remote sensing

MDPI

Article

Remote Sensing Micro-Object Detection under Global and Local Attention Mechanism

Yuanyuan Li [1], Zhengguo Zhou [1], Guanqiu Qi [2,*], Gang Hu [2], Zhiqin Zhu [1] and Xin Huang [1]

[1] College of Automation, Chongqing University of Posts and Telecommunications, Chongqing 400065, China; liyy@cqupt.edu.cn (Y.L.); s220331165@stu.cqupt.edu.cn (Z.Z.); zhuzq@cqupt.edu.cn (Z.Z.); huangxin@cqupt.edu.cn (X.H.)

[2] Computer Information Systems Department, State University of New York at Buffalo State, Buffalo, NY 14222, USA; hug@buffalostate.edu

* Correspondence: qig@buffalostate.edu

Abstract: With the rapid advancement of technology, satellite and drone technologies have had significant impacts on various fields, creating both opportunities and challenges. In areas like the military, urban planning, and environmental monitoring, the application of remote sensing technology is paramount. However, due to the unique characteristics of remote sensing images, such as high resolution, large-scale scenes, and small, densely packed targets, remote sensing object detection faces numerous technical challenges. Traditional detection methods are inadequate for effectively detecting small targets, rendering the accurate and efficient detection of objects in complex remote sensing images a pressing issue. Current detection techniques fall short in accurately detecting small targets compared to medium and large ones, primarily due to limited feature information, insufficient contextual data, and poor localization capabilities for small targets. In response, we propose an innovative detection method. Unlike previous approaches that often focused solely on either local or contextual information, we introduce a novel Global and Local Attention Mechanism (GAL), providing an in-depth modeling method for input images. Our method integrates fine-grained local feature analysis with global contextual information processing. The local attention concentrates on details and spatial relationships within local windows, enabling the model to recognize intricate details in complex images. Meanwhile, the global attention addresses the entire image's global information, capturing overarching patterns and structures, thus enhancing the model's high-level semantic understanding. Ultimately, a specific mechanism fuses local details with global context, allowing the model to consider both aspects for a more precise and comprehensive interpretation of images. Furthermore, we have developed a multi-head prediction module that leverages semantic information at various scales to capture the multi-scale characteristics of remote sensing targets. Adding decoupled prediction heads aims to improve the accuracy and robustness of target detection. Additionally, we have innovatively designed the Ziou loss function, an advanced loss calculation, to enhance the model's precision in small target localization, thereby boosting its overall performance in small target detection. Experimental results on the Visdrone2019 and DOTA datasets demonstrate that our method significantly surpasses traditional methods in detecting small targets in remote sensing imagery.

Keywords: remote-sensing detection; multi scale feature fusion; attention mechanism; loss function

Citation: Li, Y.; Zhou, Z.; Qi, G.; Hu, G.; Zhu, Z.; Huang, X. Remote Sensing Micro-Object Detection under Global and Local Attention Mechanism. *Remote Sens.* **2024**, *16*, 644. https://doi.org/10.3390/rs16040644

Academic Editors: Richard Gloaguen, Junshi Xia, Puhong Duan, Xueqian Wang and Fulin Luo

Received: 19 December 2023
Revised: 31 January 2024
Accepted: 6 February 2024
Published: 9 February 2024

1. Introduction

In the realm of remote sensing scenes captured by drones or satellites, target detection is a crucial technology. Its primary purpose is to identify and locate specific objects or targets in images or videos taken by drones or satellites. This technology is vital in various applications, such as plant conservation [1], wildlife protection [2], and urban monitoring [3]. In recent years, with the rapid development of deep learning, significant

progress has been made in the development of target detection technology. Presently, the mainstream target detection strategies can be divided into two types: the two-stage strategy represented by the R-CNN series [4] and the one-stage strategy, with YOLO [5] being one of the most popular frameworks. In the two-stage target detection strategy, the initial step uses heuristic methods or region proposal generation techniques to obtain multiple candidate boxes, which are then filtered, classified, and regressed in subsequent steps. In contrast, the one-stage strategy processes in an end-to-end manner, transforming the target detection task into a global regression problem. This global regression strategy can simultaneously assign locations and categories to multiple candidate boxes and more clearly differentiate between targets and background. Single-stage object detectors achieve a balance between real-time metrics and performance, with the YOLO series algorithms gaining attention for their rapid iterative updates. In the current general object detection research field, researchers have widely explored various methods, including expanding model width and depth and applying multi-scale fusion techniques [6]. These techniques are typically applied to natural image processing. However, for high-resolution aerial images in remote sensing scenes, target detection techniques for natural images often cannot be directly applied. This is because aerial images face unique challenges, mainly including two aspects: firstly, due to the high flight altitude of drones or satellites, images often contain a large number of tiny micro-objects, ranging in size from 800×800 to 16×16 pixels, leading to a decrease in Average Precision (AP) when the object size decreases. This is mainly due to the network's insufficient focus on key details and features. Traditional models often face challenges in fully exploiting fine-grained details, thereby impacting the effectiveness of object detection. Consequently, it is imperative to explore more efficient methods for capturing and recognizing these diminutive objects in high-resolution images. Widely employed within the field, multi-scale feature extraction techniques offer a valuable approach to enhance the discernment of intricate details.

Recently, FE-YOLOv5 [7] implemented a feature enhancement strategy to improve model spatial perception and feature representation capabilities. QueryDet [8] introduced an innovative coarse-to-fine cascading sparse query mechanism. LMSN [9] proposed a multi-scale feature fusion and receptive field enhancement module to promote lightweight multi-scale object detection. EFPN [10] developed a feature texture transfer module and a novel foreground–background balance loss function. iS-YOLOv5 [11] optimized the information path of feature fusion and improved the SPP module. Although these techniques retain shallow features for subsequent learning to some extent, the prominence of large targets often dominates due to the simultaneous presence of details of both large and small targets, leading to weakened small target features. Therefore, contextual information becomes crucial in visual recognition tasks, and the attention mechanism has become a popular technique for capturing this information. CANet [12] used a cross-layer attention network to merge deep and shallow features through bidirectional feature fusion. AFPN [13] designed three novel attention modules to enhance the model perception of foreground and contextual information. Typically, these attention modules operate independently at their respective levels, and the interlayer contextual correlation is rarely deeply explored. Existing methods still show inadequacies in balancing local details and global structure in improving the accuracy of small target detection in remote sensing images. Another challenge is the moving characteristics of tiny targets, which reduce the tolerance for bounding box localization. Even a slight offset in the bounding box can lead to a significant drop in Intersection over Union (IoU), causing a large number of false detections. Traditional loss functions have limitations when handling small target localization precision, which may affect the model's overall performance in object detection, especially for small targets. Although many novel loss functions such as GWD [14] and KLD [15] have been proposed recently, they still require further optimization and enhancement in terms of tolerance and computation details for target bounding box offsets in remote sensing target detection scenarios.

In this study, we introduce new modules and structures to further enhance performance and flexibility in remote sensing target detection models. This model, which we named GALDET, primarily aims to provide new solutions for two major issues in remote sensing target detection. Firstly, in GALDET, we introduced our designed CGAL module, which significantly enhances feature extraction capability, enabling the model to more effectively detect and parse fine-grained information carried by tiny targets. It allows local details and global context information to be fused through a specific mechanism, enabling the model to consider both local details and global structure for a more precise and comprehensive understanding of images. Unlike previous research, our GALDET introduces a structure based on Convolutional Neural Networks (CNN) with four decoupled prediction heads to mitigate the impact of target scale variation. This includes three existing decoupled prediction heads and one newly added prediction head, which uses low-level, high-resolution feature maps to increase sensitivity to remote sensing objects. Simultaneously, to further enhance the model's tolerance for bounding box offsets, we propose an improved loss function—the Ziou loss. By optimizing the details of the loss function's calculation, it can enhance the model's precision in small target localization, thereby improving the overall performance of the model in small target detection tasks. Compared to the original YOLO model, our modified model GALDET performs better in processing high-resolution images in remote sensing scenes. In terms of Average Precision (AP) value, it shows an improvement of about 6.3% to 6.5% compared to recent advanced methods.

The main contributions of our research are as follows:

1. To effectively capture the fine-level details embedded in targets, this study introduces an innovative feature fusion module named CGAL, based on Global and Local Attention Mechanisms. This module effectively balances the utilization of local details with global contextual information.

2. Considering the diversity in target scales, we have developed a decoupled detection framework featuring a four-head structure. This is aimed at more effectively tackling the challenges posed by scale variations in remote sensing targets.

3. Addressing the challenges of detecting small objects, we designed a novel loss function that enhances the tolerance scale for bounding box offsets and refines the calculation process, thereby increasing detection precision.

Through these innovative enhancements, our GALDET model has achieved significant performance improvements in the task of target detection in remote sensing scenes. This advancement brings broader application prospects in practical fields such as plant conservation, wildlife protection, and urban monitoring.

The structure of this paper is as follows: Section 2 provides an overview of related work. In Section 3, we offer a comprehensive description of our proposed method. Section 4 presents the experimental results and accompanying discussion. Section 5 delves into a discussion of the limitations of our method and outlines future research directions. Finally, Section 6 provides the concluding remarks for this paper.

2. Related Work

2.1. Data Augmentation

Data augmentation is a critical method for expanding the training sets in machine learning and deep learning, which is essential for improving model performance. This technology is not only widely utilized in the field of object detection [16], but also excels in the domain of image dehazing [17]. Common data augmentation techniques include flipping and mirroring (suitable for targets with no fixed orientation, such as trees), rotation (for targets that may appear from multiple angles, like vehicles or animals), scaling (to address changes in the size of targets, such as pedestrians or vehicles), and cropping (to enhance the model's capability to process partially visible targets, like crowds in surveillance images). Recently, several multi-image data augmentation strategies, such as MixUp [18], CutMix, and Mosaic [19], have been introduced. MixUp trains the network to exhibit simple linear

behavior between training samples through convex combination on pairs of examples and labels, thus performing regularization. CutMix replaces obstructed areas with segments from other images. Mosaic, extending CutMix, blends four different images into one larger image, providing the model with additional scene and object information. In our GALDET model, we incorporated a combination of MixUp and Mosaic to enrich the model's feature learning and enhance its generalization ability.

2.2. Object Detection

The process of object detection and recognition can primarily be divided into two types: one-stage detection and two-stage detection. In two-stage detection, recognition and detection are independent steps; in contrast, one-stage detection performs both tasks concurrently. There are two main types of two-stage object detectors: the sliding window type and the region-based type. The latter includes two phases: generating region proposals and then classifying and refining their locations. This category of detectors includes models like RCNN [20], Fast RCNN [21], and Mask RCNN [4]. Unlike two-stage detectors, one-stage detectors directly perform classification and regression tasks, omitting the region proposal generation step, thus offering higher detection efficiency and requiring less computational power. OverFeat [22] was an early application of one-stage object detectors based on Convolutional Neural Networks (CNN), followed by the emergence of models like YOLO [5] and SSD [23]. Recent one-stage object detection algorithms also include YOLOv8 [24] and YOLOv5 [25], which provide different scale models according to various scene requirements.

Object detection architecture consists of three key components: the Backbone, the Neck, and the Head. The Backbone, such as VGG [26], CSPDarknet53 [27], and Swin Transformer [28], is responsible for feature extraction. In this study, we enhance the CSPDarknet53 [27] with the addition of the CGAL module, optimizing feature extraction by deepening the network and integrating global and local attention strategies. The Neck, serving as a bridge between the Backbone network and Head, further processes feature maps to enhance object detection accuracy. Typical structures include bottom-up and top-down pathways, such as FPN [6] and PANet [29]. The Head part uses feature maps from the Backbone for object localization and classification and can be divided into single-stage and two-stage detectors, like the YOLO and RCNN series. Our proposed multi-decoupled prediction head structure processes features at different scales, enhancing the model's perceptual abilities and capacity to handle complex visual tasks, thereby improving the accuracy of object detection.

2.3. Attention Mechanism

The attention mechanism, inspired by the human visual system, is used to dynamically select and weight relevant data and has been widely applied in object detection and deep learning tasks. This mechanism enhances computational efficiency and performance by focusing on important parts of the input and ignoring irrelevant information. The groundbreaking RAM [30] model demonstrated the effectiveness of attention mechanisms in neural networks, introducing dynamic positioning and information selection methods. The STN [31] further developed the application of attention by using a sub-network to select important regions. SENet [32] improved feature selection precision by introducing attention mechanisms within feature channels. Others, such as Dual self attention [33] and ECANet [34], successfully expanded the attention mechanism, employing different strategies for precise feature selection and weighting. The attention mechanism is divided into global and local attention: global attention helps to capture scene context information, while local attention focuses on details, enabling the model to accurately capture minute features. In our CGAL module, these two forms of attention are seamlessly integrated, allowing the model to capture information on both a global and local scale. Such a fusion not only strengthens the model's comprehension of complex scenarios but also significantly boosts its learning efficiency.

3. The Proposed Method

3.1. Overall Structure

The structural details of our remote sensing object detection method are elaborately presented in Figure 1. This method is built upon a profound algorithmic framework, with its cornerstone being the CGAL feature fusion module, complemented by the neck and the quad-decoupled prediction head structures optimized for multi-scale feature handling. Recognizing the insensitivity of traditional loss functions towards targets in remote sensing object detection, we have incorporated the Ziou loss function in addition to classification and objectiveness losses. In the sections that follow, we will delve into the pivotal components of the algorithm, encompassing the CGAL feature fusion module, the quad-decoupled prediction head structure for multi-dimensional feature processing, and our innovative Ziou loss function.

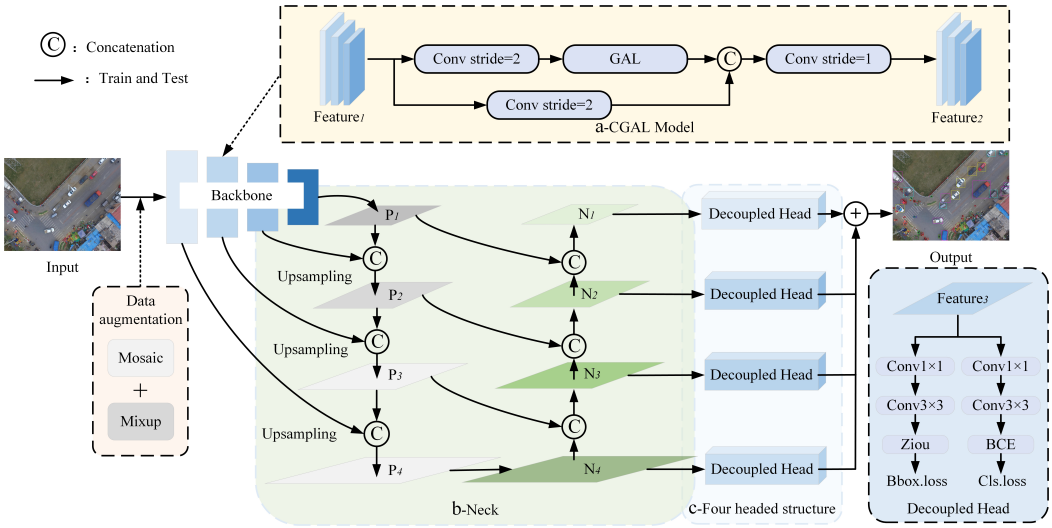

Figure 1. Illustrates the architecture of the proposed method: (**a**) The backbone structure integrated with the proposed CGAL module (yellow part); (**b**) the neck designed to accommodate multi-scale features (green part); (**c**) the four decoupled prediction heads leveraging multi-scale feature maps from the neck (blue part).

3.2. Feature Fusion Module CGAL

In Section 3.2, we delve into the feature fusion module, CGAL. Initially, in Section 3.2.1, we provide a concise recapitulation of the attention mechanism. Subsequently, in Section 3.2.2, we introduce our novel Global and Local Attention (GAL) mechanism. This mechanism, by conducting attention computations at both local and global levels on the input image and applying attention weights to varied windows and key-value pairs, can extract a more profound feature representation. Lastly, in Section 3.2.3, we elucidate how we devise our feature fusion module, CGAL, leveraging GAL as the pivotal component.

3.2.1. Attention Mechanism

Given the query $Q \in R^{N_e \times C}$, the key $K \in R^{N_k \times C}$, and the value $V \in R^{N_\odot \times C}$ as inputs, the attention function transforms each query $Q \in R^{N_q \times C}$ into a weighted sum of values $V \in R^{N_v \times C}$, where the weights are calculated as the normalized dot product between the query and the corresponding key. This can be formally represented in the following concise matrix notation:

$$\text{Attention}\,(Q, K, V) = \text{soft}\max\left(\frac{QK^T}{\sqrt{C}}\right)V$$

By introducing a scalar factor $\sqrt{C}$, we can effectively address concerns associated with weight concentration and the vanishing gradient problem. This innovative approach not

only mitigates these issues but also contributes to enhancing the overall robustness and stability of the model.

3.2.2. Global and Local Attention (GAL) Mechanism

Considering an input $X \in R^{N \times H \times W \times C}$, where N denotes the batch size,H represents the height of the image, W signifies the width of the image, and C indicates the number of channels. To simplify the process, we discuss the Global and Local Attention (GAL) mechanism for a single input where $(N = 1)$, as illustrated in Figure 2.

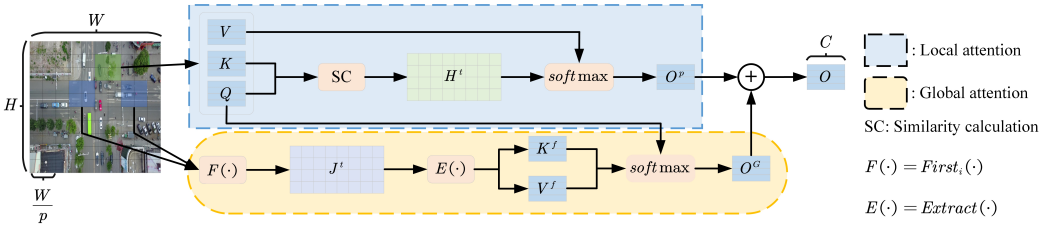

Figure 2. This diagram provides a visual representation of the GAL mechanism, utilizing the directional orientation of arrows and modular narrative to facilitate a more comprehensive understanding of the entire GAL mechanism.

Firstly, we partition the input image, reshaping it from dimension $X \in R^{H \times W \times C}$ to dimension $X^{\alpha} \in R^{p^2 \times \frac{HW}{p^2} \times C}$. This results in $p \times p$ distinct non-overlapping windows, with each window containing $\frac{HW}{p^2}$ feature vectors. Subsequently, through linear transformations, we derive the tensor forms of window-level queries Q, keys K, and values V.

$$Q = X^{\alpha}W^{q} + S^{q}, K = X^{\alpha}W^{k} + S^{k}, V = X^{\alpha}W^{v} + S^{v}$$

The variables $W^{q}, W^{k}, W^{v} \in R^{C \times C}$ represent the linear transformation weights for the queries Q , keys K, and values V, respectively. Meanwhile, $S^{q}, S^{k}, S^{v} \in R^{p^2 \times \frac{HW}{p^2} \times C}$ denote the linear transformation bias values for the queries Q , keys K, and values V, respectively.

Implementing Local Attention Mechanism; After acquiring the tensor representations of the queries Q, keys K, and values V, we apply average pooling to the queries Q and keys K across the spatial dimension, resulting in window-level queries Q^{t} and keys K^{t} . Following this, we determine the similarity score matrix H^{t} using the similarity between the window-level queries Q^{t} and keys K^{t} .

$$H^{t} = Q^{t}\left(K^{t}\right)^{T}$$

Subsequently, we utilize the function softmax to normalize the similarity score matrix H^{t} and multiply it by the value V to obtain the attention weight O^{P} between each query Q and all keys K in the given window.

$$O^{P} = \text{Attention}\left(Q^{t}, K^{t}\right)V = \text{soft max}\left(\frac{Q^{t}\left(K^{t}\right)^{T}}{\sqrt{C}}\right)V$$

where $O \in R^{H \times W \times C}$. This is used to concentrate on the interplay of information within the local window and manage intricate details and spatial relationships.

Implementation of Global Attention.; For each query Q , we choose the top "i" values from the similarity score matrix H^{t} across various local windows, resulting in the correlation matrix $J^{t} \in R^{p^2 \times i}$. This represents the "$i$" keys K most associated with each query Q.

$$J^{t} = \text{First}_{i}\left(H^{t}\right)$$

where first, $_i(\cdot)$ represents choosing the top i values from $J^t \in R^{S^2 \times i}$. Subsequently, we utilize J^t to extract the pertinent sections from the key K and value V tensors.

$$K^f = \text{Extract}(K, J^t), V^f = \text{Extract}(V, J^t)$$

$K^f, V^f \in R^{p^2 \times \frac{iHW}{p^2} \times C}$, In which Extract$(A, B)$ denotes the extraction of sections related to B from A. Upon extracting K^f and V^f, we execute a flattening process, resulting in a size designated as $K^f, V^f \in R^{p^2 \times m \times \frac{iHW}{p^2} \times \frac{C}{m}}$, representing the count of multi-head attentions. Following this, the attention weights are determined between the query tensor Q and the flattened tensors of key K^f and value V^f.

$$O^G = \text{Attention}\left(Q, K^f\right) V^f = \text{soft}\max\left(\frac{Q\left(K^f\right)^T}{\sqrt{C}}\right) V^f$$

By doing so, we manage to restrict the attention mechanism to the key-value pairs most pertinent to each query within $p \times p$ distinct non-overlapping windows, thus fulfilling our objective of harnessing global attention. In the end, we combine the results from both local and global attention and perform a reshaping operation to revert to the original input dimensions, thereby producing the final output.

$$O = O^P + O^G$$

Wherein $O \in R^{H \times W \times C}$. The GAL mechanism adeptly integrates in-depth analysis of local fine-grained features with the processing of global contextual information, offering a profound modeling approach for the input image. The local attention segment of this mechanism conducts meticulous processing on the information within local windows, allowing the model to concentrate on and delve into local nuances and spatial associations, thereby identifying intricate details in complex images. Concurrently, the global attention segment of the GAL mechanism focuses on capturing the macro contextual information of the entire image. This global attention aims to holistically seize the overarching patterns and structures within the image, assisting the model in attaining deeper semantic comprehension based on local details. In the concluding phase, the outputs from these two segments are integrated. This process is not merely an overlay of information; instead, it involves a unique mechanism that harmonizes local intricacies with a global context in a closely intertwined manner. This ensures that the model, during image processing, takes into account both the fine-grained details and the broader structural aspects, leading to a precise and comprehensive understanding of the image.

3.2.3. Feature Fusion Module: CGAL

Leveraging GAL as its core component, we have proposed an innovative feature fusion module known as CGAL. As depicted in Figure 1, the structure of the CGAL module is primarily composed of GAL and convolutional (CONV) blocks. Initially, the module incorporates two CONV blocks with a stride of two, aiming to broaden the receptive field of the network by reducing the spatial dimensions of the feature map. This design ensures that the network comprehensively captures the overarching characteristics of objects, subsequently enhancing their feature extraction capabilities. Following this, the GAL block takes on a central role within the module. Implementing both local and global attention mechanisms, the GAL block delineates the interrelations among input features, producing a refined, weighted feature representation. The module concludes with a CONV block set to a stride of one, a configuration that aids in the preservation of local object features. These paired CONV blocks further process the features refined by the GAL block, deepening the network's capabilities. In essence, the CGAL module, by amplifying the network's depth and receptive field and synergizing local and global attention strategies,

significantly bolsters its feature extraction prowess, enabling the model to delve deeper into the intricate details inherent to remote sensing targets.

3.2.4. Multi-Decoupled Prediction Heads Structure

In this study, we designed a multi-decoupled prediction head structure based on Convolutional Neural Networks (CNNs), as depicted in Figure 1. To address the challenges posed by object scale variations, we devised a structure comprising four decoupled prediction heads, three of which are conventional, with the fourth being a novel addition in our research. This newly introduced prediction head utilizes low-level, high-resolution feature maps, thereby exhibiting enhanced sensitivity to remote sensing objects. Our core rationale is that by amalgamating feature maps of different scales from the Backbone network, we can harness the semantic information extracted across multiple scales, effectively capturing the multi-scale attributes of targets. Such multi-scale characteristics are pivotal for processing remote sensing images, as they assist the model in comprehensively recognizing objects of varying sizes, morphologies, and types within these images. By incorporating additional decoupled prediction heads, our aim is to bolster the accuracy and robustness of object detection. This strategy, both theoretically and empirically, has been proven to optimize object detection performance, enhancing the model's perceptual capabilities across diverse scales, shapes, and categories. In summary, our approach, which leverages multi-scale features combined with enhanced decoupled prediction heads, effectively elevates the precision and stability of remote sensing object detection.

3.2.5. Ziou Loss Function

In remote sensing object detection applications, given that the sizes of remote sensing targets are generally small, the position and size of an object's bounding box play a significant role in the detection outcome. Traditional loss functions might not adequately address the precision required for small object localization, potentially leading to reduced detector performance when dealing with such targets. To tackle this challenge, we optimized the conventional loss function and introduced a novel loss function termed the Ziou loss. This loss function aims to enhance the model's accuracy in localizing small objects by employing a more precise loss calculation, thereby improving the model's detection performance on small targets. Given the predicted bounding box BA and the ground truth bounding box B^{gt}, the definition of the Ziou loss function is as follows:

$$L_{\text{ZIou}} = 1 - IOU + \alpha\nu + \frac{\rho^2\left(b, b^{gt}\right)}{c^2} + \frac{\rho^2\left(h, h^{gt}\right)}{c_h^2} + \frac{\rho^2\left(w, w^{gt}\right)}{c_w^2}$$

Here, b and b^{gt} represent the center points of B and B^{gt}, respectively; $IOU = \left|B \cap B^{gt}\right| / \left|B \cup B^{gt}\right|$; $\alpha = \frac{\nu}{(1-IOU)+\nu}, \nu = \frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h}\right)^2$; h and h^{gt} respectively represent the heights of B and B^{gt}; w and w^{gt} respectively represent the widths of B and B^{gt}; $\rho(\cdot) = \left\|b - b^{gt}\right\|_2$, denotes the Euclidean distance; c represents the diagonal length of the minimal bounding box C that encapsulates both the predicted box B and the target box B^{gt}; c_h and c_w respectively represent the height and width of the minimal bounding box C that encapsulates both the predicted box B and the target box B^{gt}.The visual representation of the specific parameters and their interrelationships mentioned in the formula expression is depicted in Figure 3.

Our proposed method initially adjusts the aspect ratio of the predicted box B using α and ν to gradually converge it to an appropriate range. Upon reaching this range, the edges of the predicted box B are further refined using the orientation loss L_{asp} and distance loss L_{dis} until they approach the correct values. To expedite this convergence process, we also directly regress the Euclidean distance between the center points of the predicted box B and the actual box B^{gt}.

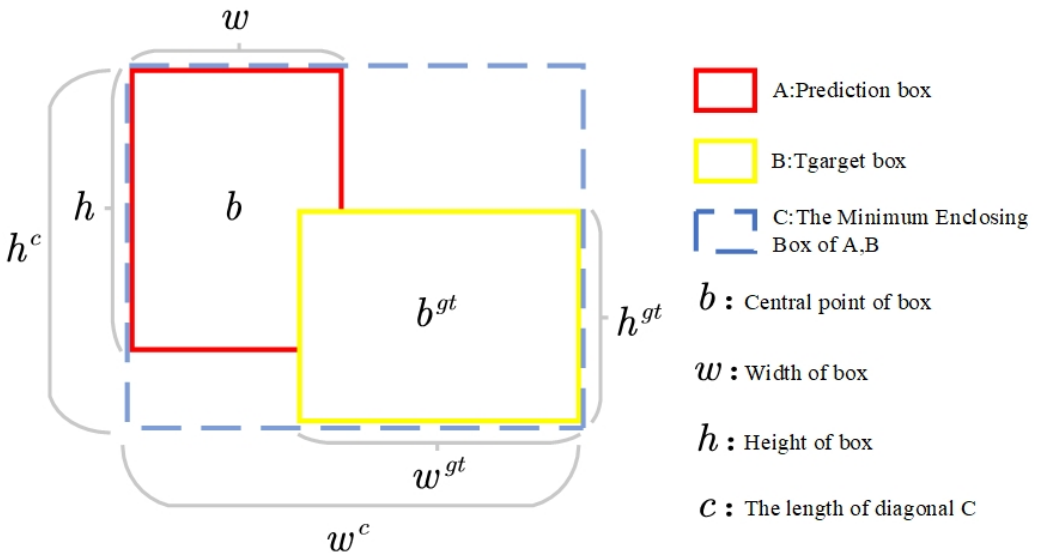

Figure 3. Provides a visual representation of the various parameters.

4. Experiment

This section is divided into four subsections: Dataset and Evaluation Metrics; Implementation Details; Experimental Results and Comparative Analysis; and Ablation Study.

4.1. Dataset and Evaluation Metrics

To evaluate the performance of our proposed method, we selected two challenging and popular benchmark datasets in aerial image detection, namely VisDrone [35] and DOTA [36]. For the VisDrone dataset, we adopted the COCO-style Average Precision (AP) [37] as the evaluation metric and specifically reported the AP values for small, medium, and large-sized objects, with a particular focus on the model's performance in detecting small objects. For the DOTA dataset, following the mainstream research practices [38], we reported the AP for each category and the overall mAP of the model.

VisDrone The VisDrone-2019 dataset consists of 8599 images captured by drone platforms across various locations and altitudes. This dataset is characterized by objects that are small in size, densely distributed, and may be partially occluded. Moreover, different scenes demonstrate variations in illumination and perspective effects. The dataset encompasses over 540k annotated bounding boxes, categorized into ten predefined classes: pedestrian, person, bicycle, car, van, truck, tricycle, awning tricycle, bus, and motorcycle. The training and validation subsets are made up of 6471 and 548 images, respectively, sourced from different locations yet under analogous environmental settings.

DOTA The DOTA-1.0 dataset is specifically designed for object detection in aerial images and is extensively utilized in research related to remote sensing image object detection. Distinctive for its high resolution and the presence of a multitude of densely packed small objects, it presents significant challenges for remote sensing image object detection. The dataset comprises 2806 large-scale aerial images (approximately 4000 $\times$ 4000 pixels) with a total of about 188,282 annotated object instances. These instances are classified into 15 distinct categories, including airplanes, ships, storage tanks, baseball fields, basketball courts, ground track facilities, cars, helicopters, sports fields, harbors, bridges, roundabouts, swimming pools, tennis courts, and runways.A visual display of the Visdrone2019-DET and DOTA-1.0 data set samples is shown in Figure 4.

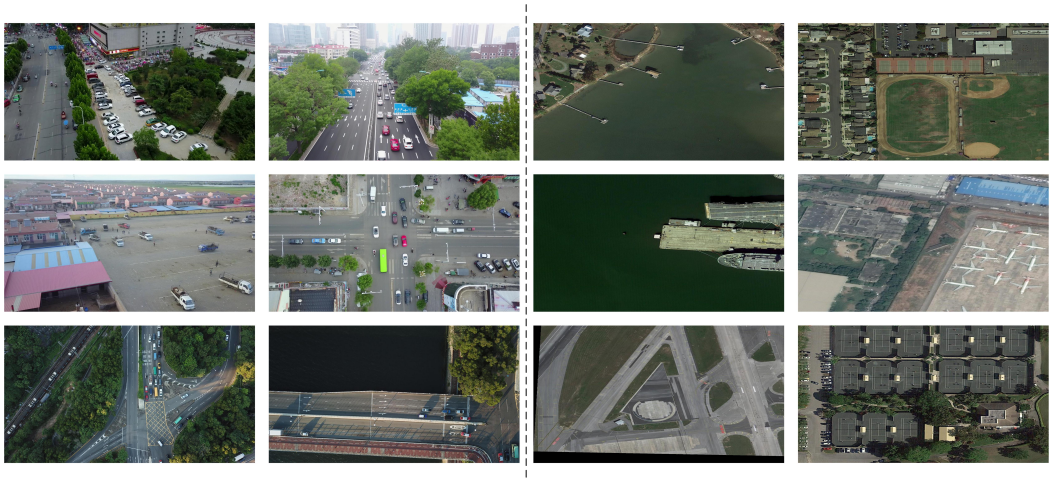

Figure 4. Vividly illustrates the three central challenges confronted in the DOTA and VisDrone datasets: diverse object sizes, high object density, and occlusions. Both datasets encapsulate comprehensive geospatial data, encompassing features such as topography, terrain variations, and architectural structures, offering a substantial foundation for the exploration of geospatial information's role in object detection. (Examples on the left are from the VisDrone dataset, while those on the right are from the DOTA dataset, separated by a black dashed line in the middle).

4.2. Implementation Details

We implemented the GALDET algorithm on the Pytorch 2.0.1 platform. For the Visdrone2019-DET dataset, the input image size was adjusted to 640 × 640 for both training and inference phases. Training was conducted over 300 epochs using the Stochastic Gradient Descent (SGD) approach. The set weight decay and momentum were 0.005 and 0.8, respectively. The batch size was kept at eight with an initial learning rate of 0.1. All model training and testing were executed on an NVIDIA RTX4090 GPU. The initial phase of the training began with 5 warm-up epochs, during which the learning rate gradually increased from 0 to 0.005. For the DOTA 1.0 dataset, each image was pre-processed into 1024 × 1024 pixel patches with an overlap of 200 pixels between patches. To verify the algorithm's efficacy across different model scales, we reduced the overall model size. During training and inference, the input image size was adjusted to 1024 × 1024 while other training parameters were kept consistent. Ultimately, considering the rich annotations of the Visdrone2019-DET and DOTA-1.0 datasets, we set the masaic and mixup coefficients to 0.8 and 0.445, respectively.

4.3. Experimental Results and Comparative Analysis

To evaluate the performance of our method in detecting tiny objects in remote sensing, we compared it with several state-of-the-art detection methods in Visdrone-2019, including CNN-based and YOLO series-based approaches. Given that many remote sensing object detection methods have not released their source codes, to avoid bias in the model retraining process, we chose to directly reference the evaluation results provided in the related literature for comparison, a common practice in the field of remote sensing tiny object detection. To ensure the accuracy and validity of the comparison results, we made sure our experimental setup was consistent with the settings of the methods being compared, thus ensuring the experiments started from the same baseline. Simultaneously, we adjusted the scale of our algorithm to match that of the compared algorithms. For instance, in the Visdrone-2019 comparison, our algorithm maintained the same scale as EdgeYOLO [39], YOLO5-X [25], etc., with a network depth setting of 1.33 and a network width setting of 1.25. Similarly, in the DOTA-1.0 comparison, we matched the scale of PPYOLOE-R-s [40], YOLOV8-S [24], etc., with a network depth setting of 0.33 and a network width setting of 0.50. The evaluation results for different methods on Visdrone, DOTA, and are listed in Tables 1 and 2, respectively, with the best-performing values highlighted in bold.

Figures 5 and 6 show the performance of different remote sensing tiny object detection methods on the AP75 and AP50 metrics for a better comparison. As can be seen from the above tables and charts, the method proposed in this paper has a stronger competitive performance compared to other methods.

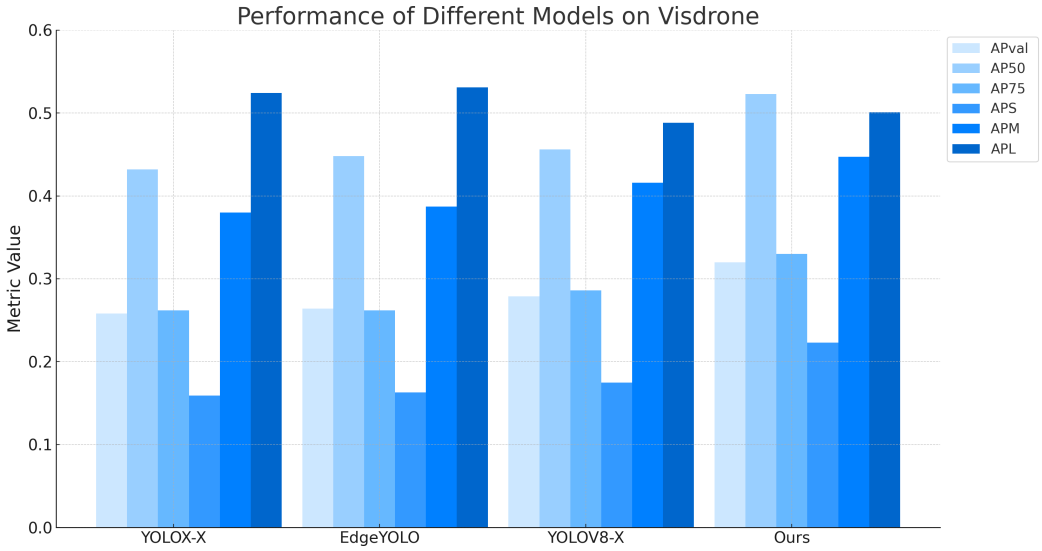

Figure 5. presents a performance comparison of various remote sensing object detection algorithms on the VisDrone2019-DET-val dataset.

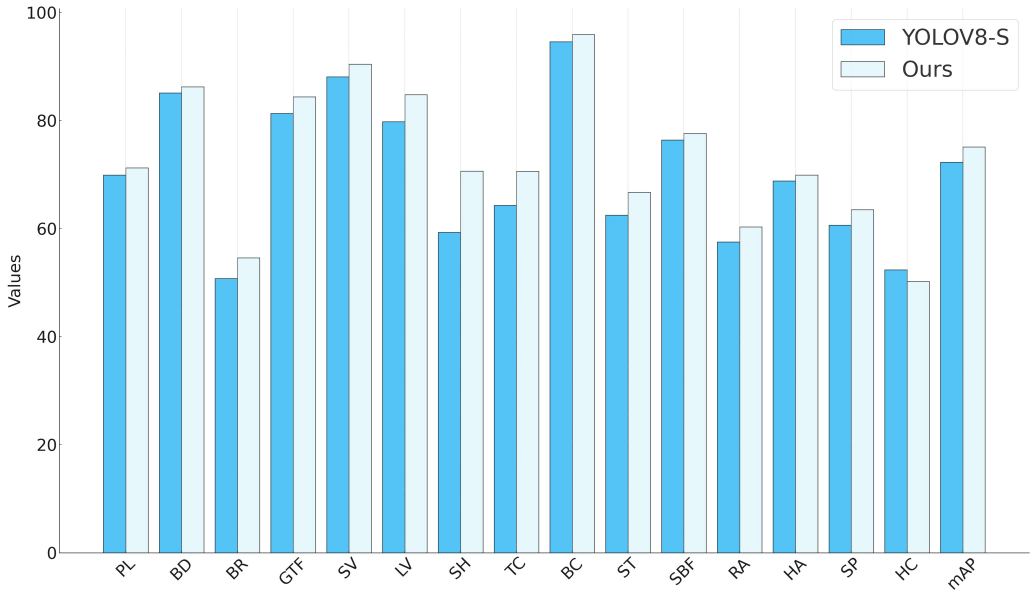

Figure 6. We have meticulously reported the improvements between each category and the overall mAP. Among them, "plane" is abbreviated as "PL", "baseball diamond" as "BD", "bridge" as "BR", "ground track field" as "GTF", "small vehicle" as "SV", "large vehicle" as "LV", "ship" as "SH", "tennis court" as "TC", "basketball court" as "BC", "storage tank" as "ST", "soccer ball field" as "SBF", "roundabout" as "RA", "harbor" as "HA", "swimming pool" as "SP", and "helicopter" as "HC".

In the Visdrone-2019 benchmark test, the method proposed in this study achieved 32.00% and 52.30% on the AP75 and AP50 metrics, respectively, which was significantly higher than other reference methods by at least 4.1% and 6.7%. Compared with CascadeRCNN+ResNeXt [41] and TOOD+SF+SAHI+FI+PO [42] in remote sensing tiny object detection, our method showed superiority in all scenarios, especially in the AP50 metric representing overall performance. Additionally, compared with the latest EdgeYOLO [39],

which incorporates enhanced data augmentation strategies, our method showed noticeable improvements in the APS and APM metrics, representing the effectiveness in tiny object detection. Figure 7 shows some visual results for the Visdrone test set.

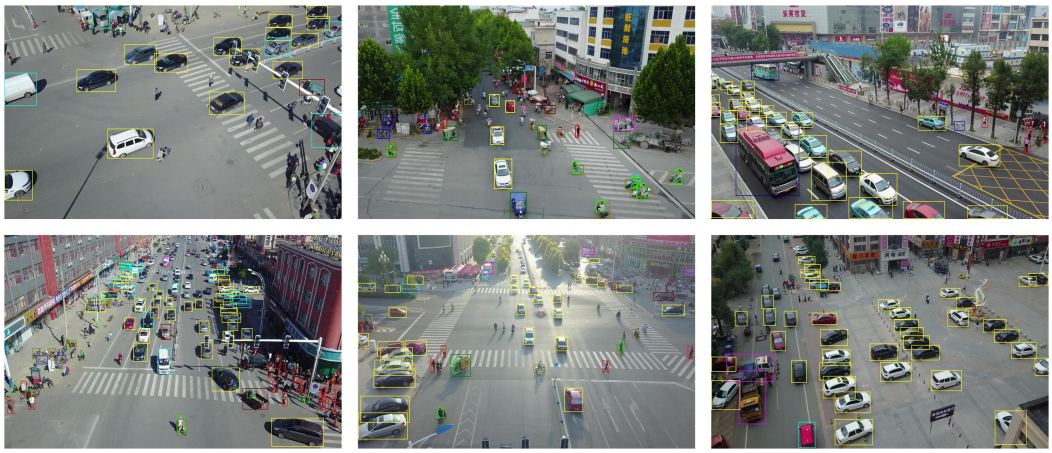

Figure 7. Some visualization results from our GALDET on Visdrone-2019 testset, different category use bounding boxes with different color (The yellow box represents cars, the red box represents pedestrians, the orange box represents non-motorized vehicles, and the blue box represents trucks). The performance is good for the localization of tiny objects, dense objects and objects blurred by motion.

Conducting a direct comparison between our method and commonly used baseline approaches in remote sensing object detection holds significant value. Such a comparative analysis serves to elucidate the importance and effectiveness of the proposed improvements. To provide a comprehensive evaluation of our approach, we will now juxtapose it with two prominent baseline methods that are frequently employed in the field of remote sensing object detection.Incorporating the methodologies of CenterNet-Hourglass104 [43] and EfficientDet±D0 [44] into our comparative assessment, Table 1 unmistakably reveals significant advancements achieved by our approach in comparison to these benchmark methods.

Table 1. Comparison of different object detectors on VisDrone2019-DET-val. In the table, "TOOD+" stands for "TOOD+SF+SAHI+FI+POTOOD+SF+SAHI+FI+PO", and "FasterRCNN+" represents "FasterRCNN+ResNeXtTOOD+SF+SAHI+FI+PO".

Model	Size	APval	AP50	AP75	APS	APM	APL
YOLO5-X [25]	640 × 640	22.60%	38.60%	-	-	-	-
TOOD+ [42]	640 × 640	-	43.50%	-	-	-	-
CascadeRCNN+ResNeXt [41]	640 × 640	24.40%	41.20%	-	-	-	-
FasterRCNN+ [42]	640 × 640	23.60%	37.40%	-	-	-	-
CenterNet-Hourglass104 [43]	640 × 640	25.60%	50.30%	22.22%	-	-	-
EfficientDet±D0 [44]	640 × 640	20.80%	37.10%	20.60%	-	-	-
M2S [45]	640 × 640	-	16.10%	29.70%	-	-	-
YOLOX-X [46]	640 × 640	25.80%	43.20%	26.20%	15.90%	38.00%	52.40%
EdgeYOLO [39]	640 × 640	26.40%	44.80%	26.20%	16.40%	38.70%	53.10%
YOLOV8-X	640 × 640	27.90%	45.60%	17.50%	28.60%	48.80%	41.60%
Ours	640 × 640	32.00%	52.30%	33.00%	22.30%	44.70%	50.10%

In the DOTA-1.0 benchmark test, our reported experimental results are based on a single-scale model and were fairly compared with previous research methods. Using the YOLOV8-S mini version as a benchmark, we validated the effectiveness of our method across different size models with a 2.81% mAP improvement. Even after scaling down the algorithm, our method still maintained an advantage in precision metrics. Moreover, compared with the latest Anchor-free architecture PPYOLOE-R-s [40], our method achieved

significant improvements in recognizing and detecting medium- and small-sized targets. Also, in comparison with other methods, our approach demonstrated strong competitiveness. Figure 8 shows some visual results on the DOTA test set.

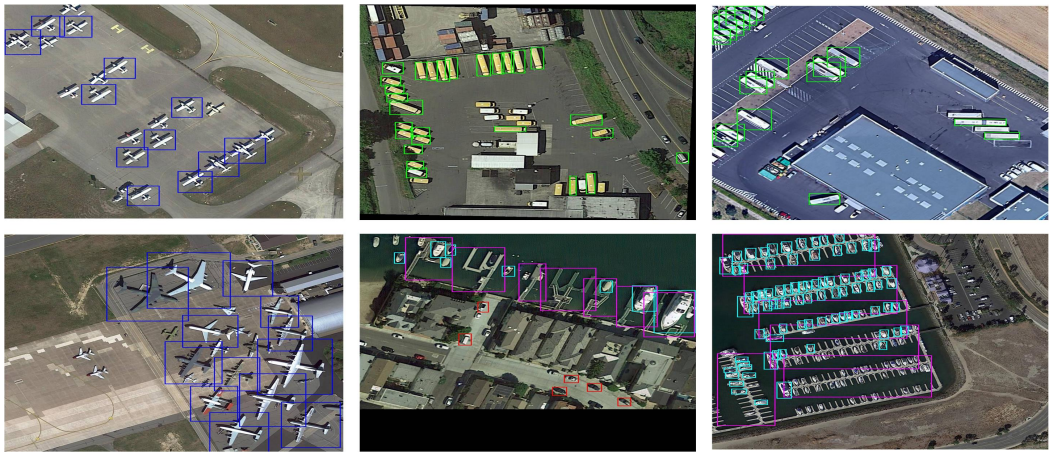

Figure 8. Some visualization results of our GALDET on the DOTA-1.0 test set challenge, similar to before, using different colored bounding boxes for different categories (The blue box represents an airplane, the purple box represents a dock, the green box represents a car, and the light blue box represents a ship). We demonstrated the ability of our model to locate small objects and dense objects via remote sensing.

Table 2. Comparison with state-of-the-art methods on DOTA dataset. We have listed the AP values of the six most representative object categories and the overall mAP value in the table.

Methods	Backbone	PL	BD	BR	GTF	SV	LV	mAP
PPYOLOE-R-s [40]	CRN-s	88.80	79.24	45.92	66.88	80.41	82.95	73.82
DRN [47]	H-104	89.71	82.34	47.22	64.10	76.22	74.43	73.23
O2-DNet [48]	H-104	89.31	82.14	47.33	61.21	71.32	74.03	71.04
DAL [49]	R-101-FPN	88.61	79.69	46.27	70.37	65.89	76.10	71.78
SCRDet [50]	R-101-FPN	89.98	80.65	52.09	68.36	68.36	60.32	72.61
S2A-Net [38]	R-50-FPN	89.11	82.84	48.37	71.11	78.11	78.39	74.12
YOLOV8-S	CSP	69.91	85.12	50.75	81.34	88.11	79.81	72.31
Ours	CSP	71.23	86.25	54.58	84.41	90.44	84.81	75.12

4.4. Ablation Study

To thoroughly validate the effectiveness of the key components designed in our study, including the CGAL module, the four decoupled prediction head structure, and the Ziou loss function, we conducted a series of ablation experiments. In these experiments, we used the standard YOLOV8-X [24] as the baseline model and incrementally added these different components to assess their impact. Considering that different comparative methods utilized various hardware platforms for testing computational complexity and time cost, we detailed the computational costs and timing information of our algorithm in the ablation experiments. When calculating frames per second (FPS), we took into account all time components, including pre-processing and post-processing. Specifically, we mainly compared the performance of the following four models:

-Baseline model: The baseline model solely uses YOLOV8-X for remote sensing small object detection.

-Baseline model +CGAL: Based on the baseline model, this configuration integrates the CGAL module proposed in this paper without leveraging pre-trained weights.

-Baseline model +CGAL+ Four Decoupled Prediction Heads: Building on the previous configuration, this one further incorporates the four decoupled prediction head structure, aiming to enhance the model's perception capabilities across various scales, shapes, and object categories.

-Complete Model: This refers to Baseline model +CGAL+Four Decoupled Prediction Heads+Ziou. By comparing baseline model with Baseline model+ CGAL, we can validate the effectiveness of the CGAL module. Comparing Baseline model+ CGAL with Baseline model +CGAL+Four Decoupled Prediction Heads allows us to verify the contribution of the four decoupled prediction head structure. Lastly, contrasting Baseline model+CGAL+Four Decoupled Prediction Heads with the complete model further confirms the role of the Ziou loss in remote sensing small object detection. Table 3 presents detailed performance metrics for each model configuration. The results clearly show that each add-on helps improve the performance of the model. It is worth noting that the introduction of the CGAL module and the four-head structure led to significant improvements in multi-scale detection, as shown in Figure 9, which shows the detection results of different models in the ablation study through bar charts. Figure 10 visually compares the detection results of different model configurations in ablation experiments. Some salient observations include: (1) Upon integrating the CGAL module, there's a notable enhancement in the detection capability for small objects; (2) With the incorporation of the four decoupled prediction head structure, the recognition accuracy for small objects shows substantial improvement compared to the ground truth; (3) After employing the Ziou loss during training, the positioning of the detection boxes becomes more precise, yielding a visually more coherent result.

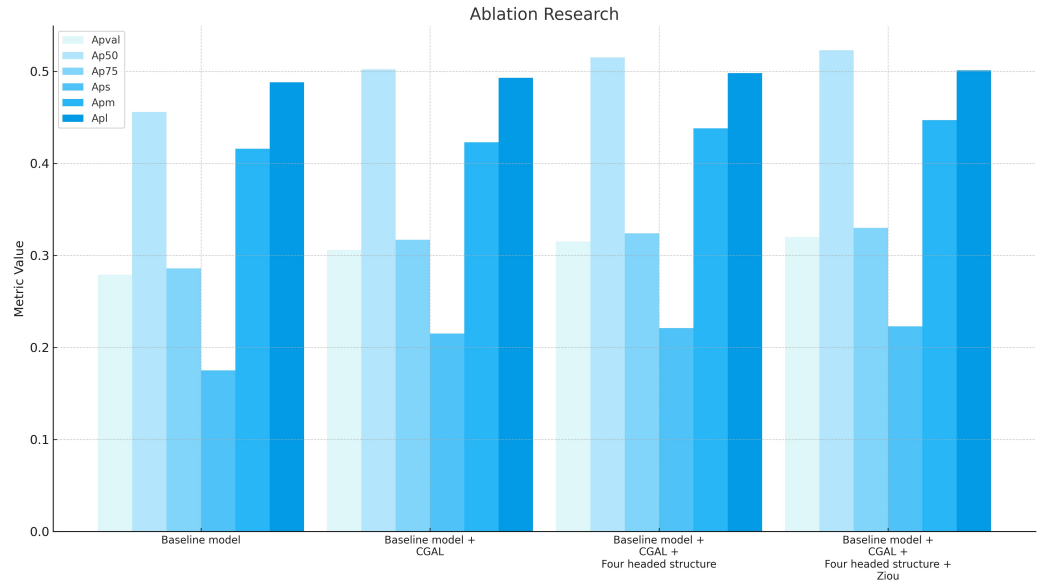

Figure 9. Bar chart comparing the detection performance of different models in the ablation study.

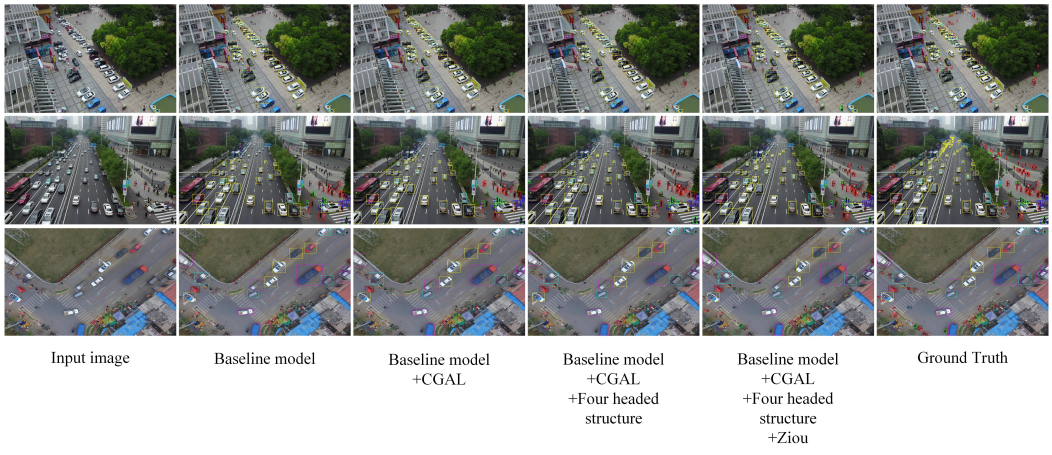

Figure 10. Visual comparison of detection results obtained by different models in the ablation study.

Table 3. Based on VisDrone2021-DET val and 640 × 640 resolution as input for ablation research.

Setting	Apval	Ap50	Ap75	Aps	Apm	Apl	GFLOPs	FPS
baseline	27.90%	45.60%	28.60%	17.5%	41.6%	48.8%	365.3	60.97
+CGAL	30.60%	50.20%	31.7%	21.5%	42.3%	49.3%	386.6	46.04
+Four headed structure	31.5%	51.5%	32.4%	22.1%	43.8%	49.8%	392.4	43.90
+Ziou	32.00%	52.30%	33.00%	22.3%	44.7%	50.1%	394.7	44.39

Analysis

The computational model proposed in this study is particularly applicable to the fields of drone aerial surveillance and remote sensing satellite imagery. In these application scenarios, the demands for real-time performance and computational complexity are exceptionally stringent. Our model significantly improves accuracy (with a 6.7% increase in the AP50 value) by sacrificing a minimal amount of real-time performance (a decrease of 16.58 frames per second). This improvement is crucial for the rapid and accurate processing of voluminous aerial image data.

In the context of drone surveillance systems, the balance between real-time performance and accuracy is vital for effective task execution. For instance, in search and rescue operations or border surveillance, the ability to quickly and accurately identify ground objects can be pivotal to the success of the mission. Our model, while maintaining high accuracy, incurs an additional computational load of only 29.4 GFLOPs. This is a manageable increase for drone systems, which often have limited computational resources.

Similarly, in remote sensing satellite applications, the enhanced accuracy is essential for extracting valuable information from extensive aerial images captured at high altitudes. Whether in environmental monitoring, agricultural planning, or urban development assessment, the improved target detection capability significantly enhances the value and effectiveness of the data. Despite the increased computational demand, this is within acceptable limits considering the high computational capabilities of satellite platforms.

5. Limitation and Future Work

Our study, while yielding promising results, is not without its limitations. Firstly, the current version of our GALDET system exhibits certain constraints in terms of detection accuracy, which we acknowledge requires further refinement. Additionally, the real-time processing speed, while adequate for many applications, may benefit from optimization to enhance overall efficiency. Another limitation pertains to the dataset diversity, as our research predominantly focused on a specific type of object detection challenge. Expanding our dataset to encompass a broader array of scenarios and object types is a potential avenue for future research.

Looking ahead, we envision several avenues for enhancing the capabilities of our GALDET system. Firstly, we will dedicate our efforts to fine-tuning our detection algorithms to achieve higher precision and robustness, thus addressing the limitations in detection accuracy. Additionally, optimizing the system's real-time performance will remain a priority, allowing for broader practical applications. We also anticipate diversifying our dataset to encompass a wider range of image detection challenges, fostering a more comprehensive understanding of GALDET's potential. Moreover, exploring novel machine learning techniques and integrating cutting-edge algorithms, such as advanced image dehazing [51] and atmospheric correction mechanisms [52], will be integral to our future research endeavors. Through these concerted efforts, we aspire to further solidify GALDET's position as a state-of-the-art solution for diminutive object detection in the field of remote sensing.

6. Conclusions

In this research paper, we introduce GALDET, a novel method for target detection in remote sensing images. Distinctively, GALDET harnesses GAL to extract salient im-

age features. The methodology is built upon three main pillars: a feature fusion module, conceptualized from GAL; a multi-decoupled prediction header grounded in CNN principles; and the Ziou loss function, meticulously crafted to bolster the model's proficiency in pinpointing smaller targets. Empirical outcomes underscore the indispensable nature of our architectural components. Notably, when juxtaposed with conventional strategies, GALDET shines, evidencing exceptional prowess in discerning minuscule targets on both the Visdrone2019 and DOTA datasets.

Author Contributions: Methodology, Y.L.; Software, Z.Z. (Zhengguo Zhou); Validation, G.H.; Formal analysis, G.Q.; Investigation, Z.Z. (Zhiqin Zhu); Resources, X.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research is jointly sponsored by National Natural Science Foundation of China (82205049, 62276037), Natural Science Foundation of Chongqing (cstc2020jcyj-msxmX0259, cstc2021jcyj-bsh0199), Special key project of Chongqing technology innovation and application development: CSTB2022TIAD-KPX0039, Basic Research and Frontier Exploration Project of Yuzhong District, Chongqing, Grant/Award Number: 20210164.

Data Availability Statement: Data are contained within the article.

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

References

1. Hird, J.N.; Montaghi, A.; McDermid, G.J.; Kariyeva, J.; Moorman, B.J.; Nielsen, S.E.; McIntosh, A.C. Use of unmanned aerial vehicles for monitoring recovery of forest vegetation on petroleum well sites. *Remote Sens.* **2017**, *9*, 413. [CrossRef]
2. Kellenberger, B.; Volpi, M.; Tuia, D. Fast animal detection in UAV images using convolutional neural networks. In Proceedings of the 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Worth, TX, USA, 23–28 July 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 866–869.
3. Audebert, N.; Le Saux, B.; Lefèvre, S. Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks. *ISPRS J. Photogramm. Remote Sens.* **2018**, *140*, 20–32. [CrossRef]
4. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
5. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
6. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
7. Wang, M.; Yang, W.; Wang, L.; Chen, D.; Wei, F.; KeZiErBieKe, H.; Liao, Y. FE-YOLOv5: Feature enhancement network based on YOLOv5 for small object detection. *J. Vis. Commun. Image Represent.* **2023**, *90*, 103752. [CrossRef]
8. Yang, C.; Huang, Z.; Wang, N. QueryDet: Cascaded sparse query for accelerating high-resolution small object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 13668–13677.
9. Li, L.; Li, B.; Zhou, H. Lightweight multi-scale network for small object detection. *PeerJ Comput. Sci.* **2022**, *8*, e1145. [CrossRef] [PubMed]
10. Deng, C.; Wang, M.; Liu, L.; Liu, Y.; Jiang, Y. Extended feature pyramid network for small object detection. *IEEE Trans. Multimed.* **2021**, *24*, 1968–1979. [CrossRef]
11. Mahaur, B.; Mishra, K. Small-object detection based on YOLOv5 in autonomous driving systems. *Pattern Recognit. Lett.* **2023**, *168*, 115–122. [CrossRef]
12. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
13. Min, K.; Lee, G.H.; Lee, S.W. Attentional feature pyramid network for small object detection. *Neural Netw.* **2022**, *155*, 439–450. [CrossRef] [PubMed]
14. 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, PMLR, Virtual, 18–24 July 2021; pp. 11830–11841.
15. Yang, X.; Yang, X.; Yang, J.; Ming, Q.; Wang, W.; Tian, Q.; Yan, J. Learning high-precision bounding box for rotated object detection via kullback-leibler divergence. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 18381–18394.
16. Qi, G.; Zhang, Y.; Wang, K.; Mazur, N.; Liu, Y.; Malaviya, D. Small object detection method based on adaptive spatial parallel convolution and fast multi-scale fusion. *Remote Sens.* **2022**, *14*, 420. [CrossRef]
17. Zhu, Z.; Wei, H.; Hu, G.; Li, Y.; Qi, G.; Mazur, N. A novel fast single image dehazing algorithm based on artificial multiexposure image fusion. *IEEE Trans. Instrum. Meas.* **2020**, *70*, 1–23. [CrossRef]

18. Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. mixup: Beyond empirical risk minimization. *arXiv* **2017**, arXiv:1710.09412.
19. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
20. Zhang, Y.; Ming, Y.; Zhang, R. Object detection and tracking based on recurrent neural networks. In Proceedings of the 2018 14th IEEE International Conference on Signal Processing (ICSP), Beijing, China, 12–16 August 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 338–343.
21. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
22. Sermanet, P.; Eigen, D.; Zhang, X.; Mathieu, M.; Fergus, R.; LeCun, Y. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv* **2013**, arXiv:1312.6229.
23. 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 Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part I 14; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
24. Solawetz, J. What is YOLOv8? The Ultimate Guide. 2023. Available online: https://blog.roboflow.com/whats-new-in-yolov8/ (accessed on 18 December 2023).
25. Jocher, G.; Stoken, A.; Borovec, J.; Chaurasia, A.; Changyu, L.; Hogan, A.; Hajek, J.; Diaconu, L.; Kwon, Y.; Defretin, Y.; et al. *Ultralytics/yolov5: v5. 0-YOLOv5-P6 1280 Models, AWS, Supervise. ly and YouTube Integrations*; Zenodo: Geneva, Switzerland, 2021.
26. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
27. Wang, C.Y.; Liao, H.Y.M.; Wu, Y.H.; Chen, P.Y.; Hsieh, J.W.; Yeh, I.H. CSPNet: A new backbone that can enhance learning capability of CNN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 390–391.
28. 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 IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 10012–10022.
29. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8759–8768.
30. Mnih, V.; Heess, N.; Graves, A.; Kavukcuoglu, K. Recurrent models of visual attention. *arXiv* **2014**, arXiv:1406.6247.
31. Jaderberg, M.; Simonyan, K.; Zisserman, A.; Kavukcuoglu, K. Spatial transformer networks. *arXiv* **2015**, arXiv:1506.02025.
32. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
33. Zhu, Z.; Luo, Y.; Qi, G.; Meng, J.; Li, Y.; Mazur, N. Remote sensing image defogging networks based on dual self-attention boost residual octave convolution. *Remote Sens.* **2021**, *13*, 3104. [CrossRef]
34. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 11534–11542.
35. Zhu, P.; Wen, L.; Du, D.; Bian, X.; Ling, H.; Hu, Q.; Nie, Q.; Cheng, H.; Liu, C.; Liu, X.; et al. Visdrone-det2018: The vision meets drone object detection in image challenge results. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Munich, Germany, 8–14 September 2018.
36. 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 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3974–3983.
37. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Proceedings, Part V 13; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755.
38. Han, J.; Ding, J.; Li, J.; Xia, G.S. Align deep features for oriented object detection. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–11. [CrossRef]
39. Liang, S.; Wu, H.; Zhen, L.; Hua, Q.; Garg, S.; Kaddoum, G.; Hassan, M.M.; Yu, K. Edge YOLO: Real-time intelligent object detection system based on edge-cloud cooperation in autonomous vehicles. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 25345–25360. [CrossRef]
40. Wang, X.; Wang, G.; Dang, Q.; Liu, Y.; Hu, X.; Yu, D. PP-YOLOE-R: An Efficient Anchor-Free Rotated Object Detector. *arXiv* **2022**, arXiv:2211.02386.
41. Tang, W.; Sun, J.; Wang, G. Horizontal Feature Pyramid Network for Object Detection in UAV Images. In Proceedings of the 2021 China Automation Congress (CAC), Beijing, China, 22–24 October 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 7746–7750.
42. Akyon, F.C.; Altinuc, S.O.; Temizel, A. Slicing aided hyper inference and fine-tuning for small object detection. In Proceedings of the 2022 IEEE International Conference on Image Processing (ICIP), Bordeaux, France, 16–19 October 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 966–970.
43. Zhou, X.; Wang, D.; Krähenbühl, P. Objects as points. *arXiv* **2019**, arXiv:1904.07850.
44. Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10781–10790.
45. Guo, X. A novel Multi to Single Module for small object detection. *arXiv* **2023**, arXiv:2303.14977.

46. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:2107.08430.
47. Pan, X.; Ren, Y.; Sheng, K.; Dong, W.; Yuan, H.; Guo, X.; Ma, C.; Xu, C. Dynamic refinement network for oriented and densely packed object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11207–11216.
48. Wei, H.; Zhang, Y.; Chang, Z.; Li, H.; Wang, H.; Sun, X. Oriented objects as pairs of middle lines. *ISPRS J. Photogramm. Remote Sens.* **2020**, *169*, 268–279. [CrossRef]
49. Ming, Q.; Zhou, Z.; Miao, L.; Zhang, H.; Li, L. Dynamic anchor learning for arbitrary-oriented object detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; Volume 35, pp. 2355–2363.
50. Yang, X.; Yang, J.; Yan, J.; Zhang, Y.; Zhang, T.; Guo, Z.; Sun, X.; Fu, K. Scrdet: Towards more robust detection for small, cluttered and rotated objects. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8232–8241.
51. Zheng, M.; Qi, G.; Zhu, Z.; Li, Y.; Wei, H.; Liu, Y. Image dehazing by an artificial image fusion method based on adaptive structure decomposition. *IEEE Sens. J.* **2020**, *20*, 8062–8072. [CrossRef]
52. Zhu, Z.; Luo, Y.; Wei, H.; Li, Y.; Qi, G.; Mazur, N.; Li, Y.; Li, P. Atmospheric light estimation based remote sensing image dehazing. *Remote Sens.* **2021**, *13*, 2432. [CrossRef]