remote sensing

MDPI

Article

Faster and Lightweight: An Improved YOLOv5 Object Detector for Remote Sensing Images

Jiarui Zhang, Zhihua Chen *, Guoxu Yan, Yi Wang and Bo Hu

National Key Laboratory of Transient Physics, Nanjing University of Science and Technology, Nanjing 210094, China
* Correspondence: chenzh@njust.edu.cn

Abstract: In recent years, the realm of deep learning has witnessed significant advancements, particularly in object detection algorithms. However, the unique challenges posed by remote sensing images, such as complex backgrounds, diverse target sizes, dense target distribution, and overlapping or obscuring targets, demand specialized solutions. Addressing these challenges, we introduce a novel lightweight object detection algorithm based on Yolov5s to enhance detection performance while ensuring rapid processing and broad applicability. Our primary contributions include: firstly, we implemented a new Lightweight Asymmetric Detection Head (LADH-Head), replacing the original detection head in the Yolov5s model. Secondly, we introduce a new C3CA module, incorporating the Coordinate Attention mechanism, strengthening the network's capability to extract precise location information. Thirdly, we proposed a new backbone network, replacing the C3 module in the Yolov5s backbone with a FasterConv module, enhancing the network's feature extraction capabilities. Additionally, we introduced a Content-aware Feature Reassembly (content-aware reassembly of features) (CARAFE) module to reassemble semantic similar feature points effectively, enhancing the network's detection capabilities and reducing the model parameters. Finally, we introduced a novel XIoU loss function, aiming to improve the model's convergence speed and robustness during training. Experimental results on widely used remote sensing image datasets such as DIOR, DOTA, and SIMD demonstrate the effectiveness of our proposed model. Compared to the original Yolov5s algorithm, we achieved a mean average precision (mAP) increase of 3.3%, 6.7%, and 3.2%, respectively. These findings underscore the superior performance of our proposed model in remote sensing image object detection, offering an efficient, lightweight solution for remote sensing applications.

Keywords: object detection; YOLOv5s; remote sensing; FasterConv; LADH-Head; coordinate attention mechanism; XIoU loss function

Citation: Zhang, J.; Chen, Z.; Yan, G.; Wang, Y.; Hu, B. Faster and Lightweight: An Improved YOLOv5 Object Detector for Remote Sensing Images. *Remote Sens.* **2023**, *15*, 4974. https://doi.org/10.3390/rs15204974

Academic Editors: Bo Tang, Xinghua Li, Zongxu Pan, Fan Zhang, Zhongling Huang and Wei Yao

Received: 25 August 2023
Revised: 8 October 2023
Accepted: 12 October 2023
Published: 15 October 2023

1. Introduction

Remote sensing image object detection tasks play a pivotal role in the realm of airborne and satellite remote sensing imagery, representing invaluable applications. Throughout recent decades, remote sensing technology has witnessed remarkable progress, enabling the capture of copious details that inherently reflect the contours, hues, textures, and other distinctive attributes of terrestrial targets [1]. It has emerged as an indispensable avenue for acquiring comprehensive knowledge about the Earth's surface. The primary objective of remote sensing image object detection is to precisely identify and locate objects of interest within the vast expanse of remote sensing images. Presently, this task finds extensive implementation across significant domains, including military reconnaissance [2], urban planning [3], environmental monitoring [4], soil science [5], and maritime vessel surveillance [6]. With the incessant advancement of observational techniques [7], the availability of high-quality remote sensing image datasets, encompassing richer and more intricate information, has unlocked immense developmental potential for the ongoing pursuit of remote sensing image object detection.

In the past decade, deep learning has undergone rapid advancements and progressively found applications in diverse fields, including speech recognition, natural language processing, and computer vision. Computer vision technology has been widely implemented in intelligent security, autonomous driving, remote sensing monitoring, healthcare and pharmaceuticals, agriculture, intelligent transportation, and information security [8–14]. Within computer vision, tasks can be classified into image classification [15], object detection [16], and image segmentation [17]. Notably, object detection, a pivotal branch of computer vision, has made remarkable strides during this period, largely attributed to the availability of extensive object detection datasets. Datasets such as MS COCO [18], PASCAL VOC [19], and Visdrone [20,21] have played a crucial role in facilitating breakthroughs in object detection tasks.

Nevertheless, in the realm of optical remote sensing imagery, current object detection algorithms still encounter numerous formidable challenges. These difficulties arise due to disparities between the acquisition methods used for optical remote sensing imagery and those employed for natural images. Remote sensing imagery relies on sensors such as optical, microwave, or laser devices to capture Earth's surface information by detecting and recording radiation or reflection across different spectral ranges. Conversely, natural images are captured using electronic devices (e.g., cameras) or sensors to record visible light, infrared radiation, and other forms of radiation present in the natural environment, thereby acquiring everyday image data. Unlike natural images captured horizontally by ground cameras, satellite images taken from an aerial perspective provide extensive imaging coverage and comprehensive information. In complex landscapes and urban environments, advanced structures and uneven distribution of background information can pose additional challenges [22]. Furthermore, due to the imaging method of remote sensing images, they encompass a wealth of information regarding various target objects. Consequently, these images frequently exhibit numerous instances of overlapping and varying-scaled targets, such as ships and ports, which are often arranged in a non-directional manner unnecessarily [23]. This necessitates that models designed for detecting remote sensing targets possess a highly perceptive ability in terms of accurate positioning [24] while also being sensitive to capturing informative details during the detection process. Additionally, the prevalence of small target instances in remote sensing images, some of which may consist of only a few pixels, poses significant challenges in feature extraction for the model [25], thereby resulting in performance degradation. Moreover, certain target instances in remote sensing images, such as flyovers and bridges, share strikingly similar features, intensifying the difficulties encountered in feature extraction for the model [26], consequently leading to phenomena such as false detections or missed detections. The presence of target instances in remote sensing images with extreme aspect ratios [27], such as highways and sea-crossing bridges, further exacerbates the challenges faced by the detector. Lastly, the complex background information within remote sensing images often leads to the occlusion of target regions by irrelevant backgrounds, rendering it difficult for the detector to extract target-specific features [28]. Moreover, the imaging method of remote sensing images is subject to environmental conditions on Earth's surface [29], including atmospheric interference, cloud cover, and vegetation obstruction, which may result in target occlusion and overlap, impeding the detector's ability to accurately delineate object contours [30] and consequently compromising the precise localization of target information. As a consequence, remote sensing images necessitate calibration and preprocessing measures [31]. Furthermore, in the current stage, numerous advanced detectors have achieved exceptional performance in remote sensing object detection through the design of neural network models' depth and width. However, this achievement comes at the cost of a substantial increase in model parameters. For instance, in remote sensing devices such as unmanned aerial vehicles and remote sensing satellites, it is impractical to equip them with mobile devices possessing equivalent computational power. As a result, the lightweight design of remote sensing object detection lags its progress in natural image domains. Hence,

effectively addressing the balance between model detection performance and lightweight design becomes an immensely valuable research question.

Deep learning-based object detection algorithms can be broadly classified into two categories. The first category consists of two-stage object detection algorithms that rely on candidate regions. These algorithms generate potential regions [32,33] and then perform classification and position regression [34,35], achieving high-precision object detection. Representative algorithms in this category include R-CNN [36], Faster R-CNN [37], Mask R-CNN [38], and Sparse R-CNN [39]. While these algorithms achieve high accuracy, their slower speed prevents real-time detection on all devices. The second category comprises single-stage object detection networks based on regression. These algorithms directly predict the position and class of objects from input images using a single network, avoiding the complex process of generating candidate regions and achieving faster detection speeds. The main representative networks in this category include SSD [40] and the YOLO [41–46] series. Among them, the YOLO series of single-stage detection algorithms is widely used. Currently, YOLOv5 strikes a balanced performance in the YOLO series.

The YOLO object detection model, proposed by Redmon et al. [47], achieves high-precision object detection performance while ensuring real-time inference. However, the individual training of each module in the YOLO model compromises the model's inference speed, thus the concept of joint training was introduced in YOLOv2 [48] to enhance the model's inference speed. The Darknet-53 backbone network architecture, first introduced in YOLOv3 [49], combines the strengths of Resnet to ensure highly expressive feature representation while avoiding gradient issues caused by excessive network depth. Additionally, multi-scale prediction techniques were employed to better adapt to objects of various sizes and shapes. In YOLOv4 [50], the CSPDarknet53 feature extraction backbone network integrated a cross-stage partial network architecture (CSP), effectively addressing information redundancy within the backbone network and significantly reducing the model's parameter count, thereby improving the overall inference speed. Moreover, the introduced Spatial Pooling Pyramid module in YOLOv4 helps expand the receptive field of the feature maps, further enhancing detection accuracy. As for YOLOv5, it strikes a balance in detection performance within the YOLO series. By employing CSPDarknet as the backbone network for feature extraction and adopting the FPN (Feature Pyramid Network) [51] approach for semantic transmission in the neck region, YOLOv5 incorporates multiple feature layers with different resolutions at the top of the backbone network. Convolutional and upsampling operations are utilized to fuse the feature maps and align scales. Furthermore, the PANet (Path Aggregation Network) [52] facilitates top-down localization. The YOLOv5 model has achieved favorable outcomes in natural image object detection tasks, but its effectiveness diminishes when applied to remote sensing satellite image detection due to challenges in meeting both real-time requirements and accuracy.

As a result, this study focuses on enhancing the YOLOv5 model for target detection in remote sensing images by proposing a faster and lightweight approach. The improvements include introducing a new backbone module called FasterConv, which replaces the C3 module in the original backbone network. This replacement reduces computational redundancy and optimizes memory access, thereby enhancing the inference speed of YOLOv5 across multiple devices. Additionally, a novel lightweight asymmetric detection head named LADH-Head is designed, inspired by the decoupled heads in YOLOX. By dividing the network paths based on the task type, this design significantly reduces parameters and GFLOPs compared to the original decoupled head module, leading to improved inference speed. Furthermore, the integration of Coordinate Attention into the Neck addresses the relatively high proportion of small objects in remote sensing images. The updated C3CA module models the coordinate information in the image, enabling a better understanding of spatial structure and positional relationships, thus improving performance in detecting small objects. To enhance detail and boundary preservation during upsampling operations, the content-aware reassembly upsampling module (CARAFE) [53] replaces the nearest-neighbor interpolation upsampling module in the original model. Finally, the original loss

function is refined by replacing CIoU with XIoU, resulting in enhanced model robustness, superior regression results, and improved convergence speed. Experimental results validate the outstanding performance of the proposed improved YOLOv5 model in remote sensing image object detection tasks. The paper presents the following contributions and innovations:

(1) Asymmetric decoupled detector head with a lightweight structure designed by combining deeply separable convolutions. By incorporating the groundbreaking lightweight asymmetric detection head, commonly referred to as LADH-Head, YOLOv5 achieves a significant reduction in computational complexity and a remarkable enhancement in inference speed. This innovative utilization marks a key milestone in the evolution of YOLOv5, leading to improved efficiency and accelerated performance.
(2) Designing C3CA Modules for Integrated Attention Mechanisms. By seamlessly integrating the Coordinate Attention mechanism into the C3 module of the Neck module within YOLOv5, the model adeptly captures and modeling the intricate spatial information present in the image. This refined approach significantly enhances the detection of object edges, textures, and other salient features, ensuring a heightened focus on diverse positional attributes.
(3) Within the scope of this article, the integration of the FasterConv module, accomplished by seamlessly incorporating PConv [54] into the backbone network of YOLOv5, guarantees a noteworthy decrease in model parameters and computations while upholding exceptional detection performance.
(4) Moreover, the introduction of the content-aware reassembly module (CARAFE) supersedes the conventional nearest-neighbor interpolation upsampling module in the original model. This advanced technique skillfully preserves intricate details and boundary information during upsampling operations, thus elevating the overall detection performance of the model.
(5) Substituting the loss function with XIoU in YOLOv5 not only strengthens the resilience of the original CIoU loss function but also accelerates the convergence speed of the model, leading to enhanced performance.

The remaining sections of this article are organized as follows: In Section 2, we present a comprehensive review of the relevant literature on remote sensing target detection and attention mechanisms. Section 3 provides a detailed description of the improvements made to our model. In Section 4, we validate the experimental results and conduct a visual analysis of the detection performance. Section 5 entails conducting ablation experiments on the proposed model architecture to demonstrate the feasibility of the design modules. Finally, in Section 6, we draw conclusions and outline prospects for future research endeavors.

2. Related Work

2.1. Traditional Object Detection in Remote Sensing Images

In the initial stages, object detection algorithms heavily relied on manual feature design given the absence of effective image representations. Due to the limitations of image encoding, these methods necessitated intricate feature representation schemes alongside various optimization techniques to accommodate the constraints of available computational resources. The underlying process of early approaches entailed pre-processing the target images, selecting relevant areas of interest [55], extracting distinctive attributes [56], and applying classifiers for categorization [57]. Primarily, superfluous details that lacked relevance to the object detection task were effectively filtered out through advanced image pre-processing techniques, thereby streamlining the data by retaining only the most essential visual elements. To localize potential regions where objects may be present, the sliding window technique was employed. By applying the Histogram of Oriented Gradients (HOG) algorithm [58], a diverse set of features including color, texture [59], shape [60], and spatial relationships [61] were extracted from these regions. Finally, the extracted features were transformed into vector representations and classified using an appropriate classifier. However, due to the large number of candidate regions involved in feature

extraction, the computational complexity increased significantly, resulting in redundant calculations. Moreover, manually engineered features demonstrated limited resilience and proved inadequate in complex and dynamic environments. Consequently, when it comes to object detection in remote sensing imagery, traditional machine learning-based methods have gradually been superseded by more efficient deep learning approaches, which have now become the primary choice.

2.2. Object Detection Based on Deep Learning Method in Remote Sensing Images

The field of deep learning has propelled neural networks to become integral components in modern target detection methods. Leveraging the powerful feature extraction capabilities of neural networks, deep learning-based algorithms have found widespread applications in remote sensing imagery. However, traditional target detection algorithms face challenges in achieving optimal performance due to complex backgrounds, varying target sizes, object overlap and occlusion, as well as the prevalence of small-scale targets in remote sensing images [62]. To address these complexities, researchers have introduced innovative techniques. Mashformer [63] presents a hybrid detector that integrates multi-scale perception convolutional neural networks (CNN) and Transformers. This integration captures relationships between remote features, thereby enhancing expressiveness in complex background scenarios and improving target detection across different scales. Considering the diverse orientations of objects in remote sensing images, Li et al. [64] propose an adaptive point learning method. By utilizing adaptive points as fine-grained representations, this method effectively captures geometric key features of objects aligned in any direction, even amidst clutter and non-axis-aligned circumstances. Addressing the issue of object boundary detection discontinuity, Yang et al. [65] introduce a novel regression loss function called Gaussian Wasserstein distance (GWD). This function aligns the specified loss with detection accuracy, enabling efficient model learning through backpropagation. For the problem of detecting small targets, Zhao et al. [66] suggest incorporating dedicated detection heads specifically designed for such targets. They also propose a cross-layer asymmetric Transformer module that leverages minute pathways to enrich the features of small objects, thereby improving the effectiveness of small target detection while reducing model complexity. To combat specific image degradation characteristics induced by remote sensing imaging techniques, Niu et al. [67] propose an effective feature enhancement (EFE) block. This block integrates a non-local means filtering method to address issues such as weak target energy and low image signal-to-noise ratio, enhancing the quality of features. Yan et al. [68] devised a novel detection network called LssDet, which not only ensures accurate target detection but also reduces the complexity of the model. This method enhances the feature extraction capabilities specifically for small targets. Furthermore, CenterNet [69] and CornerNet [70] improve target detection speed through methodologies that focus on detecting center points and corner points, respectively.

On the whole, these advancements contribute to the ongoing improvement and effectiveness of target detection in remote sensing imagery. However, despite the significant enhancement in detection accuracy achieved by existing methods, they come at the cost of substantial computations and parameterization. This poses a challenge as the current approaches struggle to strike a harmonious balance between lightweight design and detection performance. Consequently, when applied to real-time or mobile devices, the efficacy of these methods for target detection in remote sensing images diminishes. Therefore, it becomes crucial to address the pressing issue of effectively reconciling the performance of remote sensing image detection with the imperative for model lightweight.

2.3. The Attention Mechanism

Attention mechanism is a widely employed technique in the field of deep learning which plays a similar role to human attention. It focuses on the most important and relevant parts of information during the processing stage. By mimicking human visual or attention processes, this mechanism helps models emphasize crucial information, enabling

neural networks to adapt perceptively to visual tasks and dynamically adjust their focus on inputs. Currently, attention mechanisms find extensive applications in various tasks, including image classification [71], image semantic segmentation [72], object detection [73], natural language processing [74], medical image processing [75], and image generation [76]. The Recurrent Attention Model (RAM) [77] was the first to apply attention mechanisms to deep neural networks. Attention mechanisms can be categorized into different types: channel attention, spatial attention, hybrid attention, temporal attention, and branch attention. Channel attention aims to automatically learn attention mechanisms for each channel and adjust the weights of channels accordingly. SENet [78] was the pioneering work that introduced channel attention, collecting global information through squeeze-and-excitation to capture channel-wise information and enhance feature representation and discrimination. Spatial attention involves automatically learning the importance of each spatial position within an image and adjusting the weights of positions accordingly. The Spatial Transformer Network (STN) [79] is a representative method that transforms various deformable data in space and automatically captures features from important regions. GENet [80] implicitly utilizes sub-networks to predict soft masks for selecting significant regions. Hybrid attention combines channel attention and spatial attention. Notable algorithms include DANet [81], which introduces both channel and spatial attention to capture global and contextual information by adaptively learning channel and spatial weights. Woo et al. [82] propose a lightweight attention mechanism called the Convolutional Block Attention Module (CBAM), decoupling spatial attention and channel attention to improve computational efficiency. The tremendous success of the Transformer model [83] in the field of natural language processing (NLP) has brought attention to self-attention mechanisms, which have been introduced into computer vision. Vision Transformers [84] and Swin-Transformers [85], based on attention mechanisms, achieve excellent detection accuracy and speed without using convolutional operations, showcasing the enormous potential of pure attention-based models in computer vision. However, due to the sliding window approach employed by Transformers for image processing, the computational complexity remains high, resulting in unsatisfactory performance when detecting small targets [86].

3. Method

3.1. The Original YOLOv5 Algorithm

The YOLOv5 model consists of three primary components: the Backbone, Neck, and Head modules. The Backbone network is responsible for extracting feature information from input images, while the Neck module aggregates these features to generate three feature maps at different scales. Upon these features, the Head module performs object detection. The YOLOv5 model primarily adopts the CSPDarkNet53 structure, utilizing the Convolutional (Conv) layer, C3 layer, and SPPF layer in the backbone network. The Conv layer comprises convolution, batch normalization, and the SiLU function. By employing residual connections, the C3 module effectively reduces model parameters, thereby improving the inference speed. SPPF, an improved version of the original SPP module, replaces the max-pooling layers of size 5×5, 9×9, and 13×13 with three max-pooling layers of size 5×5. This modification ensures the fusion of features with different receptive fields while further enhancing operational speed. In terms of the Neck component, YOLOv5 incorporates the Path Aggregation Network (PANet), which extends upon the Feature Pyramid Network (FPN) by introducing bottom-up pathways. After top-down feature fusion in FPN, the bottom-up pathways transmit positional information from lower levels to deeper ones, significantly enhancing localization ability across multiple scales. YOLOv5 includes five derivative models: YOLOv5n, YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. These models share the same architecture but vary in width and depth. For a comprehensive illustration of the original YOLOv5 model's overall structure, please refer to Figure 1.

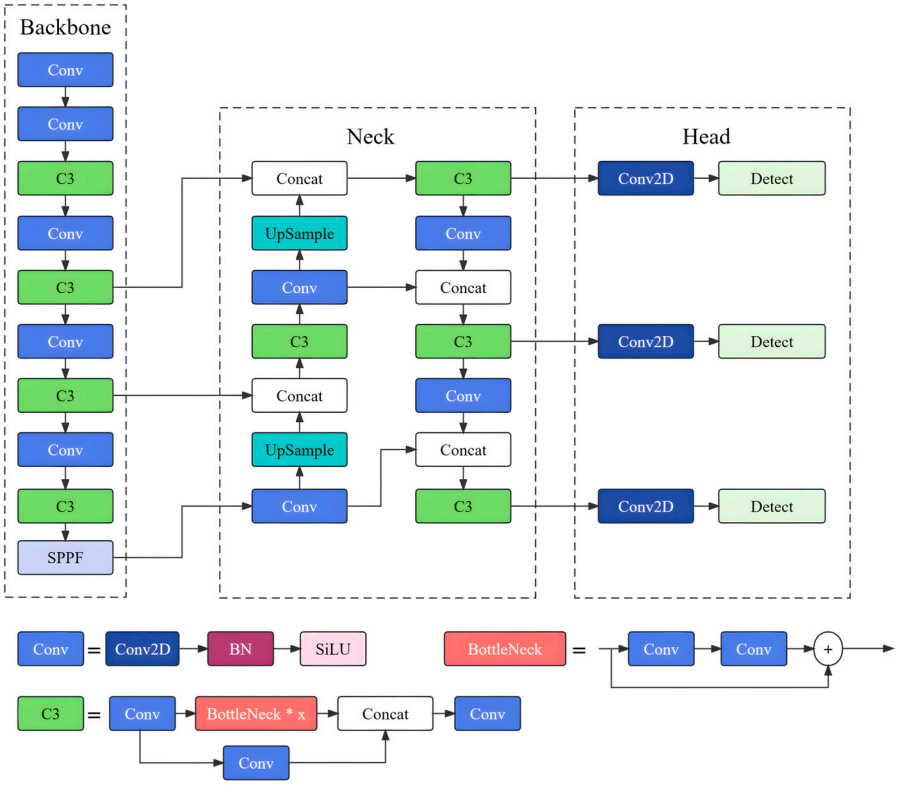


Figure 1. The structure of the YOLOv5s model.

3.2. Proposed Method

In this section, we will delve into the key enhancements of our innovative approach, which include the integration of the FasterConv module, the LADH-Head module, the C3CA module, and the utilization of the XIoU loss function. Furthermore, Figure 2 illustrates the architectonic rendition of our refined YOLOv5 model.

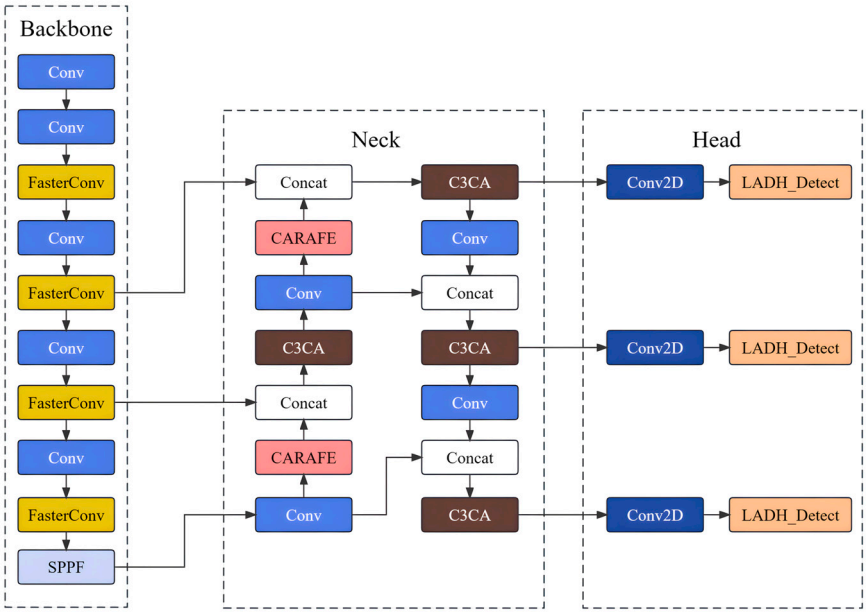


Figure 2. Network architectonic of our proposed YOLOv5s model.

3.2.1. LADH-Head

The original YOLO algorithm employs a coupled head for detection tasks, utilizing the same convolutional layers at the network's top for both classification and regression. However, these tasks have distinct focuses, leading to conflicts during the detection process. While classification primarily emphasizes sample texture, regression focuses on target edge features. Consequently, the YOLOv5 model experiences a decline in detection performance. Moreover, the prevalence of small targets and complex backgrounds in remote sensing images exacerbates the issue, reducing YOLOv5's detection accuracy.

To overcome these challenges, the concept of a decoupled head module was introduced in YOLOX [45]. This model aims to resolve conflicts arising from the coupled head's classification and regression tasks. Figure 3 [45] visually illustrates the structural differences between the decoupled head and YOLOv5.

For the input feature map, the decoupled head employs a 1×1 convolution to decrease the feature channels by 256. Subsequently, it incorporates two parallel branches for the classification (Cls.) and regression (Reg.) tasks. Each branch integrates two 3×3 convolutions to streamline the model and enhance convergence speed. Furthermore, an IoU branch is introduced within the regression branch to indicate the presence of objects in each feature point. By independently processing the classification and regression tasks, the model's detection performance is enhanced. However, although the decoupled head significantly improves the model's detection capability, it also presents notable challenges. Firstly, this design substantially increases the network's parameter count, resulting in decreased inference speed. Secondly, the decoupled head fails to fully exploit the high-level features extracted by the backbone network due to the complementary nature of its feature representation. Consequently, detection accuracy is diminished, particularly when dealing with complex backgrounds and varying target sizes in remote sensing images.

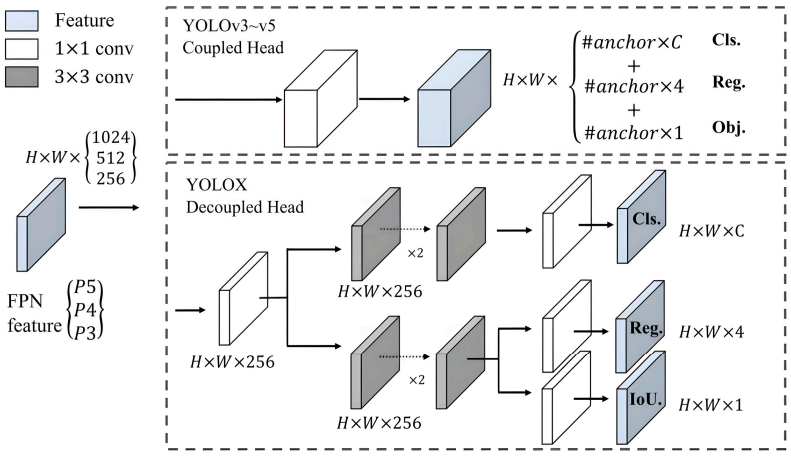


Figure 3. YOLOX Decouple Head structure.

To tackle these issues, we drew inspiration from the Asymmetric Decouple Head (ADH) [87], incorporating multi-level channel compression to develop a lightweight asymmetric detection head (LADH-Head). Figure 4 provides an overview of our proposed detection head's structure.

By segregating the network based on task types, we utilize three distinct channels to perform the relevant tasks. To expand the receptive field and increase task parameters for the IoU branch, we employ three convolutions that reduce features along the channel dimension. In place of the conventional 3×3 convolutions in each branch, we employ 3×3 depth-wise separable convolutions (DWConv) as replacements in the ADH network.

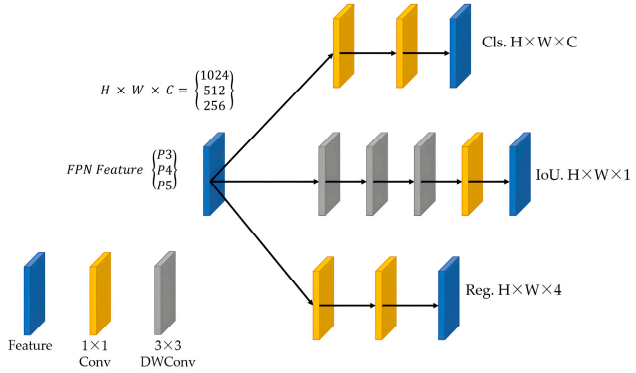


Figure 4. Network structure of the lightweight asymmetric detection head (LADH-Head).

The advantage of DWConv over standard convolution is that it significantly reduces the number of parameters by decomposing the convolution operation into Depthwise Convolution and pointwise convolution. The number of feature maps after Depthwise Convolution is the same as the number of channels in the input layer, and it is not possible to extend the feature maps. Since this operation is performed independently for each channel in the input layer, it is not possible to utilize the feature information of different channels at the same spatial location effectively. Therefore, pointwise convolution is needed to combine these feature maps to generate a new feature map, but since it cannot effectively utilize the feature information of different channels at the same spatial location, point-wise convolutional is used later for dimensionality expansion. The operation of pointwise convolution is very similar to the regular convolutional operation and is used to mix the information between channels by applying a 1×1 convolutional kernel between different channels. It does not change the spatial dimension of the feature map, only the number of channels. So the point-by-point convolution operation combines the previous maps weighted in the depth direction to generate a new feature map. Secondly, due to the reduction in the number of parameters, the depth separable convolution is computationally more efficient and is suitable for use in resource-limited environments; and finally, the information from different channels is combined through the point-by-point convolution to maintain a certain level of feature extraction capability.

This approach further reduces model parameters and enhances detection speed. Separating the classification and bounding box tasks using 3×3 depth-wise separable convolution layers prevents excessive expansion of both tasks since positive samples matched to both tasks have relatively smaller associated losses. By replacing the decoupled head in the YOLOv5 network with LADH-Head, model parameters are significantly reduced while improving detection accuracy, effectively resolving conflicts between the classification and regression tasks introduced by the original coupled head.

3.2.2. C3CA

Due to the imaging modality, remote sensing images often exhibit intricate backgrounds and significant variations in target orientation. Consequently, the original YOLOv5 algorithm suffers from a loss of essential feature information for small targets during the sampling process. This leads to the occurrence of false positives and missed detections, ultimately diminishing the overall detection performance. In light of this challenge, we propose a redesign of the Neck component within the YOLOv5 model. Our approach involves incorporating the Coordinate Attention [88] technique and introducing the C3CA module. The primary objective of this enhancement is to improve the accuracy of detecting small targets within complex backgrounds without increasing the model's parameters or introducing redundant computations.

The Coordinate Attention mechanism, which is a variant of the channel attention mechanism known as Squeeze-and-Excitation [78], plays a crucial role in our solution. Figure 5 [78] demonstrates the network structure of SENet. SENet attention mechanism

as a classical channel attention mechanism, which is mainly composed of two parts, the compression module and the excitation module, after the convolution of the feature map, and then its own spatial convolution of the $1 \times 1 \times C$ feature map obtained after multiplication, to obtain the final feature map, where the C represents the number of channels. The $1 \times 1 \times C$ feature map has a global receptive field, which contains global information, and when it is multiplied with the $H \times W \times C$ feature map, it obtains the feature map after the action of the attentional mechanism; this makes the effective feature map channel have a large weight, and the ineffective or ineffective feature map channel has a small weight to train the model to achieve better results. However, SENet only considers the relationship between the modeling channels to weight each channel, ignoring the position information, because changes in the target position information and scale information in the remote sensing image will lead to errors in the model in the detection process of the target spatial and positional feature information extraction in the generation of spatially selective feature maps bias, which will lead to a decline in the performance of the detection of small targets.

To address this limitation, our Coordinate Attention mechanism incorporates positional information into the channel attention process. This enables the network to capture directionality and position-sensitive information effectively. More specifically, our method decomposes the channel attention mechanism into two one-dimensional feature encoding processes. These processes independently capture long-range dependencies and positional information along two spatial directions. By doing so, the resulting attention maps encode both directionality and position sensitivity, thereby significantly enhancing the representation of objects of interest within the input feature map.

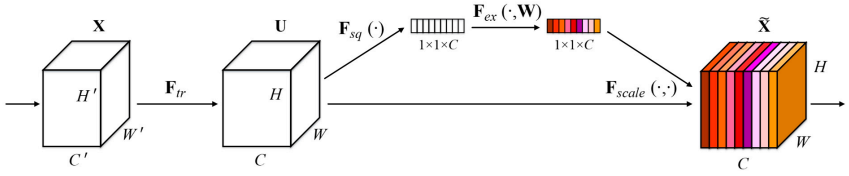


Figure 5. Network structure of SENet.

Coordinate Attention employs coordinate information embedding and the coordination between two attention processes to encode channel relationships and long-term dependencies. Figure 6 shows the structure of Coordinate Attention.

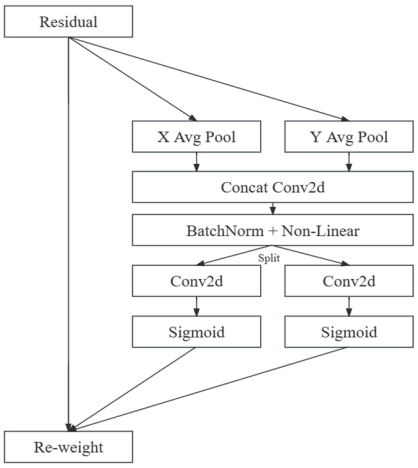


Figure 6. Network structure of Coordinate Attention.

Initially, the global pooling method in SENet enables the compression of spatial information into channels during the overall encoding process of channel attention. However,

this approach also poses challenges in preserving positional information. Consequently, the squeezing stride for c channels of the given input X is defined as:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_c(i,\ j) \tag{1}$$

Therefore, through the technique of factorization, global pooling is transformed into a set of one-dimensional feature coding operations. This transformation ensures that the attention module effectively captures precise positional information during remote interactions. For a given input X, two spatial ranges with dimensions (H, 1) or (1, W) are employed to encode all channels along both the horizontal and vertical orientations. The resulting output of the c-th channel at height h is defined as:

$$z_c^h(h) = \frac{1}{W} \sum_{0 \le i < W} x_c(h,\ i) \tag{2}$$

Similarly, the output of the c-th channel with width w can be expressed as:

$$z_c^w(w) = \frac{1}{H} \sum_{0 \le j < H} x_c(j,\ w) \tag{3}$$

Two distinct transformation techniques converge feature aggregation along diverse spatial directions, resulting in an ensemble of feature maps that possess a heightened sense of directionality. These maps play a pivotal role in facilitating the network's enhanced precision when localizing target information.

Subsequently, the Coordinate Attention Generation method harnesses the expressive representations produced from (2) and (3). Initially, z_c^h and z_c^w are seamlessly inputted into the shared 1×1 convolutional transformation function F_1:

$$\mathbf{f} = \delta\left(F_1\left(\left[z^h, z^w\right]\right)\right) \tag{4}$$

where the [•, •] represents the concatenation operation along the channel dimension. δ denotes the non-linear activation function Hswish, and $\mathbf{f} \in \mathbb{R}^{C/r \times (H+W)}$ signifies the intermediate feature mapping of spatial information in both horizontal and vertical directions. Here, r is employed to control the scaling ratio of the number of channels. Along the channel dimension, $\mathbf{f}$ is divided into two separate vectors: $\mathbf{f}^{\boldsymbol{h}} \in \mathbb{R}^{C/r \times H}$ and $\mathbf{f}^{\boldsymbol{w}} \in \mathbb{R}^{C/r \times W}$. Through two 1×1 convolutions, $\mathbf{f}^{\boldsymbol{h}}$ and $\mathbf{f}^{\boldsymbol{w}}$ are transformed into $\mathbf{F}^{\boldsymbol{h}}$ and $\mathbf{F}^{\boldsymbol{w}}$, respectively, ensuring they possess the same number of channels as the output X.

$$\boldsymbol{g}^h = \sigma\left(\boldsymbol{F}_h\left(\mathbf{f}^{\boldsymbol{h}}\right)\right) \tag{5}$$

$$\boldsymbol{g}^w = \sigma(\boldsymbol{F}_w(\mathbf{f}^{\boldsymbol{w}})) \tag{6}$$

where the σ denotes the Sigmoid function, $\boldsymbol{g}^{\boldsymbol{h}} \in \mathbb{R}^{C/2 \times H}$ and $\boldsymbol{g}^{\boldsymbol{w}} \in \mathbb{R}^{C/2 \times W}$. Finally, the Coordinate Attention module can be expressed as:

$$y_c(i,j) = x_c(i,j) \times g_c^h(i) \times g_c^w(j) \tag{7}$$

Figure 7a showcases the architectural framework of the C3CA model and the strategic integration of Coordinate Attention within the bottleneck. With seamless grace, we intertwine Coordinate Attention with the refined C3 module, artfully replacing the original C3 module in the Neck with the C3CA module. The CA module can be embedded at various positions to extract and enhance spatial positional information. We incorporate the CA module within the Bottleneck, where the two convolutional layers primarily extract and amplify image features. Initially, a 1×1 convolutional layer extracts the image features,

while a 3 × 3 convolution further enhances the extracted target features. Typically, we embed the CA module in between these two convolutional layers, as illustrated in Figure 7b. However, in this study, to improve the detection of small targets amidst complex backgrounds, we have modified the existing connection scheme by embedding the CA module after the two convolutional layers, as depicted in Figure 7c.

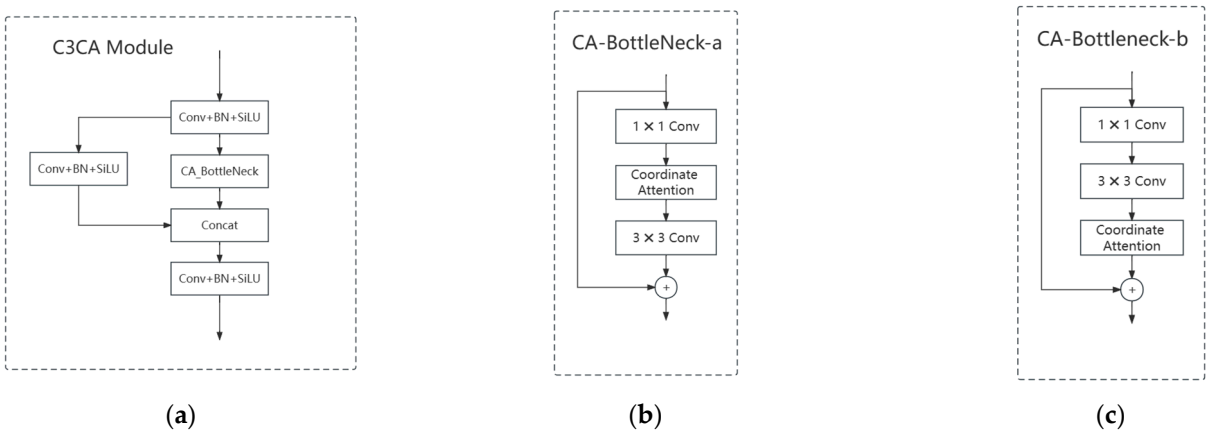


Figure 7. Network architectonic of C3CA module and CA_BottleNeck. (**a**) represents the C3CA module, (**b**) represents the conventional method of inserting Bottleneck by the attention mechanism, and (**c**) represents the embedding approach we adopt.

This augmentation elevates the target-detection capability of the YOLOv5 model when confronted with intricate backgrounds. By enabling the model to harness the full potential of positional information, we ensure the efficient extraction of target features throughout the detection process. This empowers the model to hone in on vital cues of the target object amidst complex backdrops. Notably, the incorporation of C3CA significantly enhances the model's detection performance without imposing excessive parameters or computational costs, aligning perfectly with our pursuit of a lightweight network design.

3.2.3. FasterConv

In order to construct a more lightweight YOLOv5 detection network to ensure that the model detects the performance of remote sensing images, and at the same time, better achieve the balance between the lightweight and detection performance, MobileNet, GhostNet, and other lightweight backbone networks were added to the YOLOv5s model. However, due to restrictions caused by the lightweight structure, we encountered a detection performance degradation problem which made it difficult to achieve satisfactory results for remote sensing target detection on the mobile device. Inspired by the groundbreaking FasterNet [54], PConv (Partial Convolution) provides a superior approach to achieve light-weighting of the model and accelerate inference speed. In our innovative design, we introduced the FasterConv module as a formidable replacement for the C3 module within the backbone network of the illustrious YOLOv5 model. The C3 module in YOLOv5 comprises three Conv blocks, each ingeniously harnessing the power of a 3 × 3 convolutional kernels. These Conv blocks augment the network's depth and receptive field, thereby empowering an enhanced ability to extract intricate features. Nevertheless, the Conv layers' exorbitant parameter count incurs redundant computations during inference, culminating in heightened GFLOPs and impeding the model's aspiration for swift and lightweight detection. Conversely, eminent techniques employed in distinguished architectures such as ShuffleNet [89] and GhostNet [90], namely Group Convolution and Depth-wise Separable Convolution (DWConv) [91], adeptly curtail parameters by astutely capitalizing on redundancy within filters. These approaches also come at the cost of expanding the network's width, which inevitably engenders diminished inference speed and potentially compromises detection accuracy. These lightweight and faster networks

prioritize DWConv and GConv to effectively minimize the abundance of floating-point operations (FLOPs).

However, one must exercise prudence in assuming that reducing FLOPs guarantees an upswing in inference speed. Regrettably, such nimble and sprightly networks oftentimes entail supplementary data operations such as concatenation, shuffling, and pooling, unwittingly sowing the seeds of suboptimal efficiency when it comes to FLOPS (floating-point operations per second). The remarkable PConv seamlessly unlocks the latent potential residing within device computational capabilities through judicious reduction of memory access. This ingenious feat translates into a discernible enhancement in inference speed while concurrently mitigating the burden of FLOPs. PConv artfully exploits the redundancy nestled within feature maps by selectively employing traditional convolutions (Conv) to extract spatial features from certain input channels without perturbing their unaffected counterparts. Consequently, the FLOPs associated with PConv are unequivocally diminished, unveiling a newfound era of efficiency. The FLOPs of a PConv are

$$h \times w \times k^2 \times c_p^2 \tag{8}$$

For regular convolution with input $\mathbf{X} \in \mathbb{R}^{c \times h \times w}$ and output $\mathbf{Y} \in \mathbb{R}^{c \times h \times w}$, the FLOPs are

$$h \times w \times k^2 \times c^2 \tag{9}$$

where c and c_p are the channel number, and h and w are the height and width of the input data. Taking into account that PConv selects the initial or final consecutive channels to represent the complete feature map, we deduce that both the input and output feature maps share the same set of channels. With a proportional ratio of $R = \frac{c_p}{c} = \frac{1}{4}$, the FLOPs associated with PConv are reduced by a staggering sixteen-fold in comparison to conventional convolutions. This extraordinary reduction substantially amplifies the network's inference speed while concurrently reducing the overall parameter count of the model. The working principle of PConv is illustrated in Figure 8 [54].

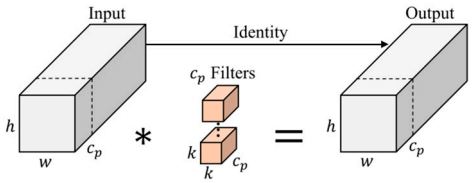


Figure 8. The working principle of PConv.

In order to strike a harmonious balance between the accuracy of detection and computational efficiency, we have devised the FasterConv module by merging the PConv structure with the original C3 module. The overall structure of FasterConv is succinctly presented in Figure 9. Within each FasterBlock, a 3×3 PConv layer takes center stage, ingeniously integrating the concept of pointwise convolution (PWConv) to maximize the utilization of all channel information. By incorporating PWConv, the network becomes more adept at focusing its attention on the central region of the target's receptive field, evoking the behavior exhibited by standard convolutions (Conv). To refine the receptive field on the input feature map, we have judiciously inserted two 1×1 Conv layers subsequent to the PConv layer. Importantly, we have established residual connections between PConv and Conv, ensuring a rich assortment of features while optimizing both rapid inference speed and minimal latency. To expedite inference, we have strategically placed BatchNormalization [92] after the intermediate Conv layer, complemented by the adoption of the GELU [93] activation function. The structural configuration of FasterBlock is visually depicted on the left side of Figure 9. Through seamless integration of FasterBlock into YOLOv5, we effectively supplant the original C3 module within the CSPDarkNet53 backbone network, leading to further acceleration in the inference speed for remote sensing image detection. This

modification achieves a reduction in model parameters without compromising detection performance, simultaneously accentuating the extraction of spatial features.

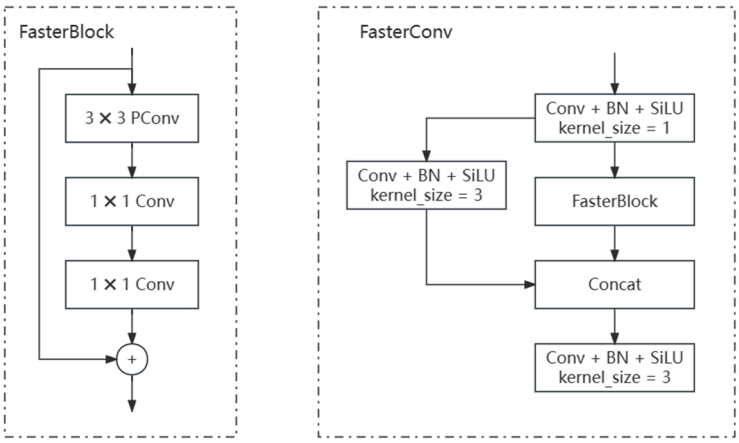


Figure 9. Network structure of FasterBlock and FasterConv model.

3.2.4. CARAFE Module

In the pursuit of multi-scale feature fusion, the YOLOv5 model employs the nearest-neighbor interpolation method for upsampling. The Feature Pyramid Network (FPN) plays a pivotal role in facilitating top-down semantic information propagation and multi-scale object detection. However, this method tends to prioritize spatial information while overlooking the crucial aspect of semantics. By solely relying on pixel positions to determine the upsampling kernel, the model's receptive field is relatively limited, resulting in the loss of comprehensive global contextual information. To address this concern and mitigate semantic information loss during upsampling in the YOLOv5 model's object detection process, an innovative experimental technique called content-aware reassembly of features (CARAFE) [53] takes center stage as a viable substitute for the conventional nearest-neighbor interpolation upsampling method.

The CARAFE module revolutionizes upsampling by reassembling features within specific regions surrounding each central position through weighted combinations. This intelligent aggregation enables a broader receptive field, effectively capturing more extensive feature information. The CARAFE module comprises two key components: the upsampling kernel prediction module and the feature reassembly module. Firstly, the upsampling kernel prediction module forecasts the upsampling kernel while simultaneously reducing the number of channels in the input features. Subsequently, the feature reassembly module encodes the compressed features and performs point-wise multiplication between the predicted reassembly kernel at the target position and the corresponding region of the original feature map, thereby accomplishing the upsampling process. Compared to the conventional nearest-neighbor interpolation upsampling method employed by YOLOv5, the CARAFE module boasts an expanded receptive field, enabling content-aware processing for specific instances, ultimately enhancing the capture of semantic details inherent in the targets. Moreover, the CARAFE module demonstrates its prowess by employing a lightweight design that facilitates faster network inference calculations.

3.2.5. Loss Function

The YOLOv5 loss function comprises three integral components: the classification loss function (L_{cls}), the confidence loss function (L_{obj}), and the localization loss function (L_{bbox}). It can be elegantly expressed as follows:

$$Loss = L_{cls} + L_{obj} + L_{bbox} \tag{10}$$

The computation of the classification loss function (L_{cls}) and the confidence loss function (L_{obj}) involves the utilization of the binary cross-entropy function (BCE Loss), and can be expressed as:

$$L_{cls} = \sum_{i=0}^{S^2} \sum_{j=0}^{B} \sum_{c \subset classs} [-\hat{p}_i(c) \ln (p_i(c)) - (1 - \hat{p}_i(c) \ln(1 - p_i(c)))] \tag{11}$$

$$\begin{aligned} L_{obj} = \lambda_{obj} \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{ij}^{obj} \left[-\hat{C}_i^j \ln \left(C_i^j \right) - \left(1 - \hat{C}_i^j \right) \ln \left(1 - C_i^j \right) \right] \\ + \lambda_{nobj} \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{ij}^{nobj} \left[-\hat{C}_i^j \ln \left(C_i^j \right) - \left(1 - \hat{C}_i^j \right) \ln \left(1 - C_i^j \right) \right] \end{aligned} \tag{12}$$

where the S represents the size of the grid, while B signifies the number of predicted boxes per grid. The notation I_{ij}^{obj} indicates whether the i-th box predicted in the j-th grid contains an object. The coefficients λ_{obj} and λ_{nobj} determine the balance between confidence losses. Additionally, C_i^j and $\hat{C}_i^j$, respectively, represent the confidences of predicted boxes and ground truth boxes. Furthermore, $p_i(c)$ and $\hat{p}_i(c)$ correspondingly convey the predicted probability and true probability of class c during the object detection process within the i-th network.

On another note, the localization loss function L_{bbox} is evaluated through the Intersection over Union (IoU) method, which holds a crucial role in bounding box regression. YOLOv5 employs the CIoU method for calculation, taking into consideration geometric factors such as overlapping area [94], center point distance [95], and aspect ratio [96] within the regression loss function. These components are measured using IoU, Euclidean geometric distance, corresponding aspect ratios, and angles to precisely quantify the extent of overlap between targets and their corresponding ground truth values.

$$\mathcal{L}_{CIoU} = 1 - IOU + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha v \tag{13}$$

The $\rho^2(b, b^{gt})$ represents the Euclidean distance between the predicted box and the ground truth box. Moreover, we employ the variable c to denote the diagonal length of the minimum enclosing box that effectively encompasses both the predicted box and the true box. Additionally, v is utilized to assess the consistency in aspect ratios between the predicted box and the ground truth box. Furthermore, the coefficient α serves as a balancing factor in this context. The Equations (14) and (15) outline the specific computation methods for evaluating v and α, respectively.

$$\alpha = \frac{v}{(1 - IOU) + v} \tag{14}$$

$$v = \frac{4}{\pi^2} \left(\tan^{-1} \frac{w^{gt}}{h^{gt}} - \tan^{-1} \frac{w}{h} \right)^2 \tag{15}$$

Here, w and h represent the width and height of the predicted box, while w^{gt} and h^{gt} correspondingly signify the width and height of the ground truth box.

However, the utilization of the arctangent function in calculating the penalty term for rectangle aspect ratio, represented by coefficient v, in the CIoU loss function compromises its robustness. This vulnerability renders it susceptible to outliers, leading to significant fluctuations in the loss function and subsequently affecting its performance. Furthermore, the range of values (0, $\pi/2$) of the arctangent function fails to meet the normalization requirement of the loss function. Our objective is to enhance the robustness of the loss function against exceptional cases without imposing additional computational burden.

In pursuit of this, we introduce a novel penalty term parameter v', as demonstrated in Equation (16):

$$v' = \left(\frac{1}{1 + e^{-\frac{w^{gt}}{h^{gt}}}} - \frac{1}{1 + e^{-\frac{w}{h}}} \right)^2 \tag{16}$$

By emphasizing varying levels of target overlap within the loss function, we effectively enhance the resilience of bounding boxes in object detection tasks, all while maintaining computational efficiency. This enhancement significantly improves the model's detection capabilities, especially in complex backgrounds. Notably, when compared to the original CIoU penalty term parameter, our proposed parameter exhibits a smoother gradient for the penalty term, resulting in reduced regression errors.

4. Experiment Results

To validate the effectiveness of our proposed method for remote sensing image object detection, we conducted experiments on three widely used datasets: DIOR [97], DOTA [98], and SIMD [99]. In this section, we provide an overview of the experimental setup, including the environment, settings, and evaluation metrics. Furthermore, we focus on validating the performance of our method primarily on the DIOR dataset, comparing it with state-of-the-art approaches. Additionally, we analyze the results on the DOTA and SIMD datasets to further investigate the feasibility of our approach. Finally, we perform ablation experiments on the DIOR dataset to assess the effectiveness of each module in our method.

4.1. Experiment Environment and Details

The experimental setting is orchestrated as depicted in Table 1, where all experiments were conducted under identical configurations.

Table 1. Experiment Environment.

Configuration	Parameters
CPU	i5-12600KF
GPU	RTX3070 8 GB
Operation System	Windows 10
Environment	Pytorch 1.11.0, Python 3.7, CUDA11.3

In our experiments, we train the proposed approach using the SGD optimizer with an initial learning rate of 0.01. The DIOR dataset was trained for 300 epochs, while both the DOTA and SIMD datasets underwent 150 epochs of training, utilizing a batch size of eight. To effectively optimize the model's learning rate, the first three epochs were dedicated to warm-up training. During this initial phase, the learning rate was set at 0.1 and subsequently reduced to 0.01 for the remaining training process. The DIOR dataset employed images with dimensions of 800 $\times$ 800, while the DOTA and SIMD datasets enlarged the image size to 640 $\times$ 640. We set the momentum to 0.937 and used a weight decay coefficient of 5e-4. All other parameters remained consistent with the original YOLOv5 configuration. We applied the same parameter settings across all experiments involving alternative methods.

4.2. Datasets

The DIOR dataset serves as a fundamental benchmark for optical remote sensing images. It comprises 23,463 images captured across 19,2474 instances in diverse scenes, encompassing 20 object classes such as airplanes (ALs), airports (ATs), baseball fields (BFs), basketball courts (BCs), bridges (Bs), chimneys (Cs), dams (Ds), expressway service areas (ESAs), expressway toll stations (ETAs), golf courses (GCs), ground track fields (GTFs), harbors (Hs), overpasses (Os), ships (Ss), stadiums (SDs), storage tanks (STs), tennis courts (TCs), train stations (TSs), vehicles (Vs), and windmills (Ws) [97]. With an image size of

800 × 800 and a spatial resolution ranging from 0.5 to 30 m, the DIOR dataset utilizes 11,725 images for training and validation, along with 11,738 images for testing. This dataset showcases significant variations and rich image transformations within and across different object categories, offering comprehensive evaluations of model robustness and detection performance by presenting diverse changes in size and background. Furthermore, it introduces multiple variations within the same object class through alterations in color and shape, effectively assessing the accuracy of target detection.

The DOTA dataset, proposed by Wuhan University, is a large-scale dataset specifically designed for optical remote sensing image object detection. It consists of 2806 aerial images, each measuring 4000 × 4000 pixels, with the spatial resolution ranging from 10 to 300 m, encompassing 15 object classes and a total of 188,282 instance objects [98]. Due to the large image sizes, direct input into the network for training is not feasible. Therefore, in our experiments, we cropped the images in the DOTA dataset to generate approximately 20,000 images sized at 1000 × 1000 pixels. These images were then divided into training and validation sets in a 7.5:2.5 ratio.

The Satellite Imagery Multivehicle Dataset (SIMD) comprises 5000 images with a resolution of 1024 × 768 pixels, and a spatial resolution ranging from 500 to 1000 feet. It encompasses around 45,096 object instances belonging to 15 different vehicle categories [99]. In contrast to the DIOR and DOTA datasets, SIMD focuses on detecting movable targets, emphasizing general-purpose vehicles with similar appearances that are also mobile. The dataset primarily includes automobiles, trucks, buses, and other vehicle types. For our experiments, we partitioned the dataset into a 4:1 ratio, allocating 4000 images for the training set and 1000 images for the validation set.

4.3. Evaluation Metrics

In this experiment, performance metrics such as precision (*AP*), mean average precision (mAP@0.5), frames per second (FPS), and parameter count (Parameters) were employed to evaluate the detection capabilities of the model. Precision (*P*) and recall (*R*) are computed based on true positives (*TP*), false positives (*FP*), true negatives (*TN*), and false negatives (*FN*):

$$P = \frac{TP}{TP + FP} \tag{17}$$

$$R = \frac{TP}{TP + FN} \tag{18}$$

By plotting precision and recall on a precision-recall curve (*P*-*R*), the area under the curve (*AP*) is defined, and the average precision (*mAP*) is calculated as the mean value of *AP* across all categories.

$$AP = \int_0^1 P(R)dR \tag{19}$$

$$mAP = \frac{1}{N}\sum AP \tag{20}$$

4.4. Experiment Result and Analysis

4.4.1. Model Performance in the DIOR Datasets

In the context of the DIOR dataset, we validate the feasibility of our proposed method in comparison to the original YOLOv5s model. The results are presented in Table 2.

Table 2. Detection result between our method and YOLOv5s on the DIOR dataset.

Method	Precision (%)	Recall (%)	mAP@0.5 (%)	mAP@[0.5:0.95] (%)	Parameters	GFLOPs	FPS
YOLOv5s	0.813	0.751	0.798	0.519	7.06M	16.1	109.0
Our Method	0.875	0.761	0.831	0.576	6.84M	15.5	111.9

Table 2 demonstrates that our method achieves an mean average precision (mAP@0.5) of 83.1%, which is a 3.3% improvement over the original YOLOv5s model. Our method also exhibits excellent detection performance at large thresholds, compared to the YOLOv5s model in mAP@ [0.5: 0.95] increased by 5.7%. Additionally, the table provides a comparison of the model's parameters and detection speed (FPS). Compared to the original model, parameters are reduced by 0.22 M. Compared to the original model, our proposed method exhibits a significant speed improvement of 2.9 ms in detection. This meets our objective of designing a fast and lightweight detection model, demonstrating the excellent detection performance of our model.

In general, the Intersection over Union (IoU) threshold and the confidence threshold are two pivotal indicators of deep learning models. For remote sensing target detection, given the higher likelihood of overlapping targets against intricate backgrounds, we set the IoU threshold and confidence threshold at 0.6 and 0.25, respectively. Bridges and ports exhibit a higher false negative rate (FN) due to their limited representation in the dataset. Consequently, an elevated FN suggests a greater number of instances being overlooked, resulting in lower average precision (AP). Additionally, categories such as vehicles, being small objects within remote sensing images, suffer from higher false positive rates (FP) due to object occlusion amidst complex backgrounds. This issue leads to false detections and consequently lowers the overall detection accuracy.

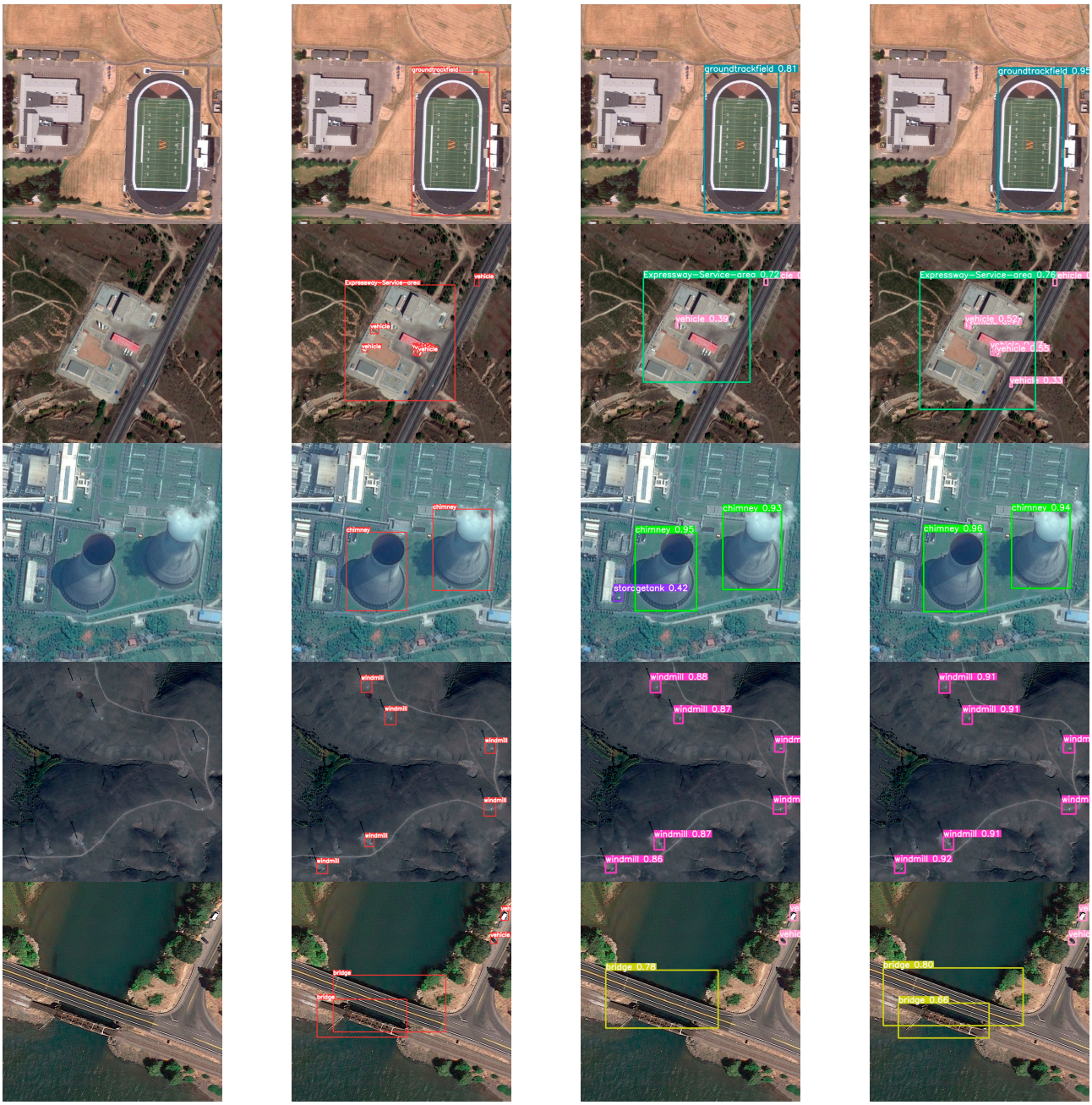
Furthermore, we conduct a comparative analysis between our proposed model and widely used detection algorithms, evaluating the detection accuracy for the 20 classes within the DIOR dataset. The results of this analysis are presented in Table 3.

Table 3. Performance comparison of different models on the DIOR dataset (The black bold font represents the best detection method for each target instance type).

Method	mAP	AL	AT	BF	BC	B	C	D	ESA	ETA	GC
CANet [100]	74.3	70.3	82.4	72.0	87.8	55.7	79.9	67.7	83.5	77.2	77.3
Faster-RCNN [37]	65.1	54.0	74.5	63.3	80.7	44.8	72.5	60.0	75.6	62.3	76.0
TRD [101]	66.8	77.9	80.5	70.1	86.3	39.7	77.9	59.3	59.0	54.4	74.6
FSoD-Net [102]	71.8	88.9	66.9	86.8	90.2	45.5	79.6	48.2	86.9	75.5	67.0
MFPNet [60]	71.2	76.6	83.4	80.6	82.1	44.3	75.6	68.5	85.9	63.9	77.3
MDCT [103]	80.5	**92.5**	85.0	93.5	84.7	53.7	90.2	74.3	79.9	68.2	68.8
CF2PN [104]	67.3	78.3	78.3	76.5	88.4	37.0	71.0	59.9	71.2	51.2	75.6
MSA-YOLO [105]	79.6	94.9	81.3	**94.1**	85.2	55.7	**90.8**	67.3	73.6	73.2	77.4
Ref. [1]	81.6	87.9	91.1	84.9	91.7	55.8	80.7	**78.9**	92.8	82.6	86.6
YOLOv5s	79.8	87.2	86.9	86.2	**92.3**	55.5	83.0	72.6	91.1	83.0	81.6
Our Method	**83.1**	91.2	91.4	89.7	92.2	**58.3**	83.8	76.7	93.8	**87.2**	**88.7**
Method	**mAP**	**GTF**	**HB**	**O**	**S**	**SD**	**ST**	**TC**	**TS**	**V**	**W**
CANet [100]	74.3	83.6	56.0	63.6	81.0	79.8	70.8	88.2	67.6	51.2	89.6
Faster-RCNN [37]	65.1	76.8	46.4	57.2	71.8	68.3	53.8	81.8	59.5	43.1	81.2
TRD [101]	66.8	73.9	49.2	57.8	74.2	61.1	69.8	84.0	58.8	50.5	77.2
FSoD-Net [102]	71.8	77.3	53.6	59.7	78.3	69.9	75.0	91.4	52.3	52.0	90.6
MFPNet [60]	71.2	77.2	62.1	58.8	77.2	76.8	60.3	86.4	64.5	41.5	80.2
MDCT [103]	80.5	**92.9**	68.4	**83.8**	92.9	77.4	**83.0**	92.8	64.7	77.4	83.0
CF2PN [104]	67.3	77.1	56.8	58.7	76.1	70.6	55.5	88.8	50.8	36.9	86.4
MSA-YOLO [105]	79.6	81.7	67.3	67.7	**95.2**	**90.9**	85.5	93.4	63.9	**81.6**	84.2
Ref. [1]	81.6	86.4	**68.7**	67.3	91.7	81.5	80.3	93.0	**77.1**	60.9	91.4
YOLOv5s	79.8	86.4	66.5	67.3	91.8	81.0	80.4	93.2	61.0	60.1	92.0
Our Method	**83.1**	86.6	66.4	68.7	91.3	85.4	81.2	**94.2**	75.0	61.2	**93.5**

The experimental results shown in Table 3 show that compared with CANet, TRD, FSoD-Net, MDCT, CF2PN, MSA-YOLO, and other single-stage and Transformer models based on the fusion attention mechanism, the proposed model has excellent detection performance, and the black bold font represents the best detection method for each target instance type. Particularly for objects such as airports (AT), bridges (B), expressway service areas (ESA), expressway toll stations (ETA), golf courses (GC), tennis courts (TC), and windmills (W). This highlights the effectiveness of our model in detecting remote sensing image objects. To provide a visual representation of our model's feasibility, we

visualize the detection results as depicted in Figure 10. The second column showcases the detection results of the original YOLOv5s model, while the last column reveals the improved detection performance of our proposed model. Our method has higher accuracy and a lower false detection rate compared to the original model, clearly demonstrating its superior accuracy.



(**a**) Original (**b**) Ground Truth (**c**) YOLOv5s (**d**) Our Method

Figure 10. Visualization results comparison of proposed method and YOLOv5s on the DIOR dataset.

4.4.2. Model Performance in the DOTA Datasets

In order to further substantiate the efficacy of our model in terms of detection accuracy, we conducted comparative experiments on the DOTA dataset, employing popular detection algorithms. The results of these experiments are presented in Table 4. Compared to SSD, Faster-RCNN, RADet [106], LO-DET [107], A2S-Det [108], R2CNN [109], and YOLOv5s, our proposed detection method significantly improves detection precision. The black bold

font represents the best detection method for each type of target instance. With a 6.7% improvement over the YOLOv5s model, it demonstrates the excellent detection capabilities of our proposed approach in complex backgrounds and diverse target scales within the DOTA dataset.

Table 4. Performance comparison of different models on the DOTA dataset (The black bold font represents the best detection method for each target instance type).

Method	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
SSD	57.9	32.8	16.1	18.7	0.05	36.9	24.7	81.2	25.1	47.5	11.2	31.5	14.1	9.1	1.0	27.2
FR-O	79.4	77.1	17.7	64.1	35.3	38.0	37.2	89.4	69.6	59.3	50.3	52.9	47.9	47.4	46.3	54.1
RADet	79.5	77.0	48.1	65.8	65.5	74.4	68.9	89.7	**78.1**	75.0	49.9	**64.6**	66.1	**71.6**	**62.2**	69.1
LO-DET	89.2	66.1	31.3	56.0	70.1	71.0	84.3	90.7	75.1	81.3	44.7	59.3	60.0	65.1	48.4	66.7
A2S-DET	89.5	**78.5**	42.8	53.9	**76.4**	74.6	86.0	90.7	83.4	**83.6**	48.6	60.5	63.5	71.3	53.1	70.4
R2CNN	81.0	65.7	35.3	**67.4**	60.0	50.9	55.8	90.7	66.9	72.4	55.1	52.2	55.1	53.4	48.2	60.7
YOLOv5s	88.3	67.6	41.4	51.6	63.8	79.4	80.7	90.8	61.7	62.5	49.3	52.1	77.5	58.8	46.1	64.5
Ours	**92.8**	73.8	**50.0**	57.4	72.5	**84.4**	**87.6**	**94.8**	70.1	74.6	**56.2**	56.7	**81.8**	63.8	50.9	**71.2**

In order to visually showcase the performance of small object detection, we have selectively visualized a collection of representative instances. Figure 11 presents the visualization results of some instances from the DOTA dataset.

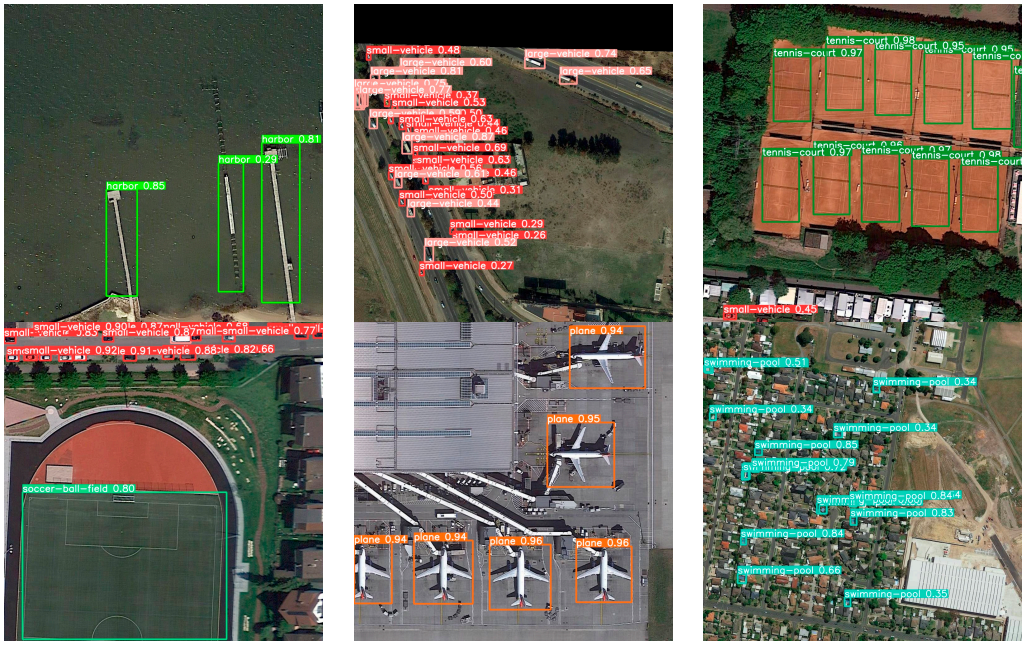


Figure 11. Visualization of proposed method detection results on the DOTA dataset.

4.4.3. Model Performance in the SIMD Datasets

We are dedicated to validating the detection performance of our model on SIMD data for dynamic small objects, while also comparing our proposed methodology with current outstanding one-stage and two-stage algorithms. Detailed experimental results can be found in Table 5. In comparison to the YOLOv5s model, our research approach exhibits a remarkable 3.2% enhancement in detection accuracy, reaching an average precision (mAP) of 81.1%. Furthermore, our method demonstrates excellent detection effectiveness across different target categories. The experiments further substantiate the successful balance we achieved between detection accuracy and speed when detecting dynamic small objects.

Table 5. Detection performance between our model and YOLOv5s on the SIMD dataset.

Method	Precision (%)	Recall (%)	Car	Truck	Van	Long Vehicle	Inference Time	mAP (%)
YOLOv5s	0.721	0.777	0.936	0.832	0.838	0.822	7.2 ms	0.779
Our Model	0.750	0.780	0.941	0.846	0.845	0.853	6.6 ms	0.811

5. Discussion

In this section, we shall validate the proposed model on the DIOR dataset through a series of ablation experiments.

Firstly, in order to validate the performance variations in detection resulting from different embedding positions of the Coordinate Attention (CA) module within the Bottleneck, we evaluate the CA-Bottleneck-a and CA-Bottleneck-b configurations. This evaluation aims to verify the superiority of the network's architectural design. The results of the ablation experiments are presented in Table 6.

Table 6. Performance evaluation of Coordinate Attention module embedded in Bottleneck.

Method	Parameters	Precision (%)	Recall (%)	Inference Time	FPS	mAP (%)
YOLOv5s	7.06 M	0.813	0.751	7.5 ms	109.0	0.798
CA-Bottleneck-a	7.07 M	0.803	0.755	7.9 ms	106.7	0.810
CA-Bottleneck-b	7.07 M	0.846	0.767	7.8 ms	108.5	0.822

Table 6 illustrates the impact of the two embedding methods on the network's detection performance. Compared to commonly used embedding techniques, our approach exhibits significant improvements in terms of accuracy, recall rate, mean average precision (mAP), and detection speed. Moreover, it also leads to a reduction in inference time. By embedding the attention mechanism after two convolutional layers, we are able to better preserve and enhance the positional information of small targets amidst complex backgrounds.

Furthermore, we compare the proposed LADH-Head with the Asymmetric Dual-Head (ADH-Head) in terms of detection speed (FPS, inference time), computational complexity (Parameters, GFLOPs), and detection accuracy (mAP). The results are summarized in Table 7.

Table 7. Performance of the LADH-Head embedded in YOLOv5s.

Method	Parameters	GFLOPs	Inference Time	FPS	mAP (%)
YOLOv5s	7.06 M	16.1	7.5 ms	109.0	0.798
ADH-Head	14.36 M	42.3	8.8 ms	90.7	0.803
LADH-Head	7.10 M	16.4	7.7 ms	106.4	0.820

The results revealed in Table 7 exhibit a remarkable enhancement in detection accuracy for our approach compared to the original YOLOv5s. Although there is a slight decrease in detection speed and computational complexity, our method showcases significant advantages over the ADH-Head in these aspects. With only 49% of the parameters of the ADH-Head, we achieve a 3% increase in detection accuracy. This outcome validates the ability of the LADH-Head to effectively balance the exponential growth in computational complexity brought by the ADH-Head while satisfying the demands for lightweight implementation and detection performance.

Finally, we will utilize the original YOLOv5s model as a baseline for verification and experimental control. In the second experiment, we will replace the C3 module in the YOLOv5s model with the FasterConv module to showcase the benefits of FasterConv in terms of inference speed and detection accuracy. Experiment three involves substituting the C3 module in the Neck with the C3CA module to evaluate the lightweight performance and feature extraction ability of the C3CA module. Moving forward, experiment four

introduces the LADH-Head module to demonstrate its effectiveness in lightweight models and detecting small targets in remote sensing imagery. In the fifth experiment, we will swap out the upsampling method in the Neck with the CARAFE module to investigate the impact of semantic information on detection performance during the upsampling process. The sixth experiment combines the second and third experiments to assess the combined impact of introducing FasterConv and C3CA on detection performance. Experiment seven integrates experiments two, three, and four to exhibit the effectiveness of incorporating the FasterConv, C3CA, and CARAFE modules on detection performance. In the eighth experiment, we incorporate the LADH-Head on top of experiment seven to evaluate the impact of changing the detection head on detection performance. Finally, in experiment nine, we introduce the XIoU loss function based on experiment eight to evaluate the model's detection performance on complex backgrounds with small targets.

The results of these nine ablation experiments provide compelling evidence of the effectiveness of our approach in enhancing the performance of remote sensing image object detection. Please refer to Table 8 for a summary of these ablation experiments. Notably, in experiment two, by employing the FasterConv module to improve the model's inference speed and introducing PConv and PWConv to enhance focus on the center position of the target receptive field, we observed a 1.4% increase in mAP compared to YOLOv5s, a reduction of 0.6 ms in inference speed, and a 9.0% decrease in parameters. Experiment three introduced the C3CA module, resulting in a 2.4% improvement in mAP compared to experiment one by helping the model better focus on positional information and extract target features more efficiently. Experiment four addressed the conflict between regression and classification tasks within the YOLOv5s detection head, resulting in a 2.2% increase in mAP. Moreover, experiment five incorporated the CARAFE module, leading to a 0.8% increase in detection accuracy by enhancing semantic information extraction and fusion capabilities. Experiments six to nine showcased the effectiveness of combining multiple modules, yielding respective improvements in model accuracy (mAP) by 2.1%, 2.5%, 2.6%, and 3.3%. Additionally, these combined experiments reduced inference speed by 1.3 ms and compared to the original model parameters were reduced by 0.22 M, aligning with the requirements of lightweight network design.

Table 8. Ablation experiment results on the DIOR dataset.

FasterConv	C3CA	LADH-Head	CARAFE	XIoU	Parameters	Inference Time	GFLOPs	mAP (%)
F	F	F	F	F	7.06 M	7.5 ms	16.1	0.798
T	F	F	F	F	6.45 M	6.2 ms	14.7	0.812
F	T	F	F	F	7.07 M	7.8 ms	16.1	0.822
F	F	T	F	F	7.10 M	7.7 ms	16.4	0.820
F	F	F	T	F	7.21 M	8.1 ms	16.6	0.806
T	T	F	F	F	6.46 M	7.2 ms	14.7	0.819
T	T	T	F	F	6.49 M	6.9 ms	15.0	0.823
T	T	T	T	F	6.84 M	6.2 ms	15.5	0.824
T	T	T	T	T	6.84 M	6.2 ms	15.5	0.831

Experiments prove that our method can effectively balance the incongruity between model detection performance and lightweight. In the military field, since detection devices such as UAVs do not have devices similar to GPUs that can perform a large number of computations, our method can be well applied to remote sensing satellites or UAV devices in the military detection field to reduce the weight of the model and redundant computations, and to help military detection devices detect remote sensing targets quickly while guaranteeing detection performance. Meanwhile, in the field of urban planning [110] and environmental monitoring, our method can provide a new lightweight network design idea to the existing detection methods, through the combination of PConv and lightweight detection head with YOLO series algorithms to better achieve detection performance on mobile devices. In addition, the provided lightweight design can be well ported to the

YOLOv8 detection method based on the anchorless frame, and the design of the joint attention mechanism can improve the detection performance degradation of the anchorless frame algorithm due to the change in position during the detection of small targets.

6. Conclusions

In this article, we present a faster and lightweight yet effective model for remote sensing image object detection. Our proposed model, which builds upon the YOLOv5s baseline network, introduces the FasterConv, LADH-Head, and C3CA modules. Additionally, we incorporate the CARAFE upsampling method and XIoU loss function to further enhance the model's detection capabilities. The FasterConv module significantly improves both the inference and detection speeds of the model while placing greater emphasis on the central receptive field of the target. This targeted focus leads to notable improvements in detection performance. In place of the original coupled head, we employ a lightweight asymmetrical detection head that bifurcates the network based on task type. This innovative approach utilizes three distinct channels to accomplish the associated tasks, effectively resolving any conflicts between classification and regression objectives. Simultaneously, the C3CA module aids the detection network in emphasizing the positional information of the targets, facilitating the model to extract target features more effectively. This enhancement enables the model to focus more precisely on crucial information pertaining to the target objects, particularly in complex backgrounds. By employing the CARAFE module, our model reconstructs feature points with similar semantic information in a content-aware manner. This process facilitates the aggregation of features within a larger receptive field, ultimately enhancing the model's ability to fuse semantic information. The XIoU loss function plays a crucial role by emphasizing the varying degrees of overlap between targets, thereby improving the robustness of bounding boxes during object detection tasks. Experimental results on the DIOR dataset demonstrate the remarkable performance of our approach. When compared to recent detection methods, our model stands out as a superior choice for achieving optimal detection performance without incurring additional computational costs. The results from our ablation experiments further underscore the effectiveness of each module, showcasing their significant contributions to improving the overall detection performance of the model.

While our approach has yielded advanced results in detecting remote sensing targets, we face the challenge of significant variation in the direction and scale of these targets within the process. Our focus primarily lies on horizontal bounding boxes within the dataset for remote sensing target detection. As a consequence, the detection process is influenced by changes in direction and scale, resulting in potential missed detections or false alarms for overlapping targets. Moreover, the integration of direction and scale detection methods poses difficulties in striking a balance between lightweight design and detection accuracy. In our future endeavors, we will continue to explore the potential of the YOLO model in detecting remote sensing targets that exhibit scale variations. Furthermore, our efforts will be directed towards improving the detection efficiency of the model on real-time devices, enhancing its performance in detecting rotating targets, and achieving a better equilibrium between lightweight design and detection capabilities. Ultimately, our aim is to enhance the real-time detection performance of the model.

Author Contributions: Conceptualization, Z.C.; methodology, J.Z.; software, J.Z.; validation, J.Z.; formal analysis, J.Z.; investigation, J.Z.; resources, Z.C.; data curation, J.Z.; writing—original draft preparation, J.Z.; writing—review and editing, G.Y., Y.W., B.H.; visualization, J.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

Acknowledgments: We wish to extend our sincere gratitude to the editors and reviewers for their invaluable feedback and support, as it has significantly enhanced the technical implementation of our article.

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

References

1. Liu, P.; Wang, Q.; Zhang, H.; Mi, J.; Liu, Y. A Lightweight Object Detection Algorithm for Remote Sensing Images Based on Attention Mechanism and YOLOv5s. *Remote Sens.* **2023**, *15*, 2429. [CrossRef]
2. Roy, P.S.; Behera, M.D.; Srivastav, S.K. Satellite Remote Sensing: Sensors, Applications and Techniques. *Proc. Natl. Acad. Sci. India Sect. A-Phys. Sci.* **2017**, *87*, 465–472. [CrossRef]
3. Liu, X.; He, J.; Yao, Y.; Zhang, J.; Liang, H.; Wang, H.; Hong, Y. Classifying urban land use by integrating remote sensing and social media data. *Int. J. Geogr. Inf. Sci.* **2017**, *31*, 1675–1696. [CrossRef]
4. Zhang, J. Multi-source remote sensing data fusion: Status and trends. *Int. J. Image Data Fusion* **2010**, *1*, 5–24. [CrossRef]
5. Zhang, C.; Su, J.; Ju, Y.; Lam, K.M.; Wang, Q. Efficient Inductive Vision Transformer for Oriented Object Detection in Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–20. [CrossRef]
6. Fu, W.; Zou, W. Review of remote sensing image classification based on deep learning. *Appl. Res. Comput.* **2018**, *35*, 3521–3525.
7. Cheng, G.; Xie, X.; Han, J.; Guo, L.; Xia, G.-S. Remote Sensing Image Scene Classification Meets Deep Learning: Challenges, Methods, Benchmarks, and Opportunities. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 3735–3756. [CrossRef]
8. Alem, A.; Kumar, S. Deep Learning Models Performance Evaluations for Remote Sensed Image Classification. *IEEE Access* **2022**, *10*, 111784–111793. [CrossRef]
9. Lu, F.; Han, M. Hyperspectral remote sensing image classification based on deep extreme learning machine. *J. Dalian Univ. Technol.* **2018**, *58*, 166–173.
10. Guo, C.; Li, K.; Li, H.; Tong, X.; Wang, X. Deep Convolution Neural Network Method for Remote Sensing Image Quality Level Classification. *Geomat. Inf. Sci. Wuhan Univ.* **2022**, *47*, 1279–1286.
11. Gu, Y.; Wang, Y.; Li, Y. A Survey on Deep Learning-Driven Remote Sensing Image Scene Understanding: Scene Classification, Scene Retrieval and Scene-Guided Object Detection. *Appl. Sci.* **2019**, *9*, 2110. [CrossRef]
12. Liu, D.; Han, L.; Han, X. High Spatial Resolution Remote Sensing Image Classification Based on Deep Learning. *Acta Opt. Sin.* **2016**, *36*, 0428001.
13. Sun, X.; Wang, B.; Wang, Z.; Li, H.; Li, H.; Fu, K. Research Progress on Few-Shot Learning for Remote Sensing Image Interpretation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 2387–2402. [CrossRef]
14. Chen, Z.; Wang, Y.; Han, W.; Feng, R.; Chen, J. An Improved Pretraining Strategy-Based Scene Classification With Deep Learning. *IEEE Geosci. Remote Sens. Lett.* **2020**, *17*, 844–848. [CrossRef]
15. Aggarwal, A.; Kumar, V.; Gupta, R. Object Detection Based Approaches in Image Classification: A Brief Overview. In Proceedings of the 2023 IEEE Guwahati Subsection Conference (GCON), Guwahati, India, 23–25 June 2023; pp. 1–6.
16. Liu, B.; Huang, J. Global-Local Attention Mechanism Based Small Object Detection. In Proceedings of the 2023 IEEE 12th Data Driven Control and Learning Systems Conference (DDCLS), Xiangtan, China, 12–14 May 2023; pp. 1439–1443.
17. Shen, T.; Xu, H. Medical Image Segmentation Based on Transformer and HarDNet Structures. *IEEE Access* **2023**, *11*, 16621–16630. [CrossRef]
18. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. *Microsoft COCO: Common Objects in Context*; Springer: Cham, Switzerland, 2014; pp. 740–755.
19. Everingham, M.; Van Gool, L.; Williams, C.K.I.; Winn, J.; Zisserman, A. The Pascal Visual Object Classes (VOC) Challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]
20. Zhu, P.; Wen, L.; Du, D.; Bian, X.; Fan, H.; Hu, Q.; Ling, H. Detection and Tracking Meet Drones Challenge. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 7380–7399. [CrossRef]
21. Zhu, P.; Wen, L.; Du, D.; Bian, X.; Ling, H.; Hu, Q.; Wu, H.; Nie, Q.; Cheng, H.; Liu, C.; et al. VisDrone-VDT2018: The Vision Meets Drone Video Detection and Tracking Challenge Results. In Proceedings of the 15th European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 496–518.
22. Gong, H.; Mu, T.; Li, Q.; Dai, H.; Li, C.; He, Z.; Wang, W.; Han, F.; Tuniyazi, A.; Li, H.; et al. Swin-Transformer-Enabled YOLOv5 with Attention Mechanism for Small Object Detection on Satellite Images. *Remote Sens.* **2022**, *14*, 2861. [CrossRef]
23. Sun, X.; Wang, P.; Yan, Z.; Xu, F.; Wang, R.; Diao, W.; Chen, J.; Li, J.; Feng, Y.; Xu, T.; et al. FAIR1M: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery. *Isprs J. Photogramm. Remote Sens.* **2022**, *184*, 116–130. [CrossRef]
24. You, Y.; Ran, B.; Meng, G.; Li, Z.; Liu, F.; Li, Z. OPD-Net: Prow Detection Based on Feature Enhancement and Improved Regression Model in Optical Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 6121–6137. [CrossRef]
25. Ma, W.; Li, N.; Zhu, H.; Jiao, L.; Tang, X.; Guo, Y.; Hou, B. Feature Split-Merge-Enhancement Network for Remote Sensing Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5616217. [CrossRef]
26. Xiao, J.; Guo, H.; Zhou, J.; Zhao, T.; Yu, Q.; Chen, Y.; Wang, Z. Tiny object detection with context enhancement and feature purification. *Expert Syst. Appl.* **2023**, *211*, 118665. [CrossRef]
27. Cheng, G.; Si, Y.; Hong, H.; Yao, X.; Guo, L. Cross-Scale Feature Fusion for Object Detection in Optical Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 431–435. [CrossRef]
28. Dou, Z.; Gao, K.; Zhang, X.; Wang, H.; Wang, J. Improving Performance and Adaptivity of Anchor-Based Detector Using Differentiable Anchoring With Efficient Target Generation. *IEEE Trans. Image Process.* **2021**, *30*, 712–724. [CrossRef]

29. Zhu, X.; Lyu, S.; Wang, X.; Zhao, Q.; Soc, I.C. TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-captured Scenarios. In Proceedings of the 18th IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 11–17 October 2021; pp. 2778–2788.
30. Liu, X.; Li, Z.; Fu, X.; Yin, Z.; Liu, M.; Yin, L.; Zheng, W. Monitoring House Vacancy Dynamics in The Pearl River Delta Region: A Method Based on NPP-VIIRS Night-Time Light Remote Sensing Images. *Land* **2023**, *12*, 831. [CrossRef]
31. Ju, M.; Niu, B.; Jin, S.; Liu, Z. SuperDet: An Efficient Single-Shot Network for Vehicle Detection in Remote Sensing Images. *Electronics* **2023**, *12*, 1312. [CrossRef]
32. Yan, B.; Wang, D.; Lu, H.; Yang, X. Cooling-Shrinking Attack: Blinding the Tracker with Imperceptible Noises. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 987–996.
33. Ji, L.; Yu-Xiao, N. Method of Insulator Detection Based on Improved Faster R-CNN. In Proceedings of the 2023 6th International Conference on Electronics Technology (ICET), Chengdu, China, 12–15 May 2023; pp. 1127–1133.
34. Zhaowei, C.; Vasconcelos, N. Cascade R-CNN: Delving Into High Quality Object Detection. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6154–6162. [CrossRef]
35. Tsung-Yi, L.; Goyal, P.; Girshick, R.; Kaiming, H.; Dollar, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2999–3007. [CrossRef]
36. Cai, C.; Chen, L.; Zhang, X.; Gao, Z. End-to-End Optimized ROI Image Compression. *IEEE Trans. Image Process.* **2020**, *29*, 3442–3457. [CrossRef]
37. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In Proceedings of the 29th Annual Conference on Neural Information Processing Systems (NIPS), Montreal, BC, Canada, 7–12 December 2015.
38. He, K.; Gkioxari, G.; Dollar, P.; Girshick, R. Mask R-CNN. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 386–397. [CrossRef]
39. Sun, P.; Zhang, R.; Jiang, Y.; Kong, T.; Xu, C.; Zhan, W.; Tomizuka, M.; Yuan, Z.; Luo, P. Sparse R-CNN: An End-to-End Framework for Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**. *Early Access*. [CrossRef]
40. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In Proceedings of the 14th European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 8–16 October 2016; pp. 21–37.
41. Xu, X.; Jiang, Y.; Chen, W.; Huang, Y.; Zhang, Y.; Sun, X. DAMO-YOLO: A Report on Real-Time Object Detection Design. *arXiv* **2022**, arXiv:2211.15444.
42. Adarsh, P.; Rathi, P.; Kumar, M. YOLO v3-Tiny: Object Detection and Recognition using one stage improved model. In Proceedings of the 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 6–7 March 2020; pp. 687–694.
43. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W.; et al. YOLOv6: A Single-Stage Object Detection Framework for Industrial Applications. *arXiv* **2022**, arXiv:2209.02976.
44. Wang, C.-Y.; Bochkovskiy, A.; Mark Liao, H.-Y. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv* **2022**, arXiv:2207.02696.
45. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. YOLOX: Exceeding YOLO Series in 2021. *arXiv* **2021**, arXiv:2107.08430.
46. Chen, Q.; Wang, Y.; Yang, T.; Zhang, X.; Cheng, J.; Sun, J. You Only Look One-level Feature. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 19–25 June 2021; p. 13034.
47. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 27–30 June 2016; pp. 779–788.
48. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6517–6525.
49. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
50. Bochkovskiy, A.; Wang, C.-Y.; Mark Liao, H.-Y. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
51. Lin, T.-Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944.
52. Li, H.; Xiong, P.; An, J.; Wang, L. Pyramid Attention Network for Semantic Segmentation. *arXiv* **2018**, arXiv:1805.10180.
53. Wang, J.; Chen, K.; Xu, R.; Liu, Z.; Loy, C.C.; Lin, D. CARAFE: Content-Aware ReAssembly of FEatures. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV 2019), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3007–3016.
54. Chen, J.; Kao, S.H.; He, H.; Zhuo, W.; Wen, S.; Lee, C.H.; Chan, S.H.G. Run, Don't Walk: Chasing Higher FLOPS for Faster Neural Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023.
55. Felzenszwalb, P.F.; Girshick, R.B.; McAllester, D.; Ramanan, D. Object Detection with Discriminatively Trained Part-Based Models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 1627–1645. [CrossRef] [PubMed]
56. Mikolajczyk, K.; Schmid, C. A performance evaluation of local descriptors. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1615–1630. [CrossRef]

57. Viola, P.; Jones, M. Rapid object detection using a boosted cascade of simple features. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Kauai, HI, USA, 8–14 December 2001; pp. 511–518.
58. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–25 June 2005; pp. 886–893.
59. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
60. Yu, N.; Ren, H.; Deng, T.; Fan, X. Stepwise Locating Bidirectional Pyramid Network for Object Detection in Remote Sensing Imagery. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 1–5. [CrossRef]
61. Uijlings, J.R.R.; van de Sande, K.E.A.; Gevers, T.; Smeulders, A.W.M. Selective Search for Object Recognition. *Int. J. Comput. Vis.* **2013**, *104*, 154–171. [CrossRef]
62. Li, Z.; Wang, Y.; Zhang, N.; Zhang, Y.; Zhao, Z.; Xu, D.; Ben, G.; Gao, Y. Deep Learning-Based Object Detection Techniques for Remote Sensing Images: A Survey. *Remote Sens.* **2022**, *14*, 2385. [CrossRef]
63. Wang, K.; Bai, F.; Li, J.; Liu, Y.; Li, Y. MashFormer: A Novel Multiscale Aware Hybrid Detector for Remote Sensing Object Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 2753–2763. [CrossRef]
64. Ding, J.; Xue, N.; Long, Y.; Xia, G.-S.; Lu, Q.; Soc, I.C. Learning RoI Transformer for Oriented Object Detection in Aerial Images. In Proceedings of the 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 2844–2853.
65. Yang, X.; Yan, J.; Ming, Q.; Wang, W.; Zhang, X.; Tian, Q. Rethinking Rotated Object Detection with Gaussian Wasserstein Distance Loss. In Proceedings of the International Conference on Machine Learning (ICML), Virtual, 18–24 July 2021.
66. Zhao, Q.; Liu, B.; Lyu, S.; Wang, C.; Zhang, H. TPH-YOLOv5++: Boosting Object Detection on Drone-Captured Scenarios with Cross-Layer Asymmetric Transformer. *Remote Sens.* **2023**, *15*, 1687. [CrossRef]
67. Niu, R.; Zhi, X.; Jiang, S.; Gong, J.; Zhang, W.; Yu, L. Aircraft Target Detection in Low Signal-to-Noise Ratio Visible Remote Sensing Images. *Remote Sens.* **2023**, *15*, 1971. [CrossRef]
68. Yan, G.; Chen, Z.; Wang, Y.; Cai, Y.; Shuai, S. LssDet: A Lightweight Deep Learning Detector for SAR Ship Detection in High-Resolution SAR Images. *Remote Sens.* **2022**, *14*, 5148. [CrossRef]
69. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. CenterNet: Keypoint Triplets for Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6568–6577.
70. Law, H.; Deng, J. CornerNet: Detecting Objects as Paired Keypoints. In Proceedings of the 15th European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 765–781.
71. Li, X.; Wang, W.; Hu, X.; Yang, J.; Soc, I.C. Selective Kernel Networks. In Proceedings of the 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 510–519.
72. Luo, Z.; Zhou, C.; Zhang, G.; Lu, S. DETR4D: Direct Multi-View 3D Object Detection with Sparse Attention. *arXiv* **2022**, arXiv:abs/2212.07849.
73. Feng, X.; Han, J.; Yao, X.; Cheng, G. TCANet: Triple Context-Aware Network for Weakly Supervised Object Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 6946–6955. [CrossRef]
74. Oh, B.-D.; Schuler, W. Entropy- and Distance-Based Predictors From GPT-2 Attention Patterns Predict Reading Times Over and Above GPT-2 Surprisal. *arXiv* **2022**, arXiv:abs/2212.11185.
75. Illium, S.; Mueller, R.; Sedlmeier, A.; Popien, C.-L.; Int Speech Commun, A. Visual Transformers for Primates Classification and Covid Detection. In Proceedings of the Interspeech Conference, Brno, Czech Republic, 30 August–3 September 2021; pp. 451–455.
76. Xu, K.; Ba, J.L.; Kiros, R.; Cho, K.; Courville, A.; Salakhutdinov, R.; Zemel, R.S.; Bengio, Y. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 7–9 July 2015; pp. 2048–2057.
77. Mnih, V.; Heess, N.; Graves, A.; Kavukcuoglu, K. Recurrent Models of Visual Attention. In Proceedings of the 28th Conference on Neural Information Processing Systems (NIPS), Montreal, BC, Canada, 8–13 December 2014.
78. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2011–2023. [CrossRef]
79. Jaderberg, M.; Simonyan, K.; Zisserman, A.; Kavukcuoglu, K. Spatial Transformer Networks. In Proceedings of the 29th Annual Conference on Neural Information Processing Systems (NIPS), Montreal, BC, Canada, 7–12 December 2015.
80. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Vedaldi, A. Gather-Excite: Exploiting Feature Context in Convolutional Neural Networks. In Proceedings of the 32nd Conference on Neural Information Processing Systems (NIPS), Montreal, BC, Canada, 2–8 December 2018.
81. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H.; Soc, I.C. Dual Attention Network for Scene Segmentation. In Proceedings of the 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 3141–3149.
82. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the 15th European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.

83. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017.
84. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2021**, arXiv:2010.11929.
85. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the 18th IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 11–17 October 2021; pp. 9992–10002.
86. Yao, Z.; Ai, J.; Li, B.; Zhang, C. Efficient DETR: Improving End-to-End Object Detector with Dense Prior. *arXiv* **2021**, arXiv:abs/2104.01318.
87. Huang, L.; Li, W.; Shen, L.; Fu, H.; Xiao, X.; Xiao, S. YOLOCS: Object Detection based on Dense Channel Compression for Feature Spatial Solidification. *arXiv* **2023**, arXiv:abs/2305.04170.
88. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 19–25 June 2021; pp. 13708–13717.
89. Zhang, X.; Zhou, X.; Lin, M.; Sun, R. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. In Proceedings of the 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 6848–6856.
90. Han, K.; Wang, Y.; Tian, Q.; Guo, J.; Xu, C.; Xu, C. GhostNet: More Features from Cheap Operations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 14–19 June 2020; pp. 1577–1586.
91. Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A ConvNet for the 2020s. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 11966–11976.
92. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 07–09 July 2015; pp. 448–456.
93. Hendrycks, D.; Gimpel, K. Gaussian Error Linear Units (GELUs). *arXiv* **2020**, arXiv:1606.08415.
94. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D.; Assoc Advancement Artificial, I. Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression. In Proceedings of the 34th AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 12993–13000.
95. 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]
96. He, J.; Erfani, S.; Ma, X.; Bailey, J.; Chi, Y.; Hua, X.-S. Alpha-IoU: A Family of Power Intersection over Union Losses for Bounding Box Regression. *arXiv* **2022**, arXiv:2110.13675.
97. 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]
98. Xia, G.-S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A Large-scale Dataset for Object Detection in Aerial Images. In Proceedings of the 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 3974–3983.
99. 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]
100. Li, Y.; Huang, Q.; Pei, X.; Chen, Y.; Jiao, L.; Shang, R. Cross-Layer Attention Network for Small Object Detection in Remote Sensing Imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 2148–2161. [CrossRef]
101. Li, Q.; Chen, Y.; Zeng, Y. Transformer with Transfer CNN for Remote-Sensing-Image Object Detection. *Remote Sens.* **2022**, *14*, 984. [CrossRef]
102. Wang, G.; Zhuang, Y.; Chen, H.; Liu, X.; Zhang, T.; Li, L.; Dong, S.; Sang, Q. FSoD-Net: Full-Scale Object Detection From Optical Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5602918. [CrossRef]
103. Chen, J.; Hong, H.; Song, B.; Guo, J.; Chen, C.; Xu, J. MDCT: Multi-Kernel Dilated Convolution and Transformer for One-Stage Object Detection of Remote Sensing Images. *Remote Sens.* **2023**, *15*, 371. [CrossRef]
104. Huang, W.; Li, G.; Chen, Q.; Ju, M.; Qu, J. CF2PN: A Cross-Scale Feature Fusion Pyramid Network Based Remote Sensing Target Detection. *Remote Sens.* **2021**, *13*, 847. [CrossRef]
105. Su, Z.; Yu, J.; Tan, H.; Wan, X.; Qi, K. MSA-YOLO: A Remote Sensing Object Detection Model Based on Multi-Scale Strip Attention. *Sensors* **2023**, *23*, 6811. [CrossRef] [PubMed]
106. Li, Y.; Huang, Q.; Pei, X.; Jiao, L.; Shang, R. RADet: Refine Feature Pyramid Network and Multi-Layer Attention Network for Arbitrary-Oriented Object Detection of Remote Sensing Images. *Remote Sens.* **2020**, *12*, 389. [CrossRef]
107. Huang, Z.; Li, W.; Xia, X.-G.; Wang, H.; Jie, F.; Tao, R. LO-Det: Lightweight Oriented Object Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5603515. [CrossRef]
108. Xiao, Z.; Wang, K.; Wan, Q.; Tan, X.; Xu, C.; Xia, F. A(2)S-Det: Efficiency Anchor Matching in Aerial Image Oriented Object Detection. *Remote Sens.* **2021**, *13*, 73. [CrossRef]

109. Jiang, Y.; Zhu, X.; Wang, X.; Yang, S.; Li, W.; Wang, H.; Fu, P.; Luo, Z. R2CNN: Rotational Region CNN for Orientation Robust Scene Text Detection. *arXiv* **2017**, arXiv:abs/1706.09579.
110. Tang, W.; He, F.; Bashir, A.K.; Shao, X.; Cheng, Y.; Yu, K. A remote sensing image rotation object detection approach for real-time environmental monitoring. *Sustain. Energy Technol. Assess.* **2023**, *57*, 103270. [CrossRef]