID) module, the Cross Block and Channel Reweight Concat (CBCC) module, and the Mixed Local Channel Attention Global Integration (MAGI) module. These three modules effectively improve the performance of the detector from three aspects: feature extraction, feature fusion, and feature perception. The ability of SCM-YOLO to detect small objects in complex remote sensing environments has been significantly improved while maintaining its lightweight characteristics. The effectiveness and lightweight characteristics of SCM-YOLO are verified through comparison experiments with AI-TOD and SIMD public remote sensing small object detection datasets. In addition, we validate the effectiveness of the three modules, SPID, CBCC, and MAGI, through ablation experiments. The comparison experiments on the AI-TOD dataset show that the mAP50 and mAP50-95 metrics of SCM-YOLO reach 64.053% and 27.283%, respectively, which are significantly better than other models with the same parameter size.

Keywords: small object detection; remote sensing image; spatial local information; feature fusion; attention mechanism

check for updates

Academic Editors: Zhenming Peng and Zhengzhou Li

Received: 2 December 2024
Revised: 6 January 2025
Accepted: 7 January 2025
Published: 12 January 2025

Citation: Qiang, H.; Hao, W.; Xie, M.; Tang, Q.; Shi, H.; Zhao, Y.; Han, X. SCM-YOLO for Lightweight Small Object Detection in Remote Sensing Images. *Remote Sens.* **2025**, *17*, 249. https://doi.org/10.3390/rs17020249

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

With the rapid development of remote sensing technology, the applications of optical remote sensing imagery in emergency rescue, traffic monitoring, building inspection, and military have grown significantly [1–4]. Furthermore, object detection based on optical remote sensing images has also flourished, including the detection of automobiles and aircraft, among other applications [5,6]. Optical remote sensing images typically possess a vast field of view, which is capable of acquiring a substantial amount of information. However, this also means that the objects of interest are small objects and are lost in a complex remote sensing background [7]. The variety of object orientations, scales, and characteristics also creates difficulties in object detection in optical remote sensing imagery [8,9].

Nowadays, with the rapid development of deep learning, optical remote sensing object detection algorithms based on deep learning have also been widely developed and applied. Due to the deep mining ability of deep learning neural networks for information and the strong generalization ability for different scenes, object detection algorithms based on deep learning have achieved quite good results in the task of object detection in complex remote sensing backgrounds. In contrast, the traditional object detection algorithms usually rely on handcrafted feature extraction techniques, which cannot adequately capture the high-level semantic information in images. Furthermore, the traditional object detection algorithms often need to compute a large number of features when dealing with high-resolution images or large images, which directly leads to a significant increase in computational overhead. Therefore, the traditional object detection algorithms have inferior performance in terms of accuracy and speed. Based on the above comparison, the research in this paper focuses on the improvement of deep learning small object detection algorithms in the context of optical remote sensing. The current supervised deep learning object detection algorithms can be mainly categorized into one-stage and two-stage. The vast majority of current two-stage object detection algorithms are based on the RCNN family [10–12]. Two-stage object detection algorithms have advantages in accuracy and efficiency but tend to be more complex and consume more resources. Therefore, in order to balance detection speed and accuracy, researchers have proposed one-stage detection algorithms, such as the YOLO series [13–22], the MobileNet series, and their improvements [23–26]. In the development of one-stage object detection algorithms, researchers have proposed a variety of improvement methods to further enhance the detection performance. The attention mechanism is a particularly noteworthy example of such methods, including CBAM [27], SEnet [28], ECA [29], MLCA [30], and so on. The attention mechanism is primarily employed to assist in feature extraction by leveraging the local and global information of the image, thereby effectively improving the detection performance of objects. The effectiveness of this feature extraction method has been demonstrated by related studies [31–34]. The above improvements are very important for the improvement of one-stage object detection algorithms; however, there are still the following problems to be improved.

(1) One-stage detection algorithms, such as the YOLO family, support feature extraction by incorporating a hierarchical structure of downsampling and upsampling [35,36]. It is typically implemented through operations such as convolution and pooling [37]. The downsampling operation loses certain spatial local information, which is insufficient for the preservation and exploitation of spatial local information.
(2) In YOLO series, the Concat operation is usually used to realize feature fusion, but the Concat operation does not make flexible and efficient use of input features for feature fusion.
(3) The introduction of the attention mechanism in the model can effectively improve the feature perception ability of the detector, thus improving the detection performance. However, the introduction of the attention mechanism will also lead to an additional increase in the number of parameters.

To address these issues, we design a high-precision small object detector SCM-YOLO, which can efficiently balance detection speed, detection accuracy, and resource consumption for the complex background of optical remote sensing images, so that it can efficiently perform the remote sensing small object detection task on multiple platforms. In order to further preserve and utilize the spatial local information, we propose the Space Interleaving in Depth (SPID) module based on SPD-conv [38], which can enhance the preservation and utilization of spatial local information by performing Interleaving Concat on spatial local information during the downsampling operation.

In the process of multi-scale feature fusion, we are inspired by the multi-scale feature fusion of BiFPN [39]. We propose the Cross Block and Channel Reweight Concat (CBCC) module. The CBCC module improves the model convergence speed while performing flexible and efficient feature fusion through a two-level weighting structure. After further optimization of feature enhancement and feature fusion, inspired by GCNet [33] and SCP [40], we propose the Mixed local channel Attention Global Integration (MAGI) module based on MLCA [30], which performs global large-view feature perception prior to detection and makes full use of contextual features to enhance feature expression and further improve the detector's ability to discriminate between objects and background. The residual structure is introduced by MLCA to ensure the robustness of the module output [41]. It is worth noting that all three modules we propose are lightweight modules, which can enhance the detection capability without significantly increasing the model complexity.

The contributions of this paper are summarized as follows:

(1) A small object detector, SCM-YOLO, for optical remote sensing of complex backgrounds is designed, which shows advanced performance for small objects in remote sensing complex backgrounds while balancing detection speed and resource consumption.
(2) We propose the three lightweight generalized modules, SPID, CBCC, and MAGI, which improve comprehensive performance of SCM-YOLO from three aspects: local spatial information utilization, feature fusion, and global perception.
(3) Based on the YOLOv5 model, three lightweight generalized modules, SPID, CBCC, and MAGI, are added to validate the effectiveness and feasibility of the new detector, SCM-YOLO, on the remotely sensed small object datasets AI-TOD [42] and SIMD [43], as shown in Figure 1.

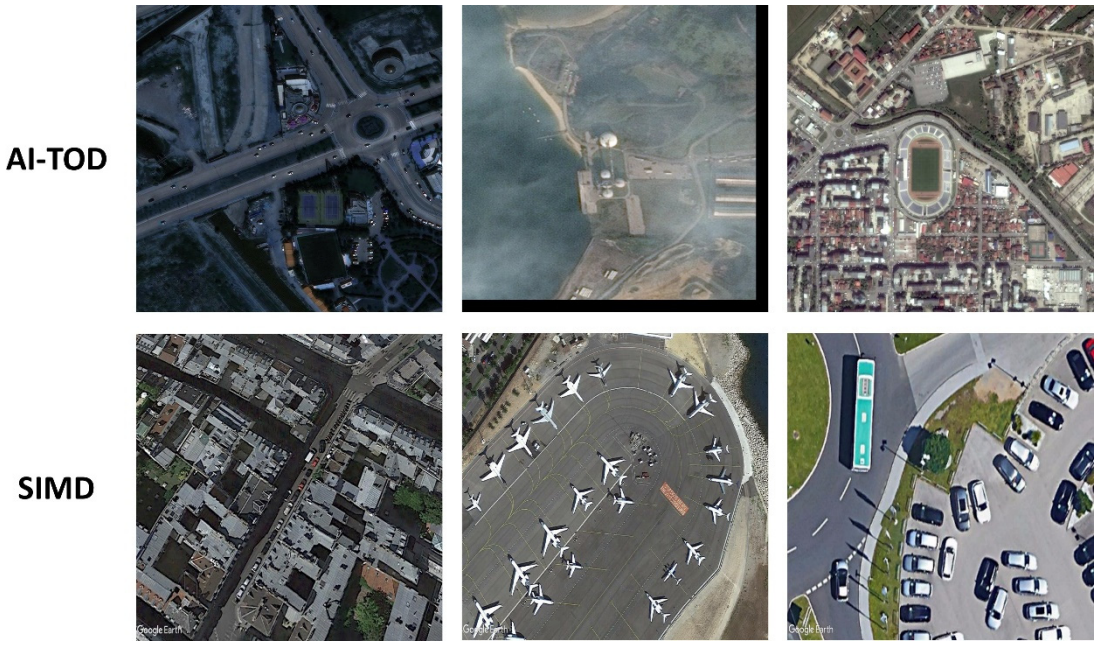


Figure 1. Examples of AI-TOD and SIMD.

The remainder of this paper is organized as follows. Section 2 summarizes the related work in terms of YOLO family, attention mechanism, and improvements of small object detection. Section 3 illustrates the overall structure of SCM-YOLO and provides the principles of the three lightweight generalized modules, SPID, CBCC, and MAGI, proposed in this study. In Section 4, we introduce the experimental details, including datasets, ablation experiments, comparison experiments, and detection effects, through which the feasibility of the proposed method is verified. Section 5 summarizes the whole paper and

looks forward to our subsequent future research directions for small object detection in the context of remote sensing.

2. Related Work

2.1. The YOLO Series Models

The current supervised deep learning object detection algorithms can be classified into two main categories: one-stage and two-stage. The one-stage detection methods are primarily represented by the YOLO series [13–22], whereas the two-stage object detection algorithms are predominantly exemplified by the RCNN series [10–12]. SCM-YOLO proposed in this paper represents an improvement on the YOLO algorithm. The development of the YOLO series algorithm is described below.

The initial iteration of the YOLO series, the YOLOv1 network [13], was developed by Redmon et al. in 2016. It employs a one-stage detection approach to markedly enhance detection speed while maintaining high accuracy. YOLOv1 achieves end-to-end one-stage detection by treating object detection as a regression problem, a notable innovation that significantly enhances the detector's utility. In 2017, Redmon et al. proposed YOLOv2 [14], which effectively improved the detection accuracy and multi-scale feature processing capability by integrating batch normalization [44] and featuring pyramid FPNs [35]. In 2018, Redmon's team introduced the use of Darknet-53 as the backbone network of the YOLOv3 model [15], which represented a further improvement in the accuracy of the detection process. In 2020, Bochkovskiy et al. proposed the YOLOv4 model [16], which incorporated the Mish activation function [45], the SPP block [46], the CSPNet [47], and the PANet [36]. Additionally, YOLOv4 employs data augmentation techniques to enhance the model's robustness and generalization performance. Despite the resulting reduction in inference speed, the enhanced detection accuracy makes YOLOv4 one of the most advanced detection models of the year.

Following the release of YOLOv4, the Ultralytics team, led by Glenn Jocher, proposed YOLOv5 [17], which represented a significant advancement in object detection. The model has been widely adopted for its balance of detection speed and accuracy in a single step, improving ease of use and flexibility. In 2022, Li et al. proposed YOLOv6 [18], which employed an enhanced CSPNet structure, incorporated more sophisticated training methodologies and data enhancement techniques, and further optimized the model's inference speed while maintaining a high level of detection accuracy. Wang et al. proposed YOLOv7 [19], which supported a broader range of tasks and offered an improved detection performance while maintaining the speed advantage. YOLOv8 [20] introduces more sophisticated architectural designs, such as an enhanced convolution module and an attention mechanism, with the objective of improving the feature extraction capability. YOLOv9 [21] integrates a transformer module with the intention of enhancing the model's capacity for long-range dependent feature capture, particularly in complex scenes. YOLOv10 [22] incorporates multimodal fusion with the goal of improving the robustness and generalization ability of the model in complex scenes. Furthermore, a novel attention mechanism was introduced to enhance the model's performance in YOLOv10. YOLOv10 [22] eliminates the necessity for NMS training, thereby achieving advanced performance while reducing computational overhead [48]. Overall, the multiple versions of the YOLO series after YOLOv5 have increased the consumption of computational resources to varying degrees, but there is no significant improvement in the ability to detect small objects for remote sensing.

While significant progress has been made with the YOLO series, the flexibility and ease-of-use features of YOLOv5 make it an excellent choice for improvement as a very mature and widely used object detection framework. In addition, YOLOv5 strikes an excellent balance between accuracy and speed. After experiments, YOLOv5 also has better

performance in remote sensing small object detection. Therefore, YOLOv5 served as the benchmark model for the experiments presented in this paper.

2.2. Attention Mechanism

In the field of object detection, the attention mechanism has emerged as a pivotal method for enhancing the performance of detection models in recent years. The purpose of the attention mechanism is to improve the accuracy and efficiency of the detection process by adjusting the weights assigned to the feature maps to direct the model's attention to the most relevant regions. In addition, numerous attention mechanisms have been proposed in recent years, such as CBAM [27], SENet [28], ECA [29], and MLCA [30]. SENet places a greater emphasis on the region of interest by evaluating the distribution of features on each channel and weighting each channel accordingly. CBAM combines the Spatial Attention Module (SAM) with the Channel Attention Module (CAM) to enhance object detection. This combination enables the model to focus on important channels while also locating key regions spatially. In 2020, Wang et al. proposed ECA, which is based on SENet. The main feature of ECA is that it weights the channels using 1D convolution, which avoids the parameter and computational overheads associated with the fully connected layer in SENet. This makes it well suited for use in resource-constrained environments. The number of parameters is reduced, while the performance of the model is still effectively improved. MLCA is a lightweight attentional mechanism that extracts and focuses key information more comprehensively through multi-level feature fusion with adaptive weighting.

As a crucial technique in the domain of deep learning, the attention mechanism offers several advantages, including enhanced model performance, flexibility, and the ability to handle long-range dependencies. However, there is a need to fully leverage the features introduced by this mechanism for more effective object detection. The MAGI module proposed in this paper represents a more comprehensive utilization of hybrid attention, thereby further enhancing the accuracy of the detection process.

2.3. Improved YOLO for Small Object Detection

TPH-YOLO [49] integrates the Transformer module, which enriches the global context information and enhances the feature representation. FE-YOLO [50], proposed by Wang et al., employs deformable convolution for feature fusion to mitigate the impact of semantic gaps and further enhance detection accuracy. A-YOLO [51] incorporates the SE and CBAM attention mechanisms to augment the detection capability and also employs the Focal–Eiou loss function [52], which markedly enhances the detection accuracy and robustness of the algorithm. YOLO-TLA [53] incorporates an additional small object detection layer into the neck network pyramid structure to facilitate the recognition of finer features. Furthermore, a C3CrossCovn module has been incorporated into the backbone network, thereby enhancing the model's compactness.

The SOD-YOLO [54] model is based on the YOLOv8 architecture and incorporates the RFCBAM downsampling module to enhance the efficiency of feature extraction. A novel multi-scale feature fusion structure, BSSI-FPN, was also devised to effectively balance the spatial and semantic information between feature mappings. Wan et al. proposed the YOLO-HR algorithm for high-resolution optical remote sensing object recognition [55]. The algorithm employs multi-layer feature pyramids to leverage the output of multiple detector heads, thereby enhancing the detection performance. Wu et al. proposed YOLO-SE [56] to refine YOLOv8 by integrating a spatial attention mechanism and multi-scale feature fusion, thus improving the accuracy and robustness of small object detection and complex scene recognition in remote sensing images. Xu et al. proposed a method for enhancing the precision and resilience of multi-scale object detection and intricate scene recognition

in remote sensing images [57]. This approach integrates the YOLOv3 and DenseNet architectures to augment the efficacy of multi-scale remote sensing object detection.

The above research work has made various improvements for small object detection based on the YOLO series of algorithms, which has promoted the development of this field. However, there are still challenges, such as insufficient retention and utilization of spatial information, insufficiently flexible feature fusion, insufficient feature detection capability, and increased model complexity, in the process of algorithmic improvement. In this paper, SCM-YOLO is proposed to address the above challenges.

3. Proposed Method

YOLOv5 is more lightweight than other models and can maintain some accuracy in small object detection tasks. Therefore, we adopt YOLOv5 as the base model of SCM-YOLO. In this section, we introduce the SCM-YOLO model architecture and the three modules of SPID, CBCC, and MAGI, respectively. The SPID module enhances the retention and utilization of spatial local information in the model, the CBCC module strengthens the flexibility and high efficiency of feature fusion, and the MAGI module improves the large-field context awareness of the model.

3.1. Overview of SCM-YOLO

SCM-YOLO is a detector for remote sensing of small objects. As shown in Figure 2, the model consists of three parts: backbone, neck, and head. The backbone is mainly structured by the CBS module (the module including Convolution, Batch Normalization, and SiLU) [17] and the CSP module (Cross Stage Partial Network) [16], and the SPID module is introduced to fully preserve spatially localized information when performing downsampling. After the feature map enters the neck part, the upsampling operation is performed, and the flexible feature fusion is performed by the CBCC module. The fusion result is fed into the MAGI module through the three-layer CSP module for global large-view feature sensing. Finally, the output feature map from the MAGI module is fed into the head section for recognition and output.

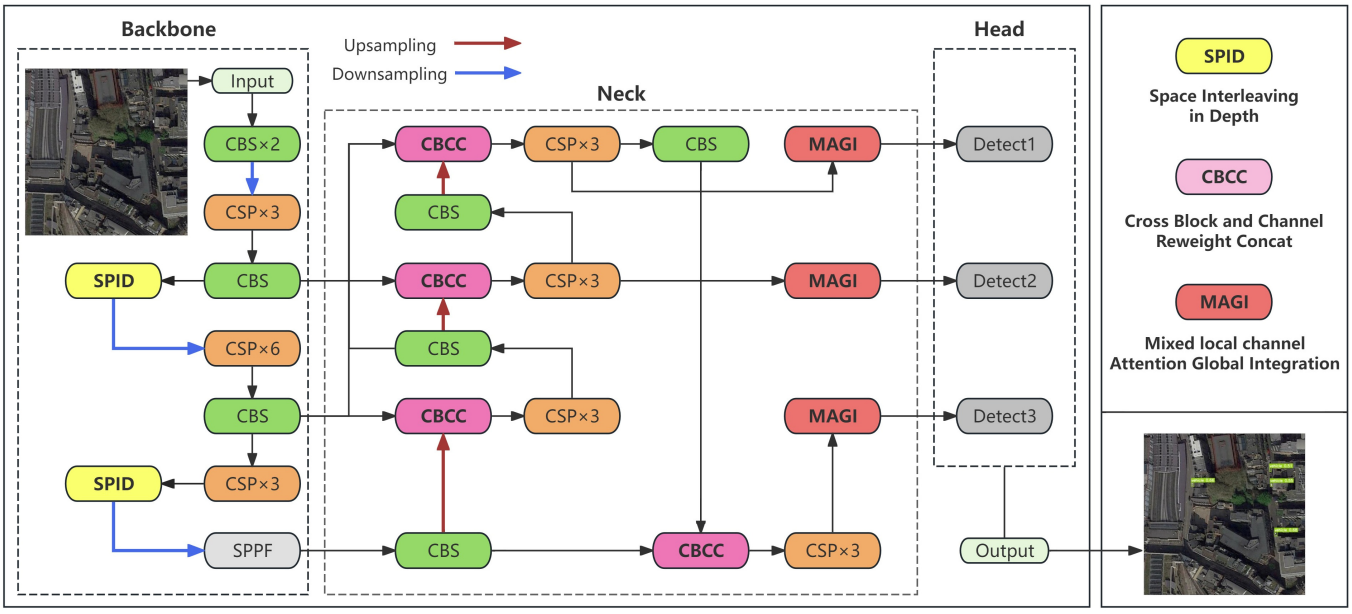


Figure 2. The overall structure of SCM-YOLO.

The CBS module and the CSP module are important modules in the YOLO series [16]. All of them are very important to the feature extraction of the model, and all of them accelerate the convergence of the model.

The workflow of the CBS module and the CSP module is elucidated in Figure 3. The CBS module is comprised of three constituent components: Conv, batch normalization, and SiLU activation function. The Conv, being the core component of feature extraction, is utilized to extract deeper features from the input feature map. It generates a feature map by performing a sliding dot product of the convolution kernel with the input feature map. Batch normalization normalizes the activation values for each input batch (i.e., adjusts the mean to 0 and the variance to 1), followed by retransformation using learnable scaling and offset parameters. The SiLU activation function enhances the nonlinear feature representation of the model through nonlinear transformations. The CSP module splits the input feature map into two parts: one is passed directly through the main path for deep processing, and the other is passed through the shortcut connection. The main path typically consists of structures such as a Conv, a batch normalization, a SiLU activation function, and some bottleneck modules, which are employed to perform complex feature transformations on the input features. The shortcut connection directly connects a portion of the input features to the output of the module, which is fused with the output of the main path to ameliorate the gradient vanishing problem of deep networks.

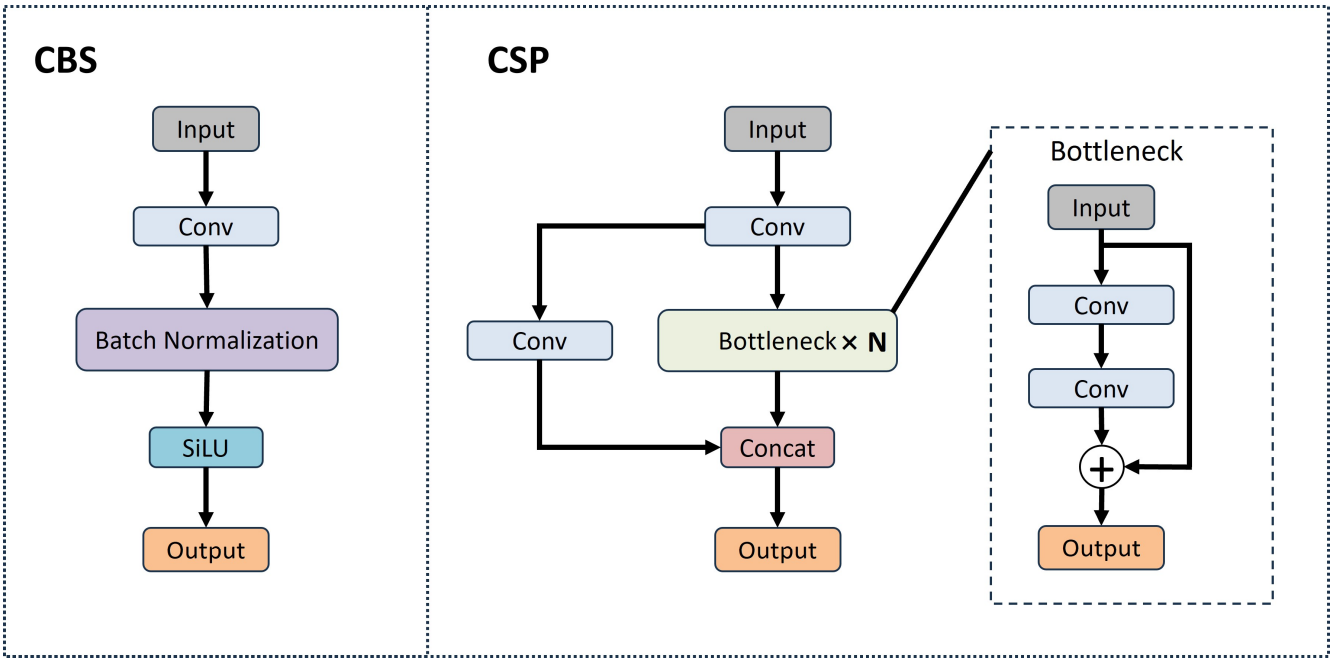


Figure 3. The structure of CBS and CSP.

3.2. Space Interleaving in Depth (SPID) Module

Due to the small scale of small objects in the complex background of remote sensing, the feature information is insufficient. Therefore, it is crucial to fully preserve and utilize the limited information. Inspired by SPD-conv [38], this paper proposes the SPID module, which introduces the Interleaving Concat operation in the spatial downsampling process. Although SPD-conv also preserves the spatial information, it makes spatially adjacent pixels far apart in the channel dimension, resulting in the loss of spatial local information, while the Interleaving Concat operation proposed in this paper allows for pixels in the same local region to remain adjacent during channel dimension reorganization, which better preserves the spatial local information and enhances the utilization of the spatial local

information. In addition, the SPID module also introduces the CBS module, which further enhances the expression of important information in the feature map while reorganizing the number of output channels to the object number of channels. The precise operation of the SPID module is illustrated in Figure 4.

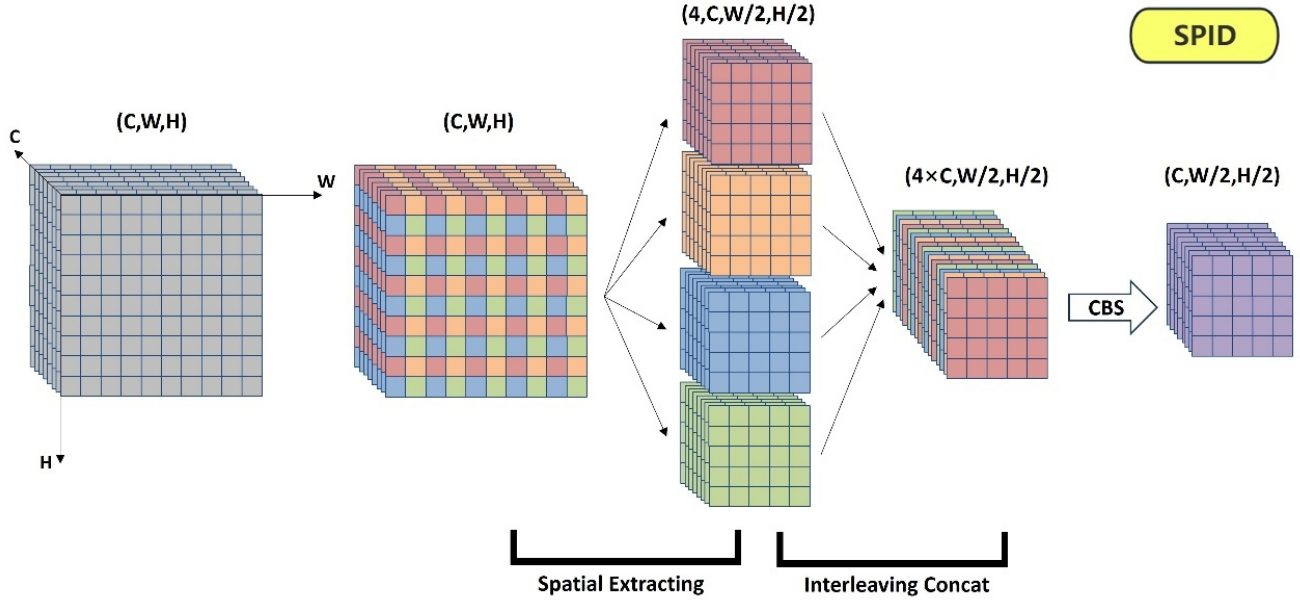


Figure 4. The structure of the SPID module (scale = 2).

For the input feature map of size (C, W, H), pixel points are first extracted in the spatial dimension across channels according to the scale parameter, as follows:

$$\begin{aligned} X_{0,0} &= X[C, 0::scale, 0::scale] \\ X_{scale-1,0} &= X[C, 1::scale, 0::scale] \\ X_{0,scale-1} &= X[C, 0::scale, 1::scale] \\ X_{scale-1,scale-1} &= X[C, 1::scale, 1::scale] \end{aligned} \tag{1}$$

where ::*scale* denotes the information extraction of the input feature map in the spatial dimension in steps of the parameter *scale*. The pixels extracted from the same starting point are reorganized into a new feature subgraph of half the size of the input feature map. The number of feature subgraphs obtained from reorganization is $scale^2$.

Then, in the channel dimension, the sub-feature maps are subjected to the Interleaving Concat operation as shown in Figure 4. After the Interleaving Concat operation, the adjacent pixels in the spatial dimension of the input feature maps are still adjacent in the channel dimension, and thus, the spatial localization information has been sufficiently preserved.

$$F_{Output} = \begin{bmatrix} [X_{0,0,C=0}, X_{scale-1,0,C=0}, X_{scale-1,0,C=0}, X_{scale-1,scale-1,C=0}] \\ [X_{0,0,C=1}, X_{scale-1,0,C=1}, X_{scale-1,0,C=1}, X_{scale-1,scale-1,C=1}] \\ \cdots \\ [X_{0,0,C=n-1}, X_{scale-1,0,C=n-1}, X_{scale-1,0,C=n-1}, X_{scale-1,scale-1,C=n-1}] \end{bmatrix} \tag{2}$$

where C represents the Cth channel of the feature subgraph. Finally, the information is fused and output through the CBS module (Figure 3).

3.3. Cross Block and Channel Reweight Concat (CBCC) Module

The semantic information represented in different levels of feature maps varies. The representation of features associated with small objects can be enhanced in a targeted

manner through the fusion of multi-level feature maps. In this paper, we propose the Cross Block and Channel Reweight Concat (CBCC) module based on BiFPN [40]. By introducing two updatable weighting tensors, W_{block} and $W_{channel}$, the CBCC module can adaptively assign weights by blocks and channels through a two-level weighting structure, which allows for more flexible and efficient feature fusion. Furthermore, the CBCC module can further improve the model convergence speed through the two-layer weighting mechanism. The CBCC module is shown in Figure 5.

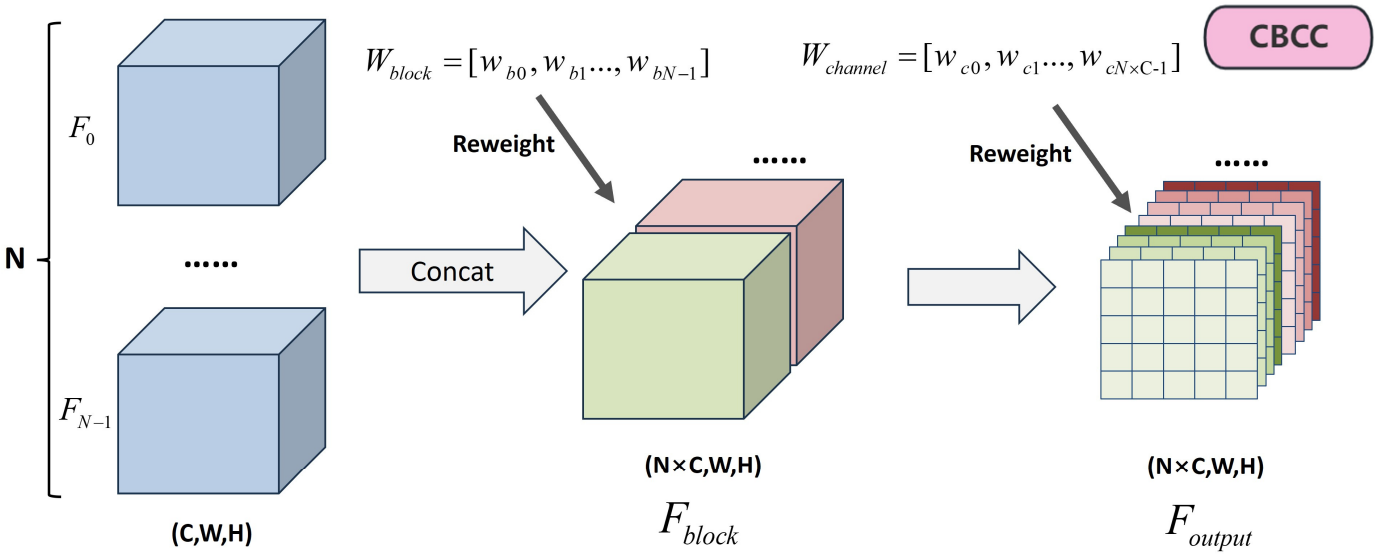


Figure 5. The structure of the CBCC module.

For N feature maps F with input size (C, W, H), the CBCC module first performs the first layer of adaptive block-level weighting on the feature blocks using W_{block} to achieve initial adaptive focusing on key information blocks. Subsequently, the Concat operation is performed to obtain a new feature map F_{block} of size (N × C, W, H). Then, the CBCC module uses $W_{channel}$ to perform a second layer of adaptive weighting on the feature channels to the feature channels and further adaptive focusing on the key channels of the F_{block} to obtain the output feature map F_{output} of size (N × C, W, H).

$$F_{Output} = f_{\text{Reweight}}\left[W_{channel}, f_{\text{Reweight}}[W_{bolck}, (F_0, F_1, \ldots, F_{N-1})]\right] \quad (3)$$

where W_{block} is an updatable block weighting vector and $W_{channel}$ is an updatable channel weighting vector.

By performing two levels of adaptive weighting on blocks and channels, the CBCC module brings more flexible and efficient feature fusion to the model. In addition, the convergence speed of the model is also improved since the channel weighting is performed on top of the block weighting.

3.4. Mixed Local Channel Attention Global Integration (MAGI) Module

After the SPID and CBCC modules, the extracted features have been able to better represent the characteristics of small objects of remote sensing. However, the global context information is not fully exploited. Therefore, the MAGI module is proposed with the combination of MLCA [30] and Atrous Conv [58]. Before the feature map is input to the head, the MAGI module uses contextual information from the global big view by branching to suppress the background, enhance the features, and improve the model's ability to distinguish between the objects and the background, further improving detection accuracy.

The MAGI module first performs Mixed Local Channel Attention (MLCA) [30] operations on the input feature maps (Figure 6) to obtain q through the hybrid attention mechanism of the global visual field, which enables the model to perceive contextual information and better understand the multiple semantic relationships in the images. Then, the input feature map is subjected to Atrous Conv to expand the perceptual field to obtain the large perceptual field feature map k. q and k are multiplied by Hadamard to obtain the qk component, which is used to perform the attention operation on the large perceptual field feature map from the global visual field. Finally, qk is used to extract features by performing query operations on v to further enhance the features and suppress the background. The final output is in the form of residuals [41] to improve the robustness of the model.

$$q = \text{softmax}\left[MLCA\left(F_{input}\right)\right], k = \text{AtrousConv}\left(F_{input}\right), v = \text{Conv}_{1\times1}\left(F_{input}\right) \quad (4)$$

$$qk = \text{Conv}_{1\times1}(q \odot k) \quad (5)$$

$$F_{output} = F_{input} + \text{Sigmoid}(qk \odot v) \quad (6)$$

where *MLCA* is an attention mechanism designed to enhance feature representation in object detection tasks by mixing local channel information. The core idea of *MLCA* is to enhance the representation of important features and suppress irrelevant or redundant information by introducing a hybrid mechanism that considers both local and global information. Overall, the *MLCA* module improves the model's ability to capture useful features while maintaining computational efficiency. Its structure is shown in Figure 7. $\text{Conv}_{1\times1}$ represents the standard convolution operations, with kernel sizes of 1×1. $\odot$ represents the Hadamard product, which is the multiplication of the elements of the corresponding positions of two matrices.

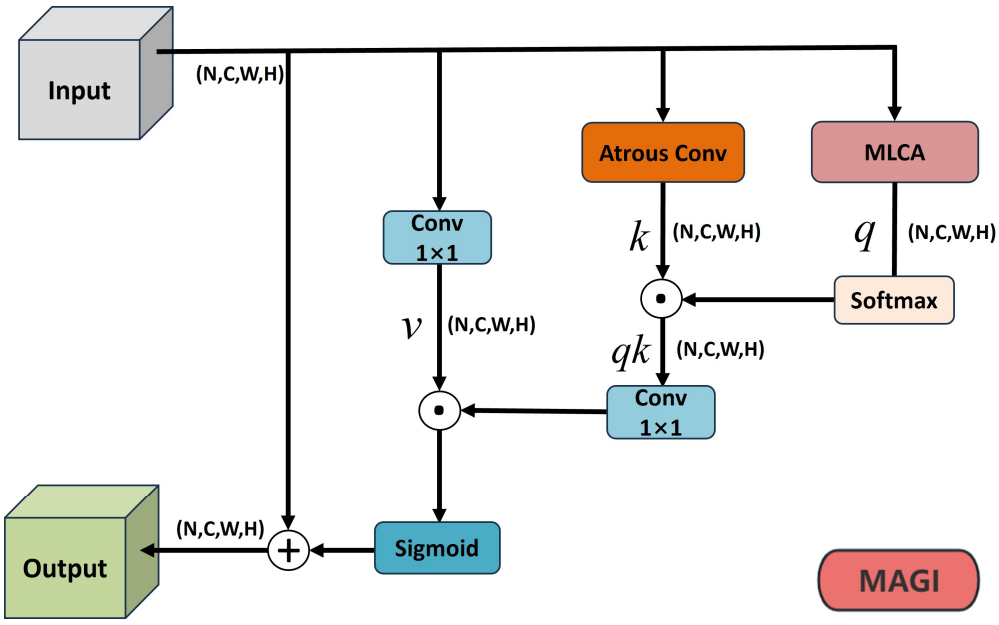


Figure 6. The structure of the MAGI module.

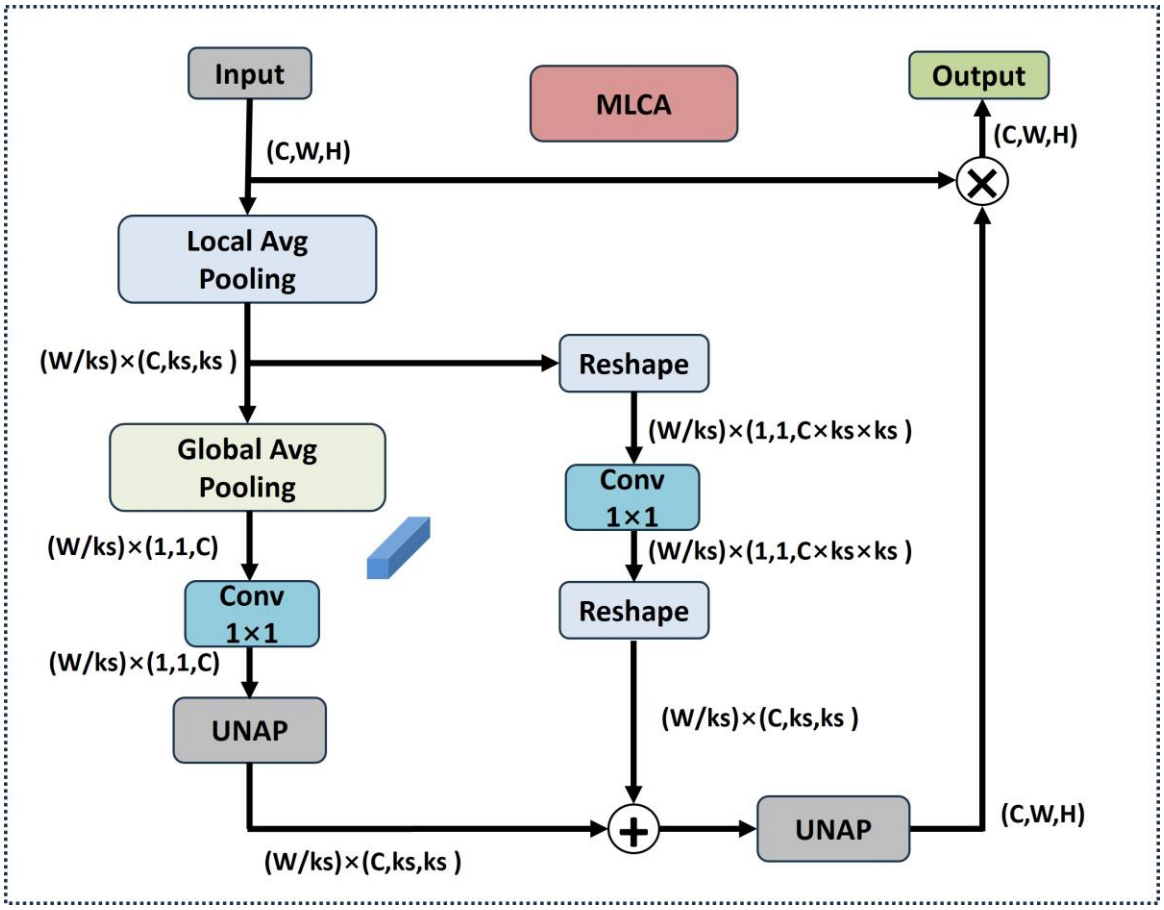


Figure 7. The structure of MLCA.

4. Experiment

4.1. Datasets

4.1.1. AI-TOD Dataset

The AI-TOD [42] (Artificial Intelligence-aided Tiny Object Detection) dataset is a difficult dataset specially designed for small object detections in remote sensing images. The average size of the objects in AI-TOD is about 12.8 pixels, which is much smaller than the general remote sensing object detection dataset. It is characterized by small objects, high resolution, complex background, and high density of objects. Therefore, the AI-TOD dataset puts high demands on the small object detection requirements of the model. AI-TOD contains 28,036 images with a total of 703,843 small objects, including airplanes, ships, vehicles, etc., for a total of eight types of small objects. In this paper, 19,462 images are used as the training set, and the remaining 8574 images are used as the validation set to evaluate the model's performance.

4.1.2. SIMD Dataset

With a larger average object size than the AI-TOD dataset, SIMD [43] is a small- and medium-sized labeled dataset, which can be used for multi-category small object detection tasks. The imagery is derived from publicly available satellite imagery from Google Earth, primarily covering multiple regions in the United States and Europe. In 5000 high-resolution (1024 × 768) satellite images, there are 45,096 labeled objects covering 15 different types of vehicular objects, including cars, trucks, buses, long vehicles, various types of airplanes, and ships. In this paper, 4000 images are used as the

training set, and the remaining 1000 images are used as the validation set to evaluate the model's performance.

4.2. Experimental Evaluation Indicators

In this paper, Precision, Recall, Average Precision (AP), and Mean Average Precision (mAP) are used to evaluate the model's performance. Model size was measured using parameters (Para). The mAPs are categorized into mAP50 (%) and mAP50-95 (%). mAP50 represents the mAP calculated at an IoU threshold of 0.5, which is an evaluation criterion that covers a wider range of IoU thresholds, thus providing a more comprehensive view of the model's detection performance. mAP50-95 represents the mAP calculated at IoU thresholds ranging from 0.5 to 0.95, which can provide a more rigorous evaluation of the model's performance.

$$\begin{cases} \text{Re}call = \frac{TP}{TP+FN} \\ \text{Pr}ecision = \frac{TP}{TP+FP} \end{cases} \tag{7}$$

where TP refers to samples with positive results and positive predictions, FP refers to samples that are actually negative but incorrectly predicted to be positive, TN refers to samples with negative results and positive predictions, and FN refers to samples that are actually positive but incorrectly predicted to be negative. Recall is the proportion of positive samples, TP, correctly identified by the model out of all samples that are actually positive. Precision is the proportion of positive samples, TP, correctly identified by the model as a proportion of all samples predicted by the model to be positive.

4.3. Experimental Results and Analysis

4.3.1. Ablation Test on AI-TOD

To analyze the importance of the three modules proposed in this paper, SPID, CBCC, and MAGI, an ablation test was performed on the AI-TOD dataset. The SPID and CBCC modules were gradually introduced into the YOLOv5 model instead of the original Conv downsampling and Concat, respectively. And the MAGI module was added before the head section to test the effectiveness. The effect of adding or subtracting components on the evaluation metrics is shown in Table 1, where $\surd$ indicates that the module is used and $\times$ indicates that the module is not used.

Table 1. Ablation experiment result on AI-TOD.

SPID	CBCC	MAGI	mAP50 (%)	mAP50-95 (%)	Para (M)
×	×	×	60.680	25.395	5.407
$\surd$	×	×	61.621	25.380	5.699
×	$\surd$	×	62.005	26.383	5.446
×	×	$\surd$	62.216	25.263	7.346
$\surd$	$\surd$	×	63.025	26.522	5.765
$\surd$	×	$\surd$	63.302	26.562	7.665
×	$\surd$	$\surd$	63.523	26.493	7.384
$\surd$	$\surd$	$\surd$	64.053	27.283	7.895

As shown in Table 1, the SPID, CBCC, and MAGI modules can significantly improve the evaluation metrics, and the model's detection performance improvement is obvious. The SPID module brings 0.941% mAP50 metrics improvement with 0.292 M parameter increment, which improves the model's retention and utilization of spatial local information. The CBCC module improves the mAP50 metrics by 1.325% by introducing the least number of parameters (0.046 M). Although the MAGI module is the module with the highest number of parameters, with 1.939 M parameters added to it, it also has the best single-module improvement effect on the mAP50, which increases by 1.536%, and the model's

sensing ability is significantly improved. In Table 1, it can also be seen that the improvement by introducing single modules is more obvious, and with the increase in the number of modules introduced at the same time, the metrics improvement gradually becomes lower, but there is still some improvement, and finally, SCM-YOLO with three modules, namely SPID, CBCC, and MAGI, achieves the highest evaluation metrics and has the best detection effect. However, its model size is also the largest, with parameters reaching 7.895 M.

4.3.2. Comparison Experiments on AI-TOD

In order to validate the performance of SCM-YOLO in detecting small objects in the complex background of remote sensing, we conducted comparison experiments on the AI-TOD dataset. Meanwhile, in order to show the lightweight of SCM-YOLO, we conducted comparison experiments using the lightweight models with the paraSmeter scales of close order of magnitude.

We analyzed the experimental results of the control model in Table 2. YOLOv5s, YOLOv8s, and SPD-YOLO all have mAP50 scores above 58%. YOLOv8s is narrowly ahead of YOLOv5s and SPD-YOLO in the mAP50 metrics to achieve the best results. At the same time, in the more stringent mAP50-95 metrics, YOLOv8s is significantly ahead of the other models, with an advantage of 3–4%, and performs very well. However, the Para metric of YOLOv8s also reaches 11.172 M, which is the model with the highest Para metrics in the comparison experiments.

Table 2. Comparison experiments of detection accuracy on AI-TOD.

Model	mAP50 (%)	mAP50-95 (%)	Recall (%)	Precision (%)	Para (M)
YOLOv5s	58.567	23.352	57.518	71.527	7.023
YOLOv8s	58.760	26.910	55.823	67.176	11.172
YOLOv10s	50.020	22.930	48.850	62.009	8.071
SPD-YOLO	58.090	23.220	59.031	67.392	8.536
SCM-YOLO	64.053	27.283	60.046	73.443	7.895

The experimental results in Table 2 also show that SCM-YOLO has better detection performance than the YOLO models with approximate Para. Compared with YOLOv5s, SCM-YOLO improves the mAP50 metrics by 5.493% and the mAP50-95 metrics by 3.931%, and the number of parameters is reduced by about 24.5%. Compared to YOLOv8s, SCM-YOLO improves the mAP50 metric by 5.293%, the recall metric by 4.223%, and the precision metric by 6.267%, with only half the number of parameters of YOLOv8s. Compared to YOLOv10s and SPD-YOLO, SCM-YOLO also improves the metrics very significantly. Meanwhile, SCM-YOLO also performs very well in the stringent mAP0-95 metric, with a lead of around 4% over YOLOv5s, YOLOv10s, and SPD-YOLO, which is the highest performance. However, YOLOv8s performs equally well on the mAP50-95 metric, only 0.373% below SCM-YOLO.

In summary, SCM-YOLO has the best detection performance for small targets in complex remote sensing contexts, and a good performance improvement can be achieved with a very small increase in Para.

In the comparison experiments of detection speed, Table 3 shows the real-time performance of the different models. It can be seen that SCM-YOLO performs well on the Para metric, indicating its good lightweight characteristics. However, due to the introduction of multiple computational modules, the GFLOPs increase to 42.7. Although the increase in computational volume makes the FPS of SCM-YOLO slightly lower than the other models, SCM-YOLO still achieves 78 FPS on the same experimental platform, which fully meets real-time requirements. This shows that SCM-YOLO has high practicality in ensuring

model performance and detection accuracy while taking into account the computational complexity and speed requirements of practical applications.

Table 3. Comparison experiments of detection speed on AI-TOD.

Model	GFLOPs	Para (M)	FPS (s)
YOLOv5s	15.8	7.023	136
YOLOv8s	28.5	11.172	133
YOLOv10s	24.5	8.071	96
SPD-YOLO	33.1	8.536	138
SCM-YOLO	42.7	7.895	78

Figure 8 shows the training results of SCM-YOLO on the AI-TOD dataset. After 50 epochs of training, SCM-YOLO still maintains a high convergence rate and eventually achieves the best metrics, especially in the mAP50 (Figure 8a) and Recall (Figure 8d) metrics. And for mAP50-95 (Figure 8b) and Precision (Figure 8c), SCM-YOLO still achieves a not outstanding enough but existential lead.

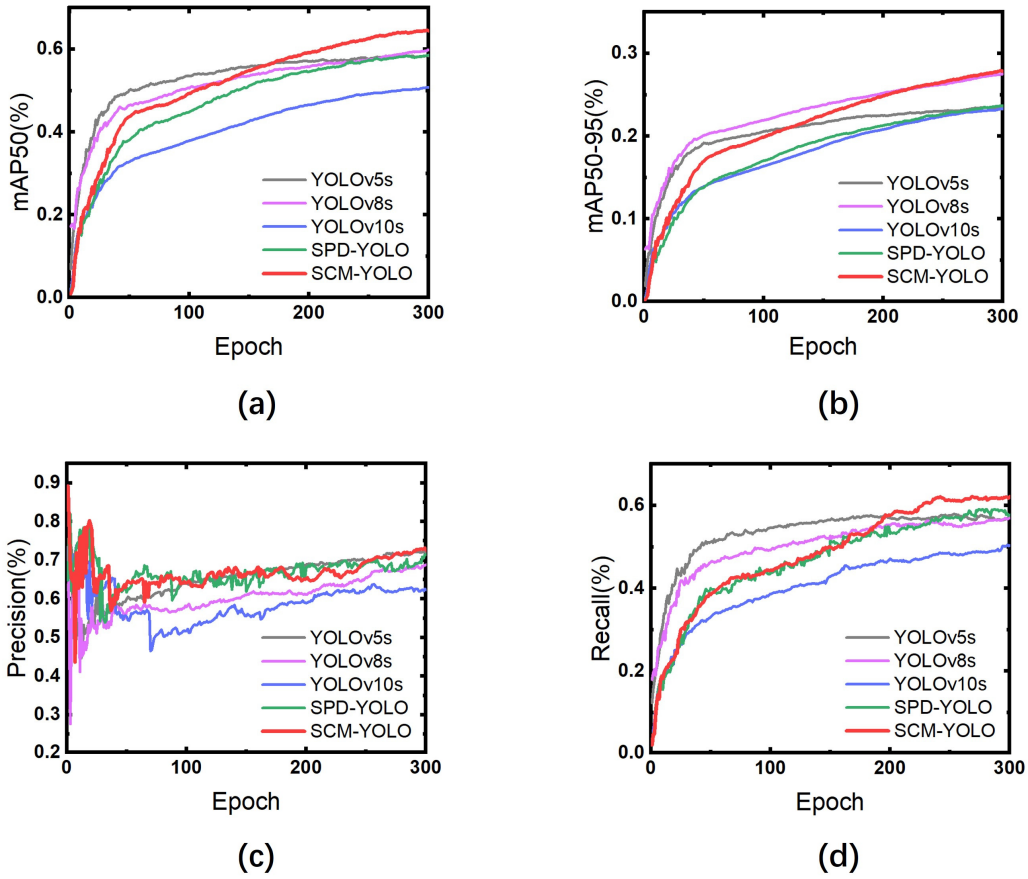


Figure 8. Model training process. (**a**) mAP50; (**b**) mAP50-95; (**c**) Precision; (**d**) Recall.

By further analyzing Figure 8, it can be concluded that the main advantages of SCM-YOLO are reflected in the second half of the training period. After the 150th epoch, SCM-YOLO still maintains a strong convergence ability compared to the other models and finally reaches the optimal metrics. The above shows that SCM-YOLO has a strong performance in deep feature characterization and effectively reduces the impact of the gradient disappearance.

Some detection results are shown in Figure 9. Although the results of SCM-YOLO are not significantly better than those of the other models in terms of intuitive results, the

accuracy of SCM-YOLO is better than that of the other models, which means the detected results are more confident.

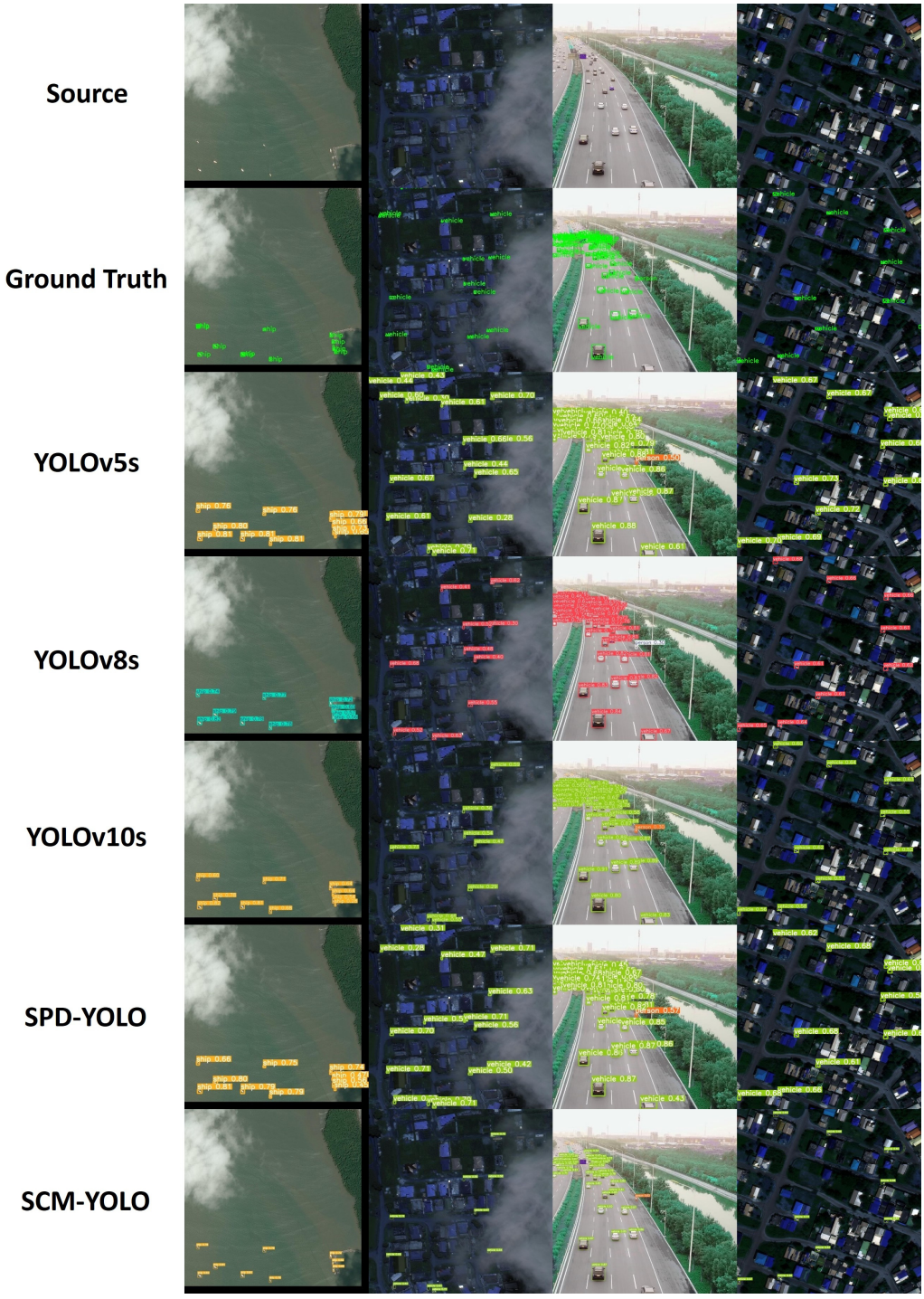


Figure 9. Visual detection results of different models.

4.3.3. Comparison Experiments on SIMD

To further validate the performance of SCM-YOLO proposed in this paper with different small remote sensing object datasets, we also conducted comparative experiments on SIMD datasets.

As shown in Table 4, on the SIMD dataset, SCM-YOLO achieves the best performance with fewer parameters while maintaining the lightness, which further confirms the superiority

of SCM-YOLO in detection performance. Meanwhile, the good performance of SCM-YOLO on different datasets also verifies that it has better versatility.

Table 4. Comparison experiment results on SIMD.

Model	mAP50 (%)	mAP50-95 (%)	Recall (%)	Precision (%)	Para (M)
YOLOv5s	81.450	64.784	78.347	80.754	7.061
YOLOv8s	82.213	64.928	81.670	77.628	11.141
YOLOv10s	81.061	64.859	78.386	76.714	8.077
SPD-YOLO	82.081	64.055	80.317	76.543	8.596
SCM-YOLO	83.158	65.402	81.866	78.735	7.690

4.4. SCM-YOLO in Challenging Scenarios

As demonstrated in Figure 10, four challenging scenarios from the AI-TOD dataset were selected to further illustrate the efficacy of SCM-YOLO. In the first scenario (a), a bright scene with atmospheric interference was analyzed, and it was observed that SCM-YOLO could still detect objects from the atmospheric interference region. In the second scenario (b), a dim scene with atmospheric interference was examined, and it was found that SCM-YOLO could detect objects with high accuracy in both the dim and atmospheric interference areas. (c) In a bright scene with a large number of objects, SCM-YOLO performs well in terms of detection rate. (d) In a dim scene with a large number of objects and some objects in the shadows of buildings, SCM-YOLO still detects objects with high accuracy and a high detection rate in the dark and shadowed areas of the buildings.

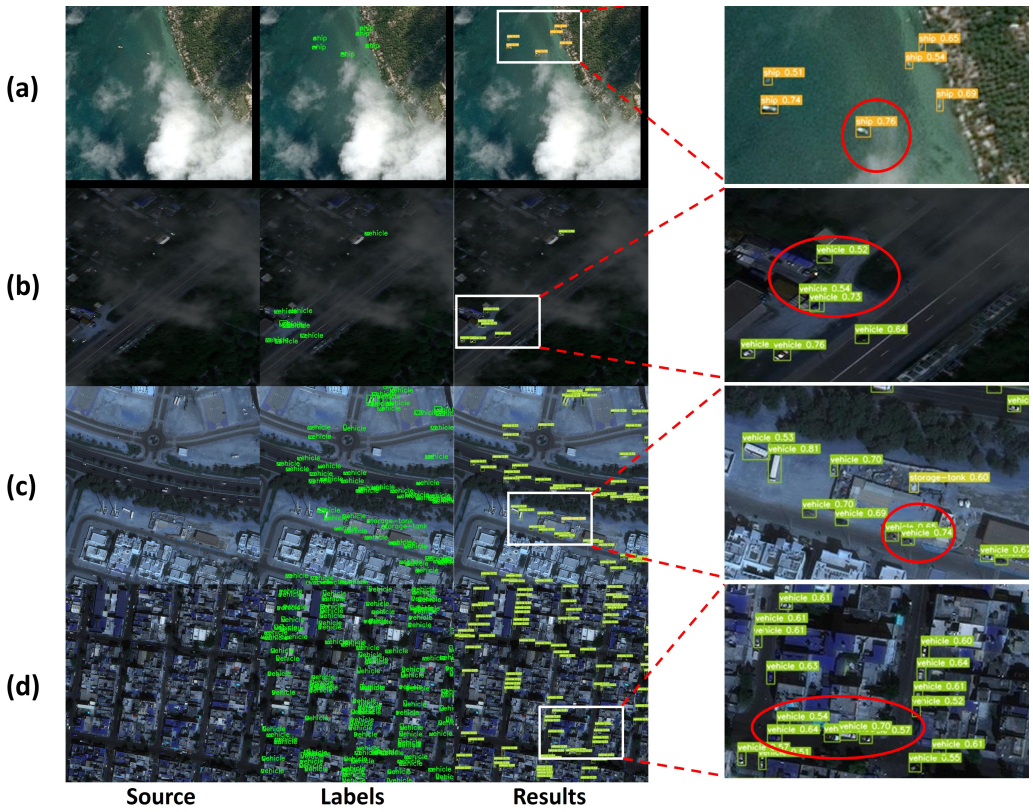


Figure 10. SCM-YOLO visual performance in challenging scenarios from AI-TOD.(**a**) bright scene with atmospheric interference; (**b**) dim scene with atmospheric interference; (**c**) bright scene with a large number of objects; (**d**) dim scene with a large number of objects.

The detection results of SCM-YOLO in challenging scenes of the AI-TOD dataset are shown in Figure 10. In these challenging scenarios, SCM-YOLO still showed good detection performance, performing excellent object detection in image degradation areas.

5. Conclusions

In this paper, a lightweight and efficient SCM-YOLO detector is proposed for optical remote sensing complex background small object detection. And experiments are conducted on the AITOD and SIMD remote sensing small object datasets. The experimental results show that SCM-YOLO has obvious advantages; its mAP50 (%) metrics and mAP50-95 (%) metrics reach 64.053% and 27.283%, respectively; and it achieves the best detection effect with lower parameter counts in comparison experiments with significant improvement. The best results were achieved in the comparison experiments with a lower number of parameters.

Specifically, the SCM-YOLO model proposed in this paper incorporates three lightweight generalized modules, namely SPID, CBCC, and MAGI, which can enhance the model's detection capability for remote sensing of small objects in terms of spatially localized feature enhancement, efficient feature fusion, and contextual feature awareness, respectively. In the feature map downsampling, the SPID module adopts the Interleaving Concat operation to enhance the connection of spatial local information in the channel dimension, which strengthens the retention and utilization of spatially localized features, while the CBCC module realizes the block-by-block and channel-by-channel adaptive weighting of the input feature map through the block-by-block and channel-by-channel weighting structure, which achieves more efficient and flexible multi-scale feature fusion and further improves the model convergence. The MAGI module perceives the global situation by branching and perceives the features in the global field of view of the feature map, which further suppresses the background and enhances the features, and then improves the ability of the model to discriminate between the objects and the background.

Although SCM-YOLO achieves good results in small object detection tasks and is lightweight to enable real-time detection under different platforms, it still has the following limitations that should be improved in subsequent research.

(1) In comparison with the SPID and CBCC modules, the MAGI module proposed in this paper introduces the largest number of parameters and exerts the greatest influence on the model size. In the domain of object detection, the incorporation of an attention mechanism has been shown to enhance the model's detection capability while concomitantly increasing the number of model parameters. A significant number of researchers are exploring more efficient and lightweight designs for the attention mechanism. In our future research, we also plan to design or introduce more lightweight attention mechanisms to improve the model's global perceptual ability. Furthermore, we will continue to explore more efficient ways of utilizing attention mechanisms than the MAGI module.
(2) The method only improves the retention and fusion of spatial local information in the feature enhancement part, and the detection ability of the model can be further improved by introducing more comprehensive feature enhancement means.
(3) The method has only been validated on the AI-TOD and SIMD remote sensing small object datasets. Its performance should be further validated on a wider range of remote sensing small object datasets.

Author Contributions: Conceptualization, W.H. and M.X.; Methodology, H.Q. and Q.T.; Software, H.Q.; Supervision, H.S.; Validation, H.Q. and X.H.; Writing–original draft, H.Q.; Writing–review and editing, Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Xi'an Institute of Optics and Precision Mechanics of CAS, grant number E41I3111.

Data Availability Statement: The datasets presented in this study can be downloaded here: https://github.com/jwwangchn/AI-TOD (AI-TOD, accessed on 2 August 2024). https://github.com/ihians/simd (SIMD, accessed on 22 August 2024).

Acknowledgments: We thank the editors and reviewers for their hard work and valuable advice.

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

References

1. Božić-Štulić, D.; Marušić, Ž.; Gotovac, S. Deep learning approach in aerial imagery for supporting land search and rescue missions. *Int. J. Comput. Vis.* **2019**, *127*, 1256–1278. [CrossRef]
2. Byun, S.; Shin, I.-K.; Moon, J.; Kang, J.; Choi, S.-I. Road traffic monitoring from UAV images using deep learning networks. *Remote Sens.* **2021**, *13*, 4027. [CrossRef]
3. Shimoni, M.; Haelterman, R.; Perneel, C. Hypersectral imaging for military and security applications: Combining myriad processing and sensing techniques. *IEEE Geosci. Remote. Sens. Mag.* **2019**, *7*, 101–117. [CrossRef]
4. Li, J.; Zhuang, Y.; Dong, S.; Gao, P.; Dong, H.; Chen, H.; Chen, L.; Li, L. Hierarchical disentangling network for building extraction from very high resolution optical remote sensing imagery. *Remote Sens.* **2022**, *14*, 1767. [CrossRef]
5. Li, K.; Wan, G.; Cheng, G.; Meng, L.; Han, J. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS J. Photogramm. Remote Sens.* **2020**, *159*, 296–307. [CrossRef]
6. Yu, Y.; Gu, T.; Guan, H.; Li, D.; Jin, S. Vehicle detection from high-resolution remote sensing imagery using convolutional capsule networks. *IEEE Geosci. Remote. Sens. Lett.* **2019**, *16*, 1894–1898. [CrossRef]
7. Shi, T.; Gong, J.; Hu, J.; Zhi, X.; Zhang, W.; Zhang, Y.; Zhang, P.; Bao, G. Feature-enhanced CenterNet for small object detection in remote sensing images. *Remote Sens.* **2022**, *14*, 5488. [CrossRef]
8. Cheng, G.; Han, J. A survey on object detection in optical remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2016**, *117*, 11–28. [CrossRef]
9. Ran, Q.; Wang, Q.; Zhao, B.; Wu, Y.; Pu, S.; Li, Z. Lightweight oriented object detection using multiscale context and enhanced channel attention in remote sensing images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **2021**, *14*, 5786–5795. [CrossRef]
10. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
11. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
12. 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.* **2016**, *39*, 1137–1149. [CrossRef]
13. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
14. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
15. Farhadi, A.; Redmon, J. Yolov3: An incremental improvement. In Proceedings of the Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1–6.
16. Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
17. Jocher, G.; Stoken, A.; Chaurasia, A.; Borovec, J.; Kwon, Y.; Michael, K.; Changyu, L.; Fang, J.; Skalski, P.; Hogan, A.; et al. YOLOv5. Available online: https://github.com/ultralytics/yolov5 (accessed on 27 August 2024).
18. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W. YOLOv6: A single-stage object detection framework for industrial applications. *arXiv* **2022**, arXiv:2209.02976.
19. Wang, C.-Y.; Bochkovskiy, A.; Liao, H.-Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.
20. Joher, G.; Chaurasia, A.; Qiu, J. YOLOv8. Available online: https://github.com/ultralytics/ultralytics (accessed on 27 August 2024).
21. Wang, C.-Y.; Yeh, I.-H.; Liao, H.-Y.M. Yolov9: Learning what you want to learn using programmable gradient information. *arXiv* **2024**, arXiv:2402.13616.

22. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. Yolov10: Real-time end-to-end object detection. *arXiv* **2024**, arXiv:2405.14458.
23. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
24. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4510–4520.
25. Howard, A.; Sandler, M.; Chu, G.; Chen, L.-C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V. Searching for mobilenetv3. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1314–1324.
26. Chu, X.; Zhang, B.; Xu, R. Moga: Searching beyond mobilenetv3. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 4042–4046.
27. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
28. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
29. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 11534–11542.
30. Wan, D.; Lu, R.; Shen, S.; Xu, T.; Lang, X.; Ren, Z. Mixed local channel attention for object detection. *Eng. Appl. Artif. Intell.* **2023**, *123*, 106442. [CrossRef]
31. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7794–7803.
32. Merget, D.; Rock, M.; Rigoll, G. Robust facial landmark detection via a fully-convolutional local-global context network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 781–790.
33. Cao, Y.; Xu, J.; Lin, S.; Wei, F.; Hu, H. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27–28 October 2019; pp. 0–0.
34. Han, W.; Zhang, Z.; Zhang, Y.; Yu, J.; Chiu, C.-C.; Qin, J.; Gulati, A.; Pang, R.; Wu, Y. Contextnet: Improving convolutional neural networks for automatic speech recognition with global context. *arXiv* **2020**, arXiv:2005.03191.
35. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
36. 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, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8759–8768.
37. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
38. Sunkara, R.; Luo, T. No More Strided Convolutions or Pooling: A New CNN Building Block for Low-Resolution Images and Small Objects. In Proceedings of the Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2022, Grenoble, France, 19–23 September 2022; pp. 443–459.
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, Seattle, WA, USA, 14–19 June 2020; pp. 10781–10790.
40. Liu, Y.; Li, H.; Hu, C.; Luo, S.; Luo, Y.; Chen, C.W. Learning to aggregate multi-scale context for instance segmentation in remote sensing images. *arXiv* **2024**, arXiv:2111.11057. [CrossRef]
41. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
42. Wang, J.; Yang, W.; Guo, H.; Zhang, R.; Xia, G.-S. Tiny object detection in aerial images. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 3791–3798.
43. Haroon, M.; Shahzad, M.; Fraz, M.M. Multisized object detection using spaceborne optical imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 3032–3046. [CrossRef]
44. Ioffe, S. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv* **2015**, arXiv:1502.03167.
45. Misra, D. Mish: A self regularized non-monotonic activation function. *arXiv* **2019**, arXiv:1908.08681.
46. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef] [PubMed]

47. Wang, C.-Y.; Liao, H.-Y.M.; Wu, Y.-H.; Chen, P.-Y.; Hsieh, J.-W.; Yeh, I.-H. CSPNet: A new backbone that can enhance learning capability of CNN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 390–391.
48. Hussain, M. YOLOv5, YOLOv8 and YOLOv10: The Go-To Detectors for Real-time Vision. *arXiv* **2024**, arXiv:2407.02988.
49. Zhu, X.; Lyu, S.; Wang, X.; Zhao, Q. TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-captured Scenarios. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021.
50. Wang, M.; Yang, W.; Wang, L.; Chen, D.; Wei, F.; KeZiErBieKe, H.; Liao, Y. FE-YOLOv5: Feature enhancement network based on YOLOv5 for small object detection. *J. Vis. Commun. Image Represent.* **2023**, *90*, 103752. [CrossRef]
51. Li, X.; Wang, S.; Wang, B. A-YOLO: Small object vehicle detection based on improved YOLOv5. In Proceedings of the Third International Conference on Intelligent Traffic Systems and Smart City (ITSSC 2023), Xi'an, China, 10–12 November 2024; pp. 208–216.
52. Zhang, Y.-F.; Ren, W.; Zhang, Z.; Jia, Z.; Wang, L.; Tan, T. Focal and efficient IOU loss for accurate bounding box regression. *Neurocomputing* **2022**, *506*, 146–157. [CrossRef]
53. Ji, C.L.; Yu, T.; Gao, P.; Wang, F.; Yuan, R.Y. Yolo-tla: An Efficient and Lightweight Small Object Detection Model based on YOLOv5. *J. Real-Time Image Process.* **2024**, *21*, 1–16. [CrossRef]
54. Li, Y.; Li, Q.; Pan, J.; Zhou, Y.; Zhu, H.; Wei, H.; Liu, C. SOD-YOLO: Small-Object-Detection Algorithm Based on Improved YOLOv8 for UAV Images. *Remote Sens.* **2024**, *16*, 3057. [CrossRef]
55. Wan, D.; Lu, R.; Wang, S.; Shen, S.; Xu, T.; Lang, X. Yolo-hr: Improved yolov5 for object detection in high-resolution optical remote sensing images. *Remote Sens.* **2023**, *15*, 614. [CrossRef]
56. Wu, T.; Dong, Y. YOLO-SE: Improved YOLOv8 for remote sensing object detection and recognition. *Appl. Sci.* **2023**, *13*, 12977. [CrossRef]
57. Xu, D.; Wu, Y. Improved YOLO-V3 with DenseNet for multi-scale remote sensing object detection. *Sensors* **2020**, *20*, 4276. [CrossRef] [PubMed]
58. Yu, F.; Koltun, V. Multi-scale context aggregation by dilated convolutions. *arXiv* **2015**, arXiv:1511.07122.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.