Measurement Science and Technology

100 year anniversary
Measurement Science and Technology

PAPER

AeroDetectNet: a lightweight, high-precision network for enhanced detection of small objects in aerial remote sensing imagery

To cite this article: Ruihan Bai *et al* 2024 *Meas. Sci. Technol.* **35** 095402

IOP Publishing Measurement Science and Technology
Meas. Sci. Technol. **35** (2024) 095402 (20pp) https://doi.org/10.1088/1361-6501/ad41fe

AeroDetectNet: a lightweight, high-precision network for enhanced detection of small objects in aerial remote sensing imagery

Ruihan Bai[1], **Jiahui Lu**[2], **Zhiping Zhang**[2], **Mingkang Wang**[2] and **Qiang Wang**[3,*]

[1] School of Civil and Transportation Engineering, Hohai University, Nanjing, Jiangsu 210098, People's Republic of China
[2] School of Civil Engineering, Tongji University, Shanghai 200000, People's Republic of China
[3] UAV Industry Academy, Chengdu Aeronautic Polytechnic, Chengdu, Sichuan 610100, People's Republic of China

E-mail: wq@cap.edu.cn, bairuihanup@163.com, 2232403@tongji.edu.cn, zzip@tongji.edu.cn and wmk126126@126.com

Received 28 December 2023, revised 21 April 2024
Accepted for publication 23 April 2024
Published 4 June 2024

Abstract

Object detection in remote sensing imagery exhibits difficulties due to complex backgrounds, diverse object scales, and intricate spatial context relationships. Motivated by the problems mentioned above, this paper introduces AeroDetectNet, a novel lightweight and high-precision object detection network custom-designed for aerial remote sensing scenarios, building upon the YOLOv7-tiny algorithm. It enhances performance through four key improvements: the normalized Wasserstein distance for consistent object size sensitivity, the Involution module for reduced background noise, a self-designed RCS-biformer module for better spatial context interpretation, and a self-designed WF-CoT SPPCSP feature pyramid for improved feature map weighting and context capture. Ablation studies conducted on a hybrid dataset composed of three open-source remote sensing datasets (including NWPU VHR-10 remote sensing images, RSOD remote sensing images, and VisDrone UAV images) have demonstrated the effectiveness of four improvements specifically for small-size object detection. Visualizations through Grad-CAM further demonstrate AeroDetectNet's capacity to extract and focus on key object features. Upon individual testing across three open-source datasets, AeroDetectNet has successfully demonstrated its ability to identify objects in images with a smaller pixel area. Through experimental comparisons with other related studies, the AeroDetectNet achieved a competitive mAP while maintaining fewer model parameters, highlighting its highly accurate and lightweight properties.

Keywords: object detection, YOLOv7-tiny, remote sensing

1. Introduction

1.1. Objective and research problem

In recent years, the rapid development of artificial intelligence, especially the continuous improvement of computer vision algorithms, has opened up new possibilities for automated management and measurement. As a result, the computer vision-based approach is receiving increased attention in technological research across diverse industries, including microelectronics [1, 2], transportation [3, 4], railway [5, 6], and civil engineering [7, 8]. Likewise, numerous research studies have integrated computer vision and artificial intelligence into remote sensing engineering for climate change research,

* Author to whom any correspondence should be addressed.

military surveillance, urban development, and disaster monitoring. However, a significant distinction between remote sense images and typical object detection scenarios is the existence of small-size objects. Specifically, small-size objects occupy fewer pixels than large ones, which poses a significant challenge for feature extraction and is susceptible to complex background information. Directly applying object detection models struggles to achieve satisfactory performance.

1.2. Background and related literature

To address this challenge, many researchers have improved the detection performance of small-size objects. Current research endeavors in the field can be categorized into two principal domains: enhancing the image quality to enrich the feature information of small-size objects and improving the performance of the object detection model.

The first category of approach emphasizes improving the image's quality. Specifically, Kisantal *et al* [9] proposed a method to enhance detection performance by copying and pasting small objects multiple times within an image. Chen *et al* [10] utilized image scaling and stitching to improve feature visibility. Romano *et al* [11] developed a super-resolution (SR) approach to address the limitations of small object resolution. The technique allowed the neural network to learn the relationship between low-resolution images and their high-resolution images, resulting in more detailed and informative features for small-size object detection. Moreover, numerous attempts have been made to utilize generative adversarial networks (GANs) to enhance small objects'resolution and feature representation. Bai *et al* [12] proposed a multi-task GAN to recover clear SR objects from small fuzzy objects. Li *et al* [13] introduced a perceptual GAN approach for identifying small objects. Although data augmentation helps address the challenges posed by small objects with limited visual information and low resolution to some degree, data augmentation techniques may inadvertently introduce noise or artifacts, negatively affecting objects' detection accuracy. Integrating extra image preprocessing modules increases resource consumption, limiting its broad applicability. As a result, more research has focused on the second category of approach.

The second category of approach is to improve the performance of the object detection model. Among others, Wang *et al* [14] propose a new evaluation metric, the normalized Gaussian Wasserstein distance (NWD), to improve the detection performance of anchor-based detectors on small-size objects. The authors argue that traditional intersection over union (IoU)-based metrics are not well-suited for small-size objects, as they are susceptible to slight variations in object location and size. Moreover, they demonstrate that the proposed NWD metric can be integrated into anchor-based detectors, such as faster R-CNN and cascade R-CNN, without requiring any modifications to the network architecture. Ren *et al* [15] adapt faster R-CNN specifically for detecting small-size objects in remote sensing images. Modifications include adjusting the RPN stage with appropriate anchors and leveraging a high-resolution feature map with top–down and skip connections. The CAB Net proposed by Cui *et al* [16] enhances small object detection by utilizing a context-aware block with pyramidal dilated convolutions integrated into a truncated backbone network. The PaddleDetection team introduced the PP-YOLOE-SOD model, which is specialized for small object detection. The model utilizes a vector-based DFL algorithm consistent with the dataset's distribution and employs a central prior optimization strategy for fine-grained small-size object detection [17]. Zhang *et al* [18] introduce a novel object detection model for remote sensing images named SuperYOLO. This method cleverly integrates multimodal data and conducts high-resolution detection on multiscale objects. It does this by leveraging SR learning, balancing detection accuracy and computational demands. Tan *et al* [19] designed a four-scale detection layer and introduced a complete intersection over the union-NMS build on YOLOv5. Zhou *et al* [20] propose an enhanced YOLOv5-S model that utilizes data augmentation, a contextual transformation module, and K-means++ to address remote sensing image detection challenges. The CA-YOLO model, an enhancement of the YOLOv5 proposed by Shen *et al* [21], is optimized for object detection in intricate remote sensing images. The model improves feature extraction ability by integrating a coordinate attention module and a spatial pyramid pooling-fast mechanism while minimizing unnecessary data interference. Li *et al* [22] introduce AC-CNN, an attention-to-context convolution neural network that integrates global and local contextual information into existing region-based convolutional neural networks (CNN) frameworks, demonstrating superior performance in extensive experiments on small object detection. Wu *et al* [23] improve the YOLO-v5 algorithm by incorporating the multiscale anchor mechanism of Faster R-CNN. The enhanced method, YOLO-v5 + R-FCN, outperforms other algorithms in datasets like NWPU VHR-10, especially in detecting small remote sensing targets such as tennis courts and vehicles. Ji *et al* [24] introduce the MCS-YOLOv4 algorithm, an enhancement of YOLOv4: the new model incorporates an additional scale detection for richer location data and introduces an expanded field-of-perception module to capture and integrate contextual features; an attention module is added to minimize the impact of irrelevant image information; and the Soft-CIOU loss function is refined to boost the detection accuracy of small objects by enhancing the contribution of these objects to the loss function. Liang *et al* [25] introduce a novel two-stage detector, an improvement of faster-RCNN, optimized for small object detection. The detector utilizes a feature pyramid architecture with lateral connections for improved small-size object sensitivity and employs specialized anchors trained with focal loss. Benjumea *et al* [26] refined the YOLOv5 by altering the structural components of the model, and a new variant called 'YOLO-Z' was introduced. This adaptation exhibits a 6.9% improvement in mAP for detecting small objects, with a negligible increase in inference time. Li *et al*

[27] introduce LV-Net, an end-to-end deep network specifically for remote sensing images. LV-Net uses two primary modules: the L-shaped module that extracts diverse scales and local details and the V-shaped module that merges encoder and decoder features to highlight salient objects. Chen *et al* [28] present the first comprehensive survey of deep learning-based small object detection, covering key aspects such as contextual information and advancements in detection networks. The LA-YOLO network by Ma *et al* [29] enhances low-altitude UAV detection using the SimAM attention mechanism and a novel fusion block, demonstrating a 5.9% increase in precision on the GUET-UAV-LA dataset over existing models. Wang *et al* [30] introduce DFANet, a network that improves semantic segmentation of remote sensing images using multi-scale feature aggregation and a conditional random field module, surpassing existing models on ISPRS datasets. Lu *et al* [31] introduce the 'attention and feature fusion SSD,' a one-stage object detection network for complex remote sensing images. It uses multilayer fusion, dual-path attention, and multi-scale enhancements to boost accuracy. Gu *et al* [32] introduce GLE-Net, a network that enhances object detection in remote sensing images using a novel global and local enhanced attention mechanism (GLE-AM) to improve the detection of small objects. Hui *et al* [33] introduce STF-YOLO, a novel UAV image detection algorithm that incorporates SwinTransformer with CNNs in a structure called STRCN and uses a lightweight classifier, CNeB, for enhanced accuracy. Chen *et al* [34] introduce BiShuffleNeXt, an efficient remote sensing scene classification model that utilizes a dual-path architecture with sandglass bottlenecks to enhance semantic and spatial information processing.

1.3. Proposed method

Despite significant progress in small-size object detection, contemporary research still has room for improvement. First, many of the proposed enhancements inadvertently increase the number of parameters in the model. Consequently, models with greater parameters demand more extraordinary memory, making them harder to implement in real-world scenarios. Moreover, existing context-based object detection models improve small-size detection performance by focusing on enhancing the model's understanding of target positioning and strengthening the relationship between targets and their environmental context at the macro level, neglecting the integration with finer details in the objects' spatial-specific and channel-specific features at the micro level. Generally, spatial features refer to the information about the location and shape of objects in an image, and channel features typically involve an image's color, texture, and other detailed information. The potential of spatial and channel features of objects is also significant. Motivated by the problems mentioned above, this study proposes AeroDetectNet, a novel lightweight and high-precision target detection network custom-designed for aerial remote sensing scenarios, building upon the foundation of the YOLOv7-tiny model. Specifically, we prioritize maintaining a lightweight model for practical applications.

(1) Use the normalized Wasserstein distance metric instead of the CIoU metric to evaluate the similarity between predicted and ground-truth bounding boxes, ensuring consistent sensitivity across objects with different sizes.
(2) Replace the convolution module in front of the model detection head with the Involution module to minimize the impact of background noise, thereby enabling a more focused analysis of the target object.

Then, we innovate and expand upon the context-based module (corresponding to the following two improvement points). Not only do we retain contextual information features through the context-based module, but more importantly, we focus on integrating the optimization of spatial or channel features into the context-based module. Integrating spatial or channel features makes the model more effective in capturing and thoroughly understanding the critical self-characteristics of small objects at the micro level; further applying spatial features enhances its ability to differentiate between targets and backgrounds in complex environments at the macro level.

(3) The RCS-biformer module was integrated into the model's backbone to improve the model's feature extraction capabilities: the channel shuffle characteristic of the RepConv ShuffleNet (RCS) convolution enhances the model's ability to represent micro features by reorganizing the different channels features, thereby facilitating information exchange between features. Combined with the macro feature of the context-based module Biformer with channel feature processing, the model has a richer feature representation.
(4) By integrating the weighted sum concat approach and the context transformer (CoT) module into the SPPCSP architecture of YOLOv7-tiny. The combined use of the weighted sum concat approach allows for the adjustment of the importance of features in each feature map according to different situations, enabling the model to emphasize spatial crucial for specific detection tasks. The CoT module allows for maintaining contextual information.

1.4. Contribution

(1) This study conducted a comparative analysis of AeroDetectNet and YOLOv7-tiny's performance in detecting small-sized targets with precision. Specifically, the focus was on each dataset's smallest pixel size categories: aircraft in the NWPU VHR-10 dataset, oiltank in the RSOD dataset, and the vehicle in the VisDrone dataset. By calculating the average ratio of the correct detected bounding box area to the image area for these categories, AeroDetectNet outperformed YOLOv7-tiny. In detail, AeroDetectNet achieved ratios of 1.71% for aircraft category, 2.97% for oiltank category, and 2.12% for vehicle category, compared to YOLOv7-tiny's 1.98%, 3.28%, and 2.39%, respectively. This finding highlights AeroDetectNet's superior ability to identify small-sized targets precisely.
(2) Compared with leading object detection algorithms, AeroDetectNet demonstrated commendable performance

with a mAP@0.5 of 0.895 while maintaining lower parameters and GFLOPs at 4.96 M and 10.3, respectively. Compared to models like YOLOv5, YOLOX, and YOLOv8, AeroDetectNet balances efficiency and accuracy. Especially when compared with more complex models such as Faster RCNN and Cascade RCNN, which score higher in mAP@0.5 (0.901 and 0.913, respectively) but significantly increase in parameter count and computational demands (reaching 41.17 M/206.71GFLOPs and 68.95 M/234.49GFLOPs, respectively), the comparative analysis emphasizes the value of AeroDetectNet in the current field of research, particularly for achieving high efficiency and precision in detection tasks.

1.5. Organization of research

The remainder of this paper is as follows: section 2 introduces the overview of the proposed AeroDetectNet model; section 3 introduces the experimental data used in this paper and evaluation metrics for model performance; section 4 presents the different experiments (including ablation experiments, model robustness experiments, and comparison experiment). The results indicate that the AeroDetectNet maintains excellent performance across various sizes, highlighting its outstanding robustness and adaptability to varying inputs. Upon individual testing across three open-source datasets (including NWPU VHR-10 remote sensing images, RSOD remote sensing images, and VisDrone UAV images), AeroDetectNet has successfully demonstrated its ability to identify objects in images with a smaller pixel area. Through experimental comparisons with other related studies, the AeroDetectNet achieved a competitive mAP while maintaining fewer model parameters, highlighting its highly accurate and lightweight properties; section 5 demonstrates the conclusion of this paper.

2. Method

2.1. Fundamentals of the YOLOv7-tiny algorithm

YOLO series algorithms are the most typical representative of the one-stage object detection algorithm, providing a balance of inference speed and detection precision. As one of the most advanced algorithms in the current YOLO series algorithms, YOLOv7 [35] was proposed in 2022, which surpasses the previous algorithms in many aspects. YOLOv7-tiny is its lightweight version with a compact network structure and fast detection speed. Like other object detection algorithms, the YOLOv7-tiny network structure mainly comprises four parts: the part of model image input, the model backbone for high-level feature extraction, the model neck for high and low-feature fusion, and the model head for final object detection.

Figure 1 shows the overall network structure of YOLOv7-tiny. First, the training images are fed into the input part of the model with the default size. Immediately afterward, the images are passed into the model backbone, including CBL, MCB, and SPPCSP modules. Among them, the CBL module is the basic unit of YOLOv7-tiny, which consists of convolution, regularization, and LeakyReLU activation functions in sequence. The MCB module concatenated the original and extracted features, realizing the fusion of local and global features. The SPPCSP integrates feature maps using a range of pooling kernel sizes to enhance the algorithm's robustness for objects with substantial variations. After going through the model backbone described above, the feature hierarchy becomes high-level instead of low-level. Typically, low-level features exhibit more pronounced semantic information, whereas high-level features emphasize geometric details. The model neck could complement the information from high and low-level features to enhance the model feature extraction capability further. Last, the model head separately corresponds to three feature maps derived from the model neck.

2.2. Overview of AeroDetectNet

The YOLOv7-tiny algorithm is renowned for its lightweight design, low computational demands, and fast detection capabilities. It is exceptionally well-suited for deployment on edge devices and aerial platforms like drones. Nevertheless, YOLOv7-tiny, due to its inherent model characteristics, continues to face challenges in meeting the precision standards mandated by complex remote sensing image scenes.

In this research, we proposed AeroDetectNet, a novel lightweight and high-precision target detection network tailored for aerial remote sensing scenarios, which is constructed on the foundation of the YOLOv7-tiny algorithm. Figure 2 depicts the comprehensive architecture of AeroDetectNet. First, a novel loss function NWD is employed to assess the similarity between predicted and ground-truth bounding boxes to ensure uniform sensitivity across objects of varying sizes. Subsequently, we have integrated the Involution module in the network's detection head, replacing the standard convolution module to enhance the network's ability to identify small and intricate targets. Furthermore, we incorporate the RCS-biformer into the backbone to enhance feature extraction capabilities in the context of intricate backgrounds. Finally, we enhance the SPPCSP architecture by introducing the WF-CoT SPPCSP module, utilizing weighted summation and concatenation methods alongside the CoT module. This augmentation bolsters the model's capacity to balance weight relationships among diverse feature maps and facilitates the acquisition of contextual information. AeroDetectNet represents a significant stride in achieving the delicate equilibrium between precision and model complexity, addressing the formidable challenges of intricate remote sensing scenarios.

2.2.1. NWD. The YOLOv7-tiny algorithm evaluates the similarity between the predicted and the ground truth bounding box based on the CIoU loss function, which is calculated by:

$$L_{CIoU} = 1 - \frac{|B \cap B_i|}{|B \cup B_i|} + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha\upsilon \tag{1}$$

$$\alpha = \frac{\upsilon}{1 - IOU + \upsilon} \tag{2}$$

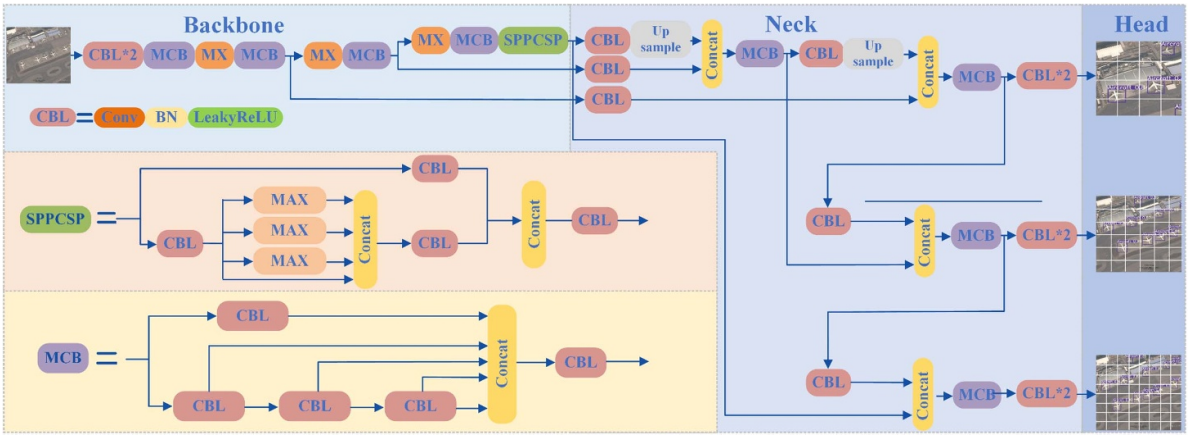

Figure 1. YOLOv7-tiny structure.

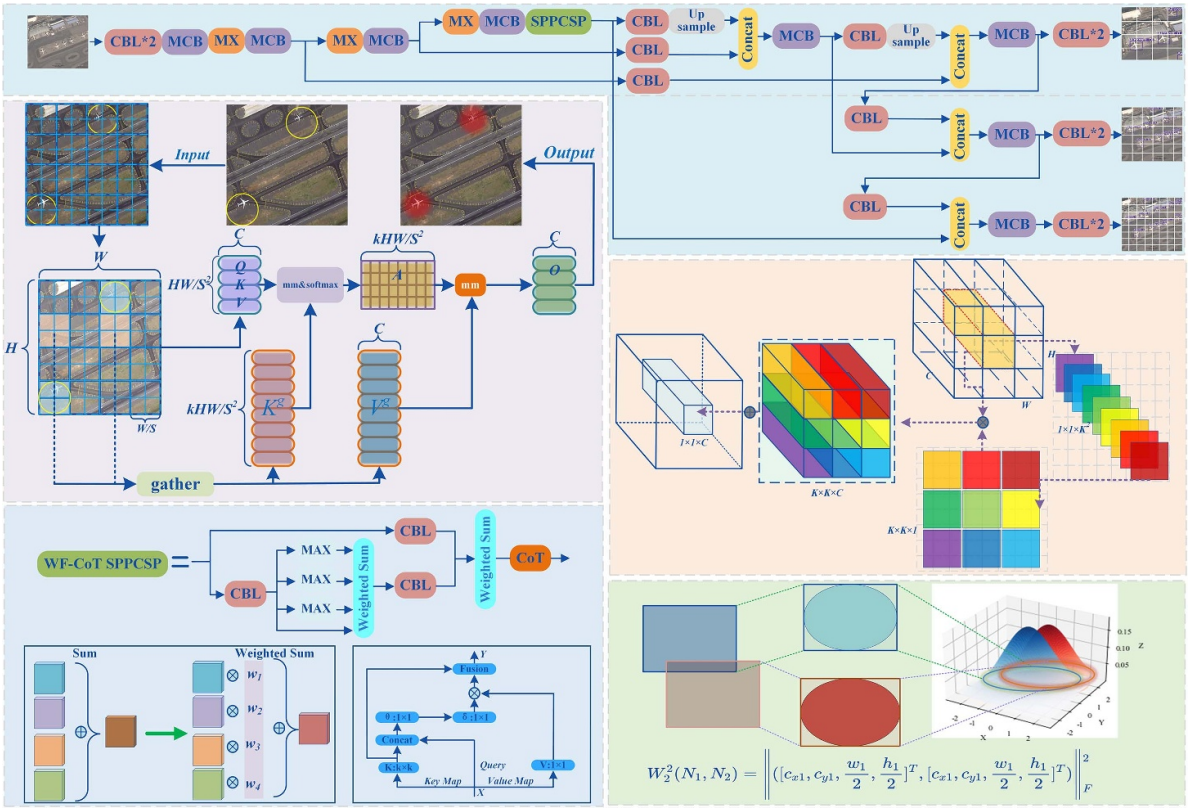

Figure 2. AeroDetectNet structure.

$$\upsilon = \frac{4}{\pi^2}\left(\arctan\left(\frac{w^{gt}}{h^{gt}}\right) - \arctan\left(\frac{w}{h}\right)\right)^2 \tag{3}$$

where B and B_i are predicted and ground-truth object bounding boxes. b and b^{gt} are the center points of the prediction frame and the real frame. ρ is the Euclidean distance between the two center points. c is the diagonal distance of the minimum circumscribed rectangle between the prediction and real frames. h, h^{gt}, w, and w^{gt} represent the prediction and real frames, respectively.

The calculation of the CIoU-based loss function depends on the overlap area between the ground truth and the predicted bounding boxes. Even minor positional deviations can lead to large changes in the CIoU values for small-size objects. As shown in figure 3(a), the red, green, and black boxes represent the ground truth bounding box A, predicted bounding box B, and predicted bounding box C, respectively. For small-size objects like baseball fields that only occupy a few pixels, the change in CIoU value is significant when the predicted bounding box shifts slightly. Conversely, for large-size objects like tennis courts or football fields, the same positional shifts lead to only minor changes in CIoU values (as depicted in figures 3(b) and (c)). During the training phase, the instability of prediction bounding boxes for

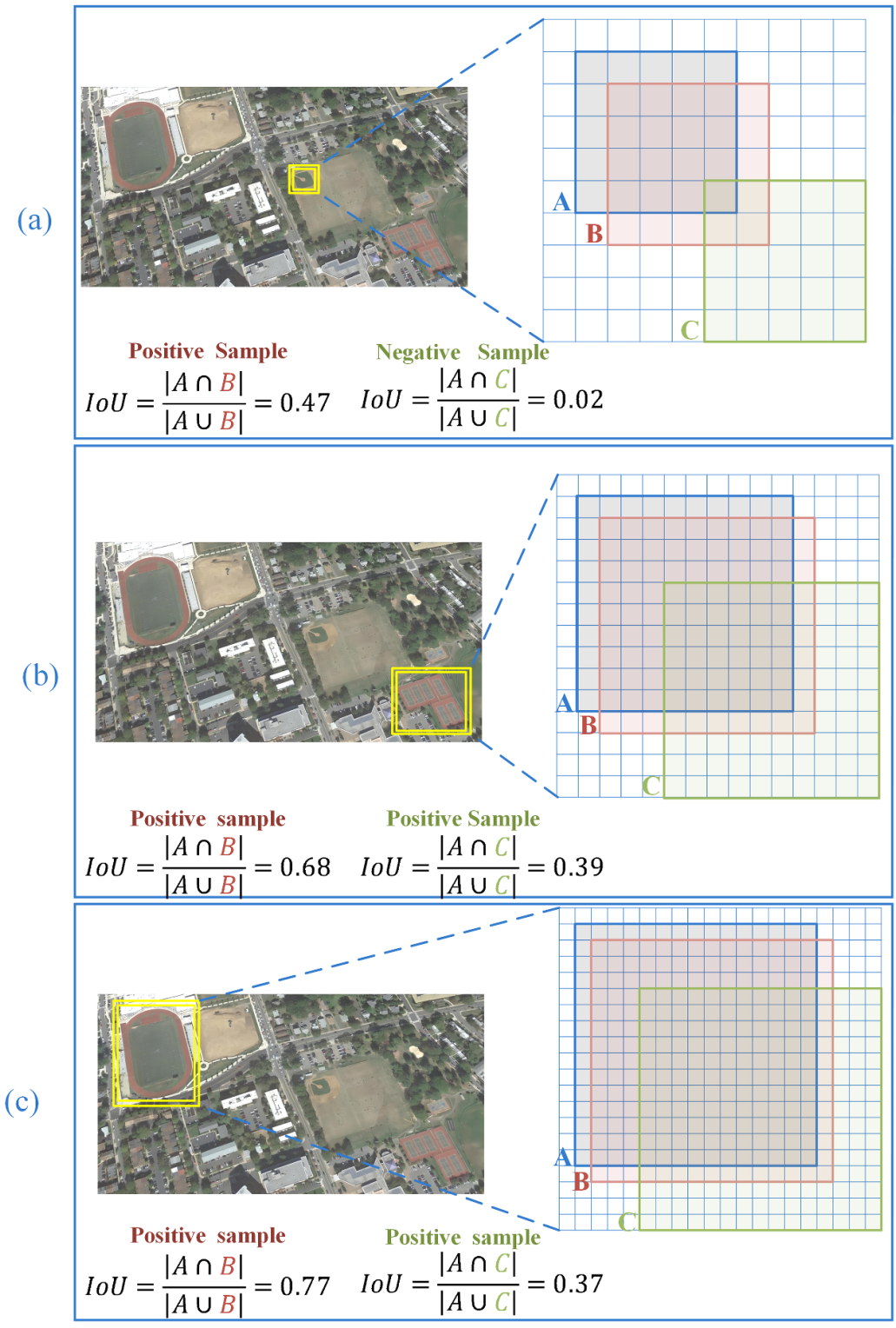

Figure 3. The sensitivity analysis of CIoU on different-size objects: (a) small-size objects, (b) medium-size objects, and (c) large-size objects.

small-sized objects triggers the alternation of positive and negative samples. The shift between positive and negative samples disrupts the model's training, leading to challenges in achieving stable convergence.

To address the limitations of the YOLOv7-tiny loss function, this paper introduces the NWD [14] as a more effective metric for small-size object detection. As shown in figure 4, the NWD metric measures the similarity between bounding boxes by converting the object bounding box into the two-dimensional Gaussian distribution and measuring the similarity through the Wasserstein distance. The process of this transformation is as follows:

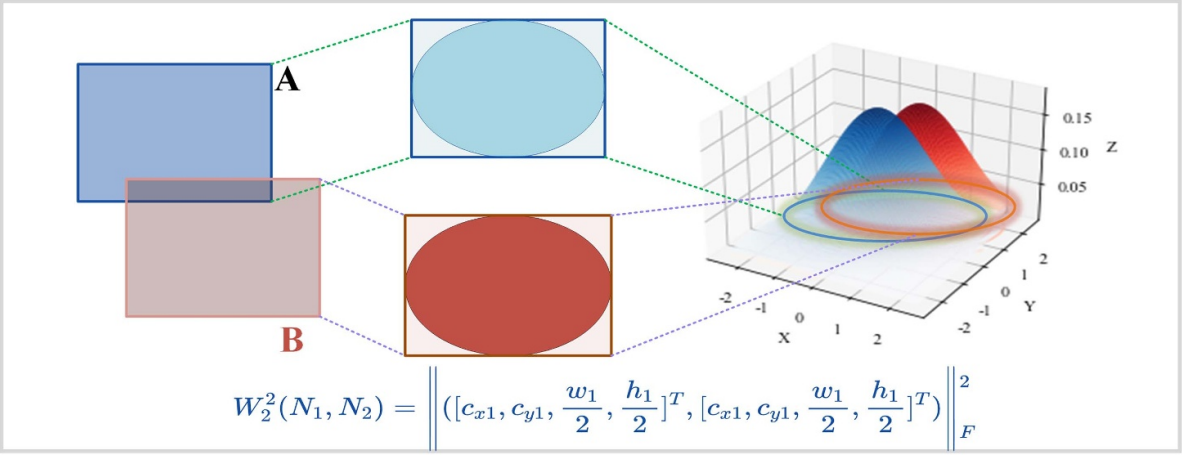

Figure 4. The calculation of the normalized Wasserstein distance (NWD) metric.

- Bounding box to ellipse: for a bounding box R (c_x, c_y, w, h), c_x, c_y, w, h indicate the bounding box's center point, width, and height, respectively. The equation for its inner tangent ellipse E (c_x, c_y, $w/2$, $h/2$) is:

$$\frac{(x-c_x)^2}{\left(\frac{w}{2}\right)^2} + \frac{(y-c_y)^2}{\left(\frac{h}{2}\right)^2} = 1 \tag{4}$$

where (c_x, c_y) represents the center coordinate of the ellipse, and ($w/2$, $h/2$) represents the length along the x-axis and y-axis.

- Gaussian distribution parameters: the probability density function of a 2D Gaussian distribution could be expressed as:

$$f\left(X|\mu, \sum\right) = \frac{\exp\left(-\frac{1}{2}(X-\mu)^T \sum -1\,(X-\mu)\right)}{2\pi\,|\sum|^{\frac{1}{2}}} \tag{5}$$

where $\boldsymbol{X}$ is the position variable, ($\boldsymbol{\mu}$, $\sum$) is the mean vector and the covariance matrix.

- Ellipse to Gaussian distribution: when satisfied with equation (6), the inner tangent ellipse E becomes the contour of a 2D Gaussian distribution $N(\boldsymbol{\mu}, \sum)$, of which:

$$(X-\mu)^T \sum^{-1} (X-\mu) = 1 \tag{6}$$

$$E(c_x, c_y, w/2, h/2) \sim N\left(\mu, \sum\right) | \mu = \begin{bmatrix} c_x \\ c_y \end{bmatrix},$$
$$\sum = \begin{bmatrix} \frac{w^2}{4} & 0 \\ 0 & \frac{h^2}{4} \end{bmatrix}. \tag{7}$$

- Bounding box to Gaussian distribution: the bounding box is converted into a two-dimensional Gaussian distribution.

The similarity of two bounding boxes is transformed into the distance distribution of two 2D Gaussian distributions. For two 2D Gaussian distributions $N_1(\boldsymbol{\mu_1}, \sum_{\mathbf{1}})$ and $N_2(\boldsymbol{\mu_2}, \sum_{\mathbf{2}})$, the Wasserstein distance between $\boldsymbol{\mu_1}$ and $\boldsymbol{\mu_2}$ can be calculated by:

$$W_2^2(N_1, N_2) = \left\| \left(\left[cx_1, cy_1, \frac{w_1}{2}, \frac{h_1}{2} \right]^T, \left[cx_1, cy_1, \frac{w_2}{2}, \frac{h_2}{2} \right]^T \right) \right\|_F^2 \tag{8}$$

where $\|.\|_F$ is Frobenius norm.

Overall, size-dependent sensitivities in CIoU-based metrics lead to unstable label assignments on small-size objects, resulting in positive and negative sample boxes having similar characteristics. NWD metrics are insensitive to objects with different sizes, and the variation of its value is smooth for small-size objects. In this study, we have selected the NWD metric to serve as the loss function for AeroDetectNet. This approach can better represent the object's spatial distribution, enhancing object detection and localization performance.

2.2.2. Involution. In aerial remote sensing image analysis, the challenge of accurately detecting small objects stems from their limited pixel coverage. The lack of pixel representation restricts the contextual information used for object recognition, making this image feature more prone to interference from the background noise. This paper introduces the Involution module [36] placed in front of the detection head layer of the network, aiming to provide an adaptive way to understand the features and surroundings of small objects in remote sensing images. Among others, the Involution module is designed to create spatially specific kernels based on the input features, allowing for adaptive weighting across different spatial locations on the feature map. By doing so, the module minimizes the impact of background noise, thereby enabling a more focused analysis of the target object. Figure 5 provides the detailed procedure of the Involution module.

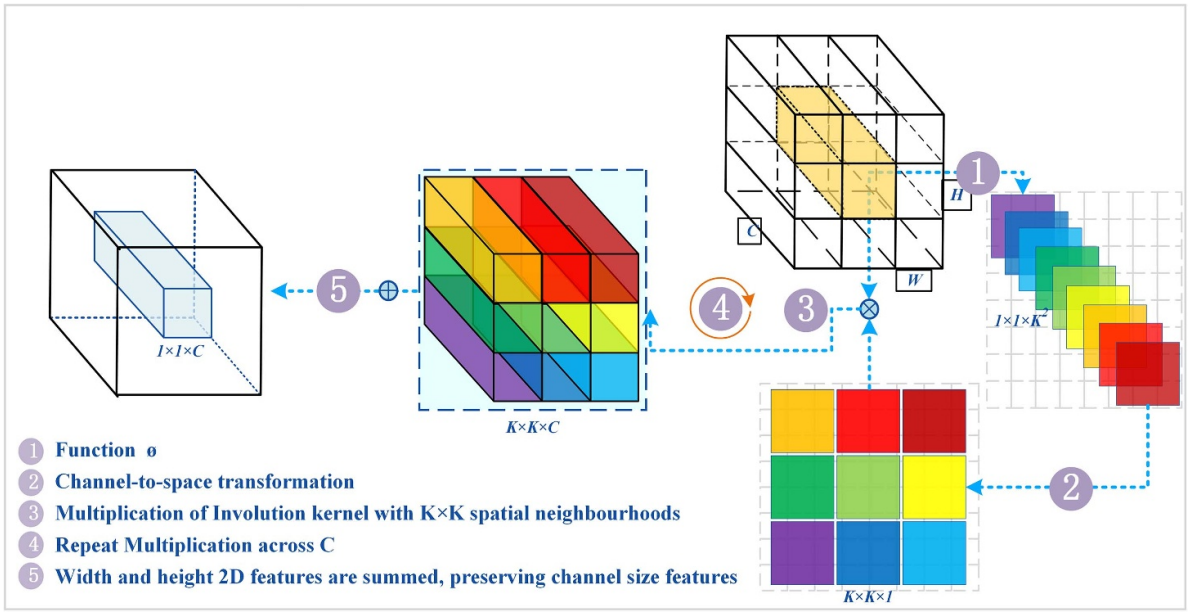

Figure 5. The principle of involution.

(1) The involution module extracts a feature vector from a specific feature point (the features corresponding to the small objects like the baseball field within the image) and processes it using a function ø. This function transforms the feature vector into a shape suitable for a kernel.
(2) Following a channel-to-space transformation, this reshaped vector becomes an Involution kernel tailored to the specific feature point. The specialized kernel is applied to neighboring feature vectors around the specific feature point.
(3) The $K \times K \times 1$ Involution kernel is multiplied by the pixels in the $k \times k$ neighborhood of the specific feature point in the original feature.
(4) Step 3 was repeated across the C channels to obtain a $K \times K \times C$ three-dimensional matrix.
(5) The 2D features in the width and height dimensions are summed, and the features in the channel dimension are maintained.

The involution module is designed to create spatially specific kernels based on the input features, allowing for adaptive weighting across different spatial locations on the feature map. By doing so, the module minimizes the impact of background noise, thereby enabling a more focused analysis of the target object. This research opts to integrate the convolution into the detection head of AeroDetectNet.

2.2.3. RCS-biformer. In small-size object detection models, the model backbone is crucial for transforming raw images into rich features. However, the use of CNNs as the model backbone limits the model's ability to grasp the global context due to the restricted receptive fields of its convolutional layers. To enhance performance, integrating global information is critical. It can create contextual links between small targets and their surroundings, effectively reducing false detection by distinguishing these targets from similarly shaped or textured background elements. Accordingly, this paper incorporates an improved biformer module [37] called RCS-biformer into the YOLOv7-tiny algorithm's backbone to address these limitations. The original biFormer block is constructed upon the bi-level routing attention. As illustrated in figure 6, the workflow of bi-level routing attention is as follows:

- Region partition and input projection: given a *2D* input feature map $X \in R^{H \times W \times C}$, the module divides it into $S \times S$ non-overlapping regions, each with HW/S^2 vectors. Transform the shape of X to get $X^{\mathrm{r}} \in R^{S2 \times HW/S2 \times C}$. Next, the query, key, and value tensors Q, K, and V are derived from X^{r} using linear projection, and the formula is as follows:

$$Q = X^{\mathrm{r}} W^{Q} K = X^{\mathrm{r}} W^{K} V = X^{\mathrm{r}} W^{V}. \tag{9}$$

- Attention relationship from region to region: obtain region-level queries and keys Q^{r} and K^{r} by applying per-region averages on Q and K, respectively. Then, the adjacency matrix A^{r} is obtained through matrix multiplication between Q^{r} and the transposed K^{r} to measure the inter-region correlation, and the formula is as follows:

$$A^{\mathrm{r}} = Q^{\mathrm{r}} (K^{\mathrm{r}})^{T}. \tag{10}$$

- Retaining top-k connections for each region: retained the top-k most relevant regions in A^{r} to get the routing index matrix I^{r}. The formula is shown in equation (13). Consequently, the ith row of I_{r} contains the indices of the k most relevant regions for the ith region.

$$I^{r} = \mathrm{topIndex}\,(A^{\mathrm{r}}). \tag{11}$$

- Gather key tensors K^{g} and value tensors V^{g}: Based on the region-to-region routing index matrix I^{r}, collect keys and

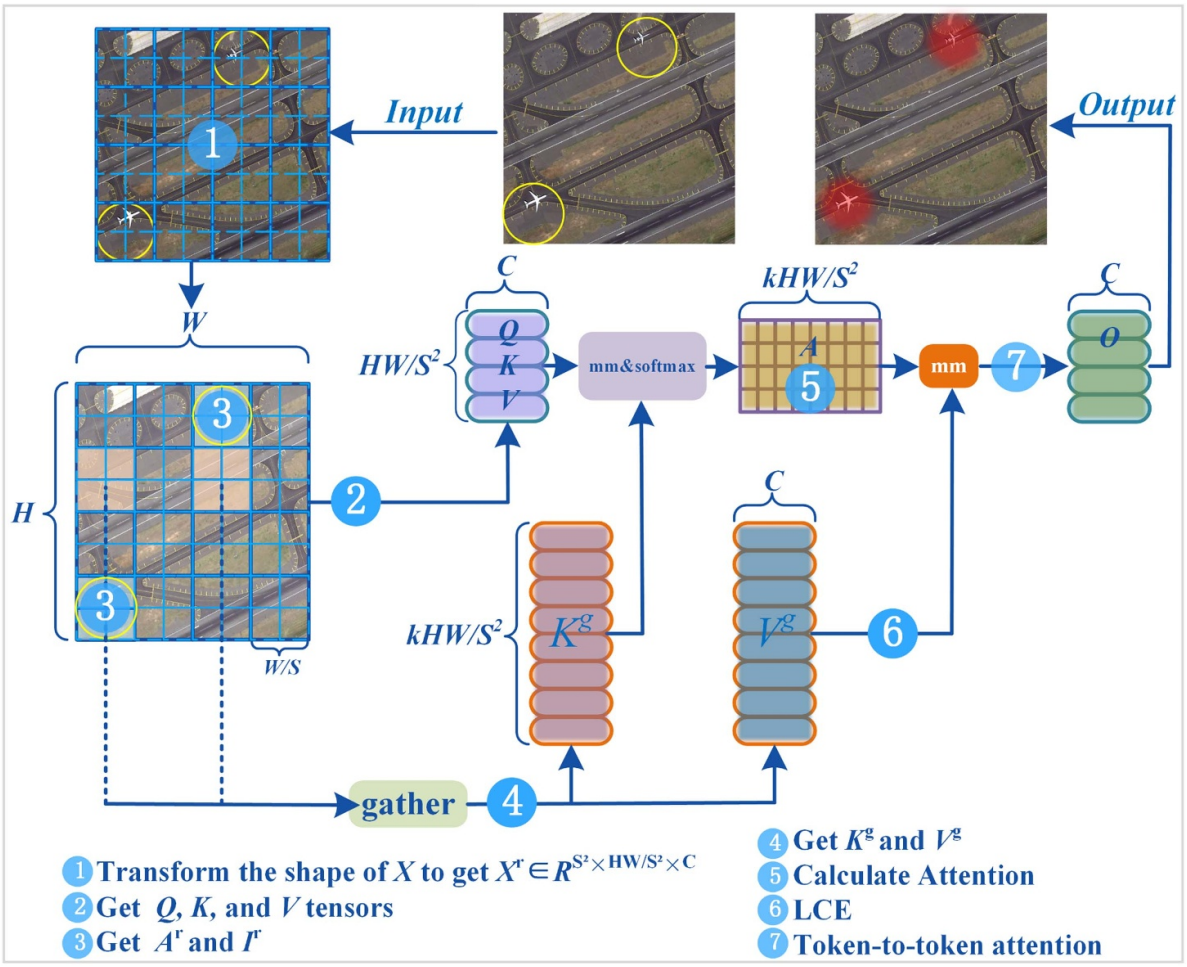

Figure 6. The workflow of bi-level routing attention.

values from the k most relevant regions for each query token. As shown in equation (12), this step is achieved using the gather function, resulting in the gathered key tensor K^{g} and value tensor V^{g},

$$K^{g} = \text{gather}\,(K, I^{r})\ V^{g} = \text{gather}\,(V, I^{r})\,. \tag{12}$$

- Compute attention weights and weighted summation: the dot product between query tokens Q and the gathered key tensor K^{g} is calculated. Then, apply the softmax function to these dot products to obtain attention weights. These weights represent the relative importance of key-value pairs in the k-routed regions for each query token. Apply the attention weights to the gathered value tensor V^{g}. Compute the attention output for each query token by summing the weighted values. The attention output O reflects the most relevant information from the input features for the query tokens.
- Local context enhancement (LCE): compute the LCE by performing depth-wise convolution on the input feature map. Then, add the LCE term to the attention output from the token-to-token attention. The final output O is calculated by

$$O = \text{Attention}\,(Q, K^{g}, V^{g}) + \text{LCE}\,(V)\,. \tag{13}$$

As illustrated in figure 7, within the original biFormer module, DWConv denotes depth-wise convolution that optimizes the model by reducing parameters and computational requirements. However, DWConv convolution performs the convolution operation independently on each channel, which may limit the feature interaction between different channels. The capture of fine-grained information often requires the collaborative work of multiple feature channels, and such interaction is missing in DWConv convolution.

This study used RCS convolution to replace the DWConv in our proposed RCS-BiFormer module. The key advantage of RCS convolution lies in its channel splitting, concatenation, and shuffling techniques, which enhance the feature description. Specifically, RCS divides the input feature map into two parts by channel splitting and performs convolution operations separately. This has the advantage of allowing the model to perform more focused processing on different subsets of features. When these channel subsets are convolved and concatenated, the model obtains a richer and more diverse representation of the features. Channel shuffling is designed to facilitate the exchange of information between different channels. In traditional convolutional operations, specific channels

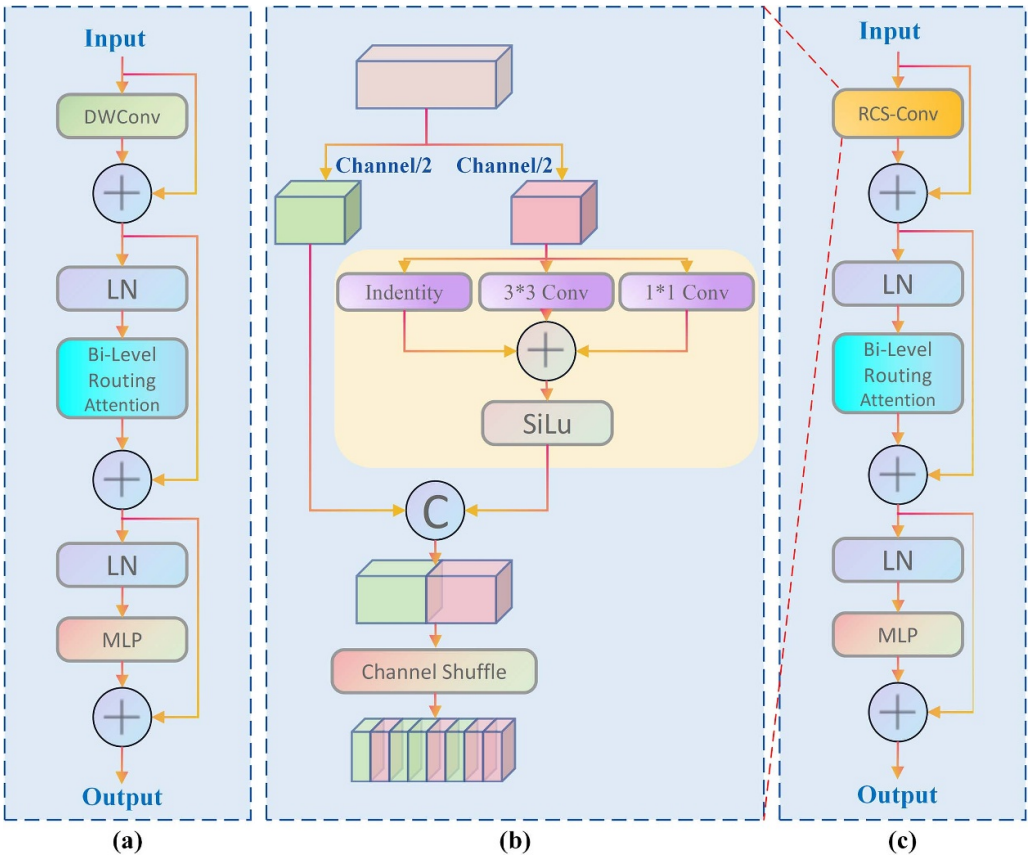

Figure 7. The network architecture diagram of RCS-biformer. (a) The network architecture of biformer, (b) the network architecture of RCS-Conv, while (c) depicts the proposed RCS-biformer by replacing DWConv with RCS-Conv.

may primarily interact only with certain fixed channels, resulting in information isolation. Channel shuffling can ensure broader communication between channels, thus enhancing the descriptiveness of features.

Channel shuffling significantly enhances the layering of channel features, enabling a more nuanced and interconnected representation at the channel-specific micro level. This process enriches the features within each channel by promoting a broader exchange of information across different channels, leading to a more comprehensive feature map that captures a diverse array of characteristics and patterns. On the other hand, biformer's context information extraction operates at the context-specific macro level, focusing on understanding the spatial relationships and structures within the image. Combining these two approaches produces a powerful interaction between micro and macro-level features. This interaction enables the model to understand the intricate details within channels (such as textures or color nuances) and the broader spatial context (like the position and relationship of objects in a scene). The synergy of micro and macro feature processing creates a more layered and contextually aware model, leading to a more robust and intelligent model capable of discerning subtle differences and similarities between objects and their environments.

2.2.4. WF-CoT SPPCSP. In the original SPPCSP architecture of the YOLOv7-tiny model, the summation approach is used to fuse different feature maps, which presents several challenges for detecting small objects. First, this approach introduces computational inefficiencies by adding redundant features, limiting the network's ability to concentrate on crucial features. Second, the feature fusion approach emphasizes self-attentive fusion, resulting in the fused feature that falls short of effectively extracting contextual information. The fused features that lack contextual information are detrimental to the subsequent detection analysis. This limitation weakens the sensitivity of the model detection head to small-size objects. Thus, as depicted in figure 8, this paper proposes a novel SPPCSP structure called weighted fusion—contextual transformer SPPCSP (WF-CoT SPPCSP) to overcome the above issues.

(1) Since input features have different resolutions, their contribution to output features is usually unequal. To solve this problem, this paper replaces the original sum approach with a weighted fusion approach. As shown in the following equation, additional weights are added to each input, allowing the network to understand the

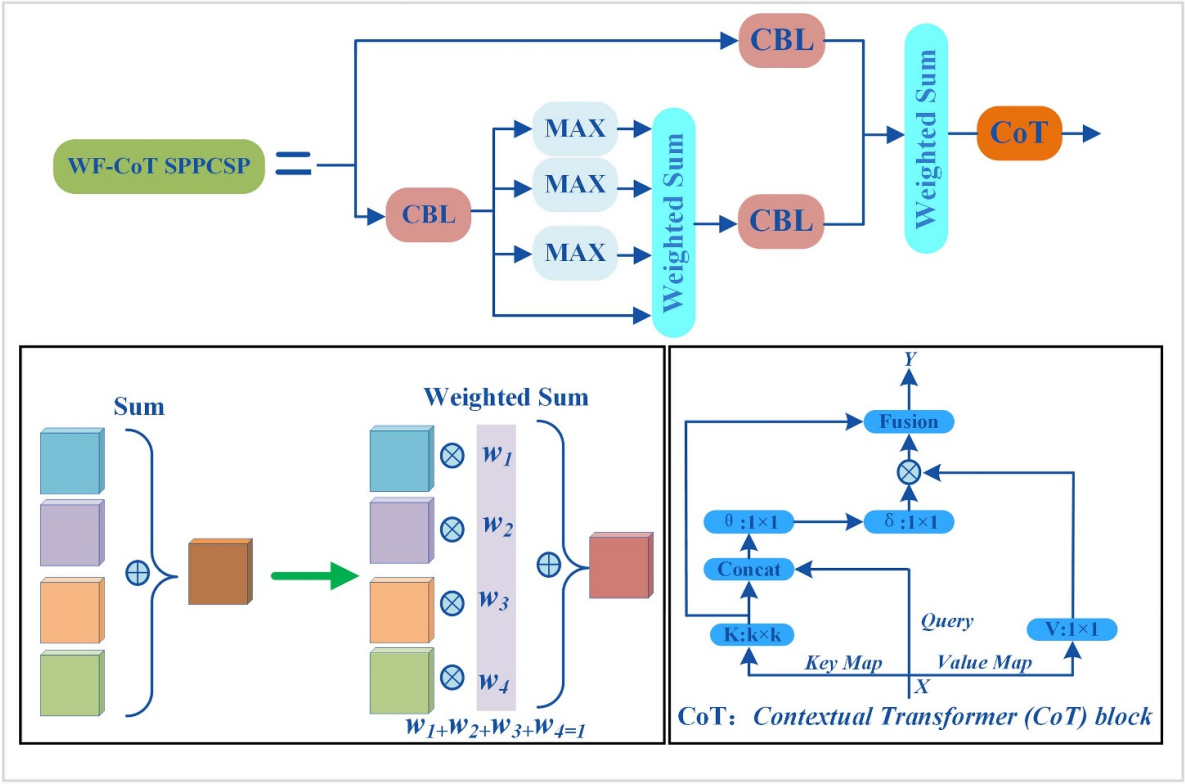

Figure 8. The network architecture diagram of improved WF-CoT SPPCSP.

importance of each input feature. The weighted sum can be expressed as:

$$Output = \sum_i \frac{e^{wi}}{\sum_j e^{wj}} Input_i \tag{14}$$

where w_i is a trainable weight. Since scalar weights are unbounded, this may lead to unstable training. This paper uses the Softmax function to make the probability of all weights normalized with values ranging from 0 to 1.

(2) Incorporation of the CoT module [38] for contextual sensitivity: the CoT module is added after the feature fusion phase. Given an input feature map X, the CoT module generates keys K, queries Q, and values V based on the input feature X. Then, the CoT module employs a $k \times k$ group convolution to spatially encode all adjacent keys within the $k \times k$ grid. The method captures the static contextual information among local neighboring keys, resulting in the contextual keys K^1. The contextual keys K^1 are concatenated with the queries Q. Then, two consecutive 1×1 convolution operations (W_θ with a ReLU activation function, W_δ without an activation function) are applied to the concatenated result to learn the dynamic multi-head attention matrix A, as described by

$$A = \left[K^1, Q\right] W_\theta W_\delta. \tag{15}$$

Using the learned dynamic contextual attention matrix A, a local matrix multiplication operation is performed with the value V to compute the dynamic contextual representation K^2. This step captures the dynamic feature interactions among input features, as described by:

$$K^2 = V \otimes A. \tag{16}$$

The final step is to fuse the static contextual representation K^1 with the dynamic contextual representation K^2. This produces the final output Y, which contains global and local information,

$$Y = \text{Fusion}\left(K^1, K^2\right). \tag{17}$$

The CoT module exploits the contextual information among neighboring keys to capture contextual relationships between fused multiscale features while augmenting visual representation by seamlessly integrating contextual and self-attention learning within a single architecture. Incorporating CoT improves the contextual sensitivity of small objects and facilitates later feature analysis.

Overall, the weighted fusion approach assigns different weights to each input feature, enabling the network to prioritize features more effectively based on their relevance to the detection task. This approach reduces computational redundancy by focusing on crucial feature maps and ensures that the feature fusion process emphasizes the most informative spatial features. 15Adding the CoT module further enriches this feature prioritization by integrating a deep understanding of the contextual relationships within the fused feature. While the

weighted fusion approach optimizes the input feature prioritization at the spatial-specific micro level, the CoT module complements this by enhancing the depth and breadth of contextual analysis at the macro level. This results in a more robust and detailed feature representation, capturing each feature's importance and intricate context. These approaches enable AeroDetectNet to distinguish small objects from complex backgrounds effectively. In this study, the newly proposed WF-CoT SPPCSP module is chosen as the feature pyramid architecture for AeroDetectNet.

3. Experiment introduction

3.1. Data introduction

This paper collected the experimental dataset from several open-source datasets, including the NWPU VHR-10 remote sensing images, RSOD remote sensing images, and VisDrone UAV images. The primary component of the experimental dataset is the NWPU VHR-10, which includes ten categories: aircraft, ships, storage tanks, baseball fields, tennis courts, basketball courts, ground tracks, ports, bridges, and vehicles. However, the number of NWPU VHR-10 datasets is relatively small, containing only around 600 images. To increase the number of datasets, the study incorporated the aircraft and storage tank categories from the RSOD dataset and the vehicle category from the VisDrone dataset. As a result, the final dataset's size expanded to approximately 1000 images. The datasets were divided into training, validation, and testing datasets in the ratio of 7:1:2. The data utilized in this investigation are illustrated in figure 9. Table 1 lists the instances of different classes in the training, validation, and testing datasets.

3.2. Evaluation metric

The confusion matrix is a widely used evaluation method that clearly and concisely represents the model performance by organizing the true and predicted values in a tabular format. In this format, rows signify the actual values, while columns correspond to the predicted values. When the model forecasts a positive outcome, it is denoted as P; conversely, it is indicated as N. Analogously, if the model's prediction is accurate, it is designated as T; if not, it is identified as F. The values within the format represent the number of objects associated with TP, FN, FP, and TN in image predictions. Precision and recall serve as the performance evaluation metrics for the model, and they are calculated using the following formulas:

$$precision = \frac{TP}{TP + FP} \tag{18}$$

$$recall = \frac{TP}{TP + FN}. \tag{19}$$

The PR curves represent the relationship between precision and recall at varying confidence thresholds. The integral of the prediction and recall curves is AP, and the average of AP for each category is mAP. The specific formula is as follows:

$$AP = \int_0^1 p(r)\,\mathrm{d}r \tag{20}$$

$$mAP = \frac{\sum_m AP}{m}. \tag{21}$$

3.3. Experimental configuration

The experimental environment of this study includes ubuntu18.04, python3.8.13, cuda-11.4, cudnn-8.2.2, and PyTorch 1.10.2. Hardware platform: RTX-3090GPU 8 g memory. The processor is intel(R)core(TM)i7- 6500 M CPU@3.20 GHZ.

4. Experimental results and analysis

4.1. Ablation experiment

In this paper, a series of improvements were applied to the YOLOv7-tiny algorithm to enhance the model performance while simultaneously reducing the model's parameter count. To systematically evaluate the effectiveness of the proposed improvement methods, ablation experiments were conducted on datasets with various improvements in this section. To ensure the fairness of the comparison experiments, the dataset division and experimental parameters were set consistently. In the ablation study section of our research, we have uniformly set the image input size to 640 × 640 pixels. This size balances preserving detail and the demand for computational resources. It is also a commonly used benchmark size in object detection, facilitating a fair comparison of our results with existing work. The results of these ablation experiments are summarized in table 2.

The first step of the ablation experiment is the NWD metric applied to the YOLOv7-tiny model as the loss function. Experimental results show that this improvement bolsters the mAP@0.5 value from 0.835 to 0.847 without altering the model parameter count and computational workload. Through category-specific AP values, we observed a stable increase in the object detection performance for small-size objects such as vehicles and ships. However, medium and large-size objects (like tennis courts and bridges) did not exhibit a similar trend. The characteristics of the CIoU and NWD metrics explain the observed results. Specifically, the CIoU metric in the YOLOv7-tiny model is highly sensitive to small-sized objects, posing challenges to model convergence during training. On the other hand, the NWD metrics show insensitivity to positional changes in the predicted box for objects across various sizes. While this characteristic aids in identifying small-size objects, it can be adverse for large-size objects. The NWD's lack of size sensitivity might hinder the recognition of subtle distinctions between objects, which can result in inaccuracies in bounding box localization. Immediately following the second step of the ablation experiment, the paper introduced the involution module into the object detection model.

Figure 9. Visualization of partial experimental datasets.

Table 1. The number of instances of different categories.

Category	The number of instances		
	Training	Validation	Testing
Aircraft	2902	155	329
Ship	223	34	79
Storage tank	1614	75	166
Baseball field	298	52	92
Tennis court	411	64	113
Basketball court	104	28	53
Surface track	123	22	40
Port	141	39	83
Bridge	67	24	57
Vehicle	2600	438	1009
Total	8483	931	2021

Compared to the model only applying the NWD metric, integrating the Involution module increases the mAP@0.5 value to 0.875 while significantly reducing the parameter count to 4865 822. Subsequently, adding the RCS-biformer module into the model that includes the NWD metric and Involution module increases the mAP@0.5 value to 0.891, and the number of parameters decreases slightly to 4859 134. Finally, the original feature pyramid SPPCSP is enhanced by incorporating the weighted sum operation approach and CoT module, allowing the mAP@0.5 value to reach 0.903, increasing the parameter count to 4956 688.

Overall, the ablation experiments in the last three steps demonstrated enhanced detection for smaller objects such as vehicles and ships. In contrast, the detection of larger entities like ports or bridges did not exhibit a consistent improvement. This phenomenon can be attributed to the improvements made to the object detection model, which predominantly benefit the detection accuracy of smaller items: as shown in figure 10, the modules of improvements 3 (RCS-Biformer), and 4 (CoT) are designed to strengthen the relationship between the object and its surrounding environmental information and enhance the connections between different divided feature maps, thereby enhancing context understanding. Nevertheless, for medium and large-size objects, this strategy may cause the separation of the whole target, which weakens the features of the object itself.

To further verify the effectiveness of our proposed model structure and to assess in detail the specific impact of each improvement on performance, we introduced additional experiments. These experiments aim to assess the individual and combined contributions of these modules to improving the YOLOv7-tiny model.

Table 3 summarizes the experimental results of YOLOv7-tiny with various combinations of improvements. Among others, experiment 1 serves as a baseline, displaying the performance of the unmodified YOLOv7-tiny model. Experiments 2 through 5 individually tested the improvements NWD, Involution, RCS-Biformer, and WF-CoT SPPCSP, with RCS-Biformer showing the most significant gain in performance as a modification. Experiments 6–10 demonstrated the combination effects of different improvement modules; combining RCS-Biformer and WF-CoT SPPCSP improved model performance but was less than ideal in reducing the model's parameter count. On the other hand, introducing Involution improved model performance while reducing parameters, and the NWD improvement increased performance without adding parameters. Experiments 9 and 10 further confirmed that these

Table 2. Ablation experimental results on the experiment dataset.

		YOLOv7-tiny (baseline)	YOLOv7-tiny + NWD	YOLOv7-tiny + NWD + involution	YOLOv7-tiny + NWD + involution + RCS-biformer	YOLOv7-tiny + NWD + involution + RCS-biformer + WF-CoT SPPCSP
AP@0.5/ @0.5:0.95	Aircraft	0.932/0.564	0.946/0.552	0.944/0.562	0.956/0.619	0.947/0.616
	Ship	0.901/0.487	0.908/0.505	0.924/0.503	0.953/0.526	0.970/0.561
	Storage tank	0.979/0.711	0.996/0.729	0.991/0.721	0.994/0.765	0.991/0.762
	Baseball field	0.984/0.711	0.986/0.710	0.995/0.699	0.990/0.729	0.984/0.734
	Tennis court	0.753/0.375	0.743/0.373	0.805/0.409	0.826/0.479	0.802/0.476
	Basketball court	0.530/0.232	0.571/0.271	0.704/0.375	0.800/0.498	0.866/0.628
	Surface track	0.978/0.724	0.977/0.734	0.992/0.773	0.899/0.670	0.943/0.695
	Port	0.835/0.444	0.871/0.465	0.870/0.447	0.839/0.474	0.845/0.472
	Bridge	0.825/0.334	0.820/0.380	0.851/0.349	0.845/0.435	0.837/0.434
	Vehicle	0.627/0.288	0.653/0.290	0.676/0.314	0.755/0.413	0.769/0.439
	mAP	0.835/0.487	0.847/0.501	0.875/0.515	0.886/0.561	0.895/0.582
Parameter		6031 950	6031 950	4865 822	4859 134	4956 688

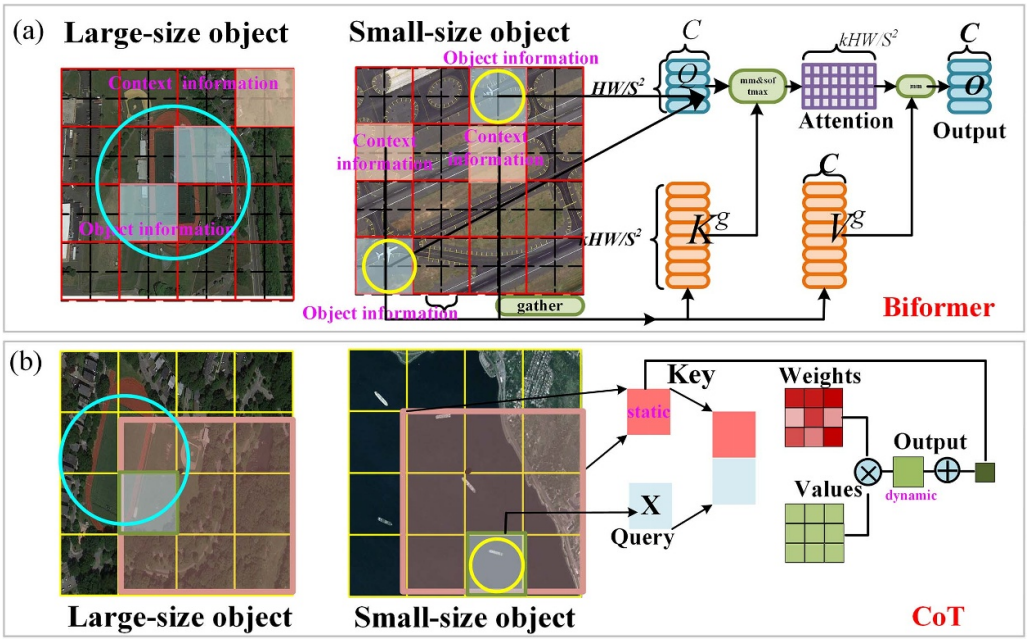

Figure 10. The effect of the improvement module on objects with different sizes: (a) biformer, and (b) CoT.

modules could maintain an improvement in model accuracy while also improving parameter efficiency. Finally, experiment 11, which represents the model proposed in this paper, confirmed that integrating all improvement modules could achieve the best performance.

4.2. Robustness of the AeroDetectNet

The impact of different input image sizes on the training outcomes during the training process of object detection algorithms is different. High-resolution input images typically contain more abundant information, which can aid the model in better learning subtle features. Conversely, low-resolution images may result in the loss of crucial features, affecting the model's performance. In addition, different input sizes significantly affect the required training time and computational resources. High-resolution input images will increase the computational load of each training step, thereby increasing the overall training time and resource consumption. To ensure the reliability of the results, we ran the experiment code multiple times. In this section, we conducted corresponding ablation experiments for the three input image sizes (640 × 640, 512 × 512, and 416 × 416 pixels). Table 4 provides a detailed summary of the ablation study results for different image input sizes.

These experiments are designed to assess the effectiveness of the proposed improvement methods when processing images of different resolutions and the specific impact of each

Table 3. The experimental results of YOLOv7-tiny with various combinations of improvements.

Experiment	YOLOv7-tiny with various combinations of improvements				mAP@0.5/@0.5:0.95	Parameter
	NWD	Involution	RCS-Biformer	WF-CoT SPPCSP		
Experiment 1(baseline)	×	×	×	×	0.835/0.487	6031 950
Experiment 2	√	×	×	×	0.847/0.501	6031 950
Experiment 3	×	√	×	×	0.847/0.508	4865 822
Experiment 4	×	×	√	×	0.858/0.525	6025 262
Experiment 5	×	×	×	√	0.851/0.515	6129 504
Experiment 6	√	√	×	×	0.875/0.515	4865 822
Experiment 7	×	×	√	√	0.870/0.547	6122 816
Experiment 8	√	√	√	×	0.886/0.561	4859 134
Experiment 9	√	×	√	√	0.891/0.568	6122 816
Experiment 10	×	√	√	√	0.888/0.563	4956 688
Experiment 11(AeroDetectNet)	√	√	√	√	0.895/0.582	4956 688

Table 4. The ablation experiment results for different input image sizes.

YOLOv7-tiny with various combinations of improvements	Input size		
	416 × 416	512 × 512	640 × 640
YOLOv7-tiny (baseline)	0.706/0.381	0.813/0.470	0.835/0.487
YOLOv7-tiny + NWD	0.736/0.386	0.822/0.478	0.847/0.501
YOLOv7-tiny + NWD + Involution	0.779/0.430	0.835/0.487	0.875/0.515
YOLOv7-tiny + NWD + Involution + RCS-biformer	0.839/0.474	0.873/0.536	0.886/0.561
YOLOv7-tiny + NWD + Involution + RCS-biformer + WF-CoT SPPCSP (AeroDetectNet)	0.867/0.511	0.885/0.563	0.895/0.582

improvement on model performance. As shown in table 4, the results indicate that our improvements maintain excellent performance across various sizes, highlighting its outstanding robustness and adaptability to varying inputs. In the subsequent experiments of this study, we standardized the image input size to 640 × 640 pixels.

4.3. Gradient-weighted class activation mapping (Grad-CAM) visualization

The Grad-CAM is a technique used for visualizing the key areas within an image that significantly influence the prediction outcome. Figure 11 highlights these critical areas, aiding in comprehending the model's operational method for processing images and making predictions. In this section, several improvement models of YOLOv7-tiny are tested for Grad-CAM visualization in this study. First, the Involution module supplies a self-attention mechanism that grasps refined contextual details, enabling the model to understand the spatial correlations among different areas of interest in an image. Furthermore, the integration of the RCS-biformer module, a proficient transformer-based structure, aids in capturing both local and global contextual information. In addition, introducing the CoT module into the feature pyramid structure module allows the model to focus more on the noteworthy aspects of the object in the image and avoid the interference of irrelevant details. Overall, the results show that the AeroDetectNet model provides a better understanding of the critical regions within the object, leading to more effective extraction of essential features.

4.4. Visualization result

The dataset used for experiments in this paper is a mixed dataset consisting of three open-source datasets. To intuitively demonstrate the improved model's superior performance, representative images from each of the three open-source data were selected for detection. Figures 12–14 shows the detection result of the original YOLOv7-tiny and AeroDetectNet, respectively. As can be seen from the results, AeroDetectNet performs well for small target objects from the three open-source data. Moreover, each dataset is used to validate the effectiveness of the improved model, and the experimental results are summarized in table 5.

4.5. Assessing average detectable object sizes within open-source datasets

This study further assesses the detection precision of AeroDetectNet on various classes of objects within three open-source datasets, particularly in terms of its capability to detect small-size objects. We chose categories from three open-source datasets where the targets represent a small ratio of the image size, including the 'aircraft' categories in NWPU VHR-10, the 'oiltank' category in RSOD, and the 'vehicle'

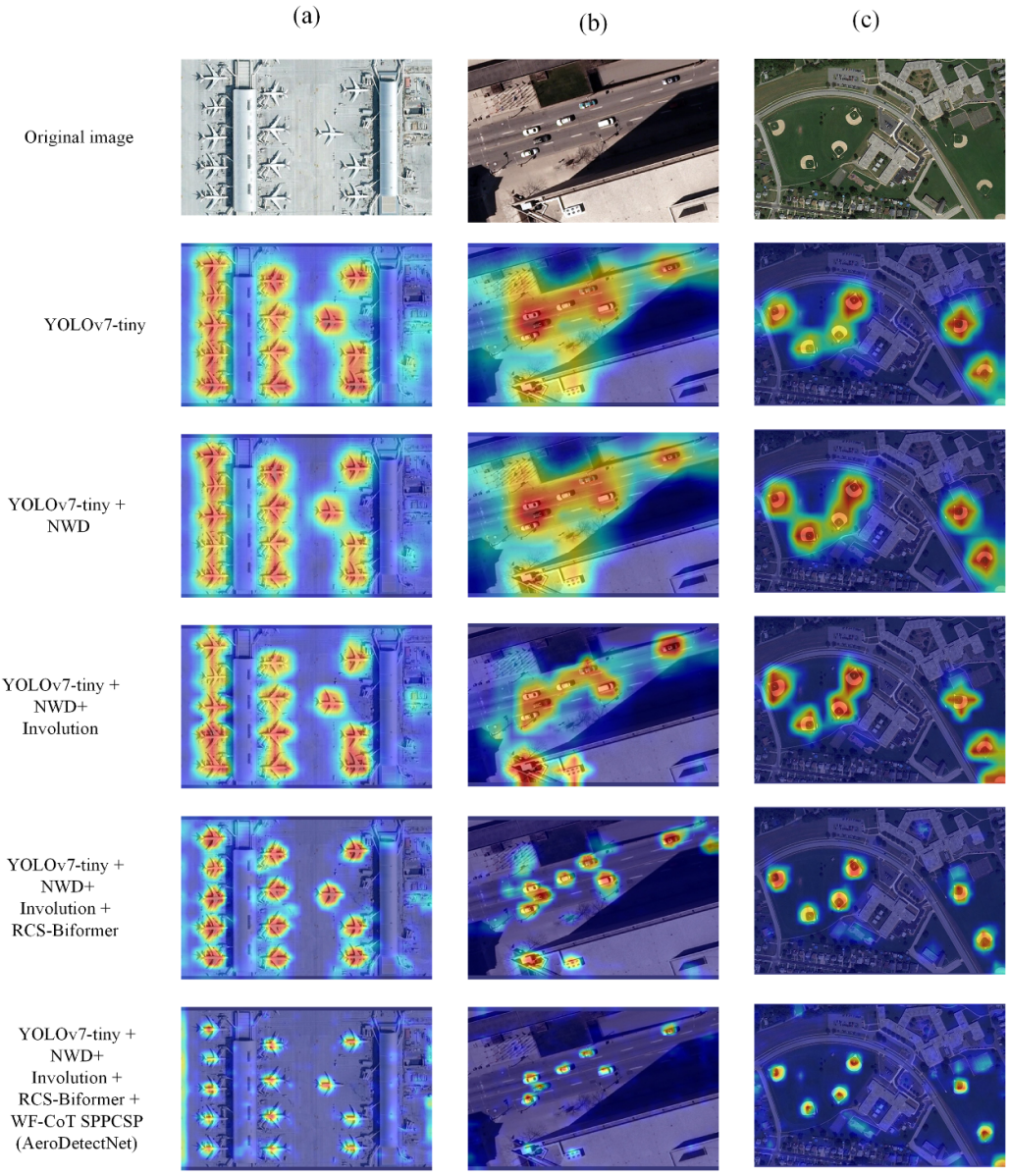

Figure 11. Grad-CAM visualization results.

category in VisDrone. We have carefully recorded the average ratio of the pixel area of all bounding boxes detected by AeroDetectNet to the pixel area of the images. The calculation method can be expressed as:

$$p = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{w_j \times h_j}{W_i \times H_i} \tag{22}$$

where w and h represent the detected bounding box's normalized width and height. W and H represent the absolute pixel values for the width and height of the target boxes that can be detected. n represents the number of images, and m represents the number of detected object boxes per image.

Table 6 shows that AeroDetectNet can detect significantly smaller pixel sizes than YOLOv7-tiny, confirming its superiority in fine-grained detection and demonstrating the model's broad applicability and reliability in practical applications.

4.6. Compare with other models

To more intuitively affirm the superiority of the proposed model, this section undertakes a comparative analysis with other object detection algorithms, including YOLOv5, YOLOX, and YOLOv8. In addition, this study reproduces models from related improvement studies for further comparative analyses, thus highlighting the value of the improved models presented in the current field of research. Uniformity in dataset division and training parameters is maintained for

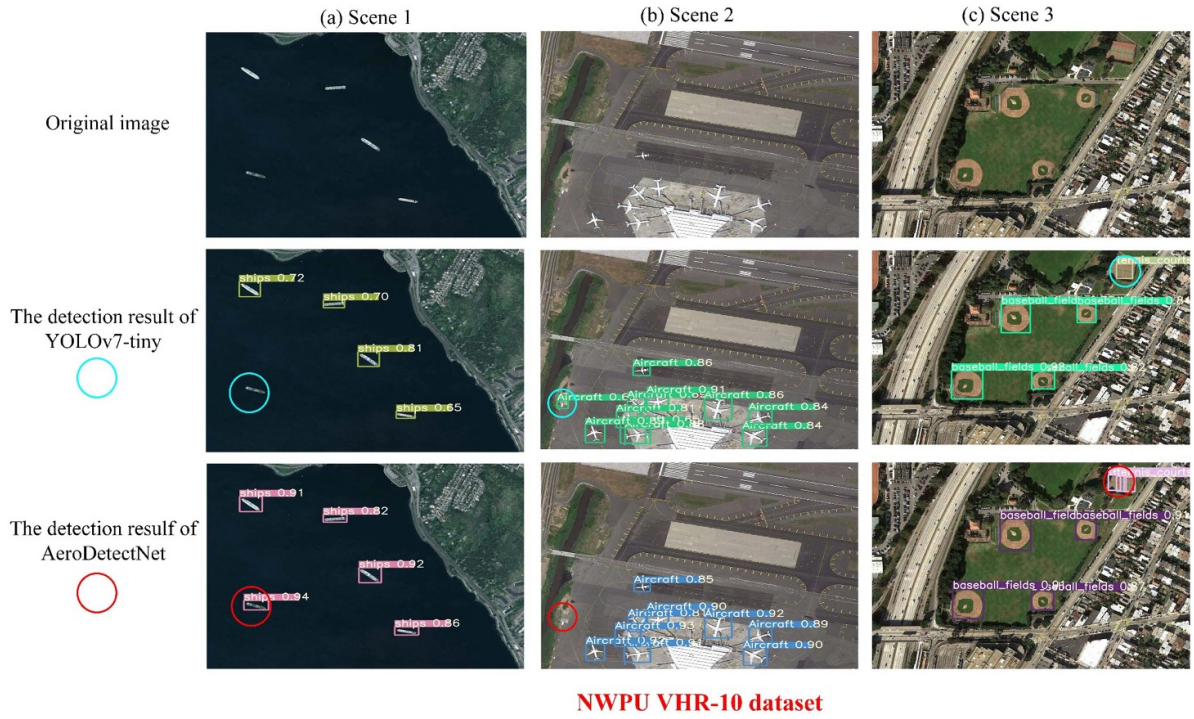

Figure 12. The detection results of the original YOLOv7-tiny and AeroDetectNet (NWPU VHR-10).

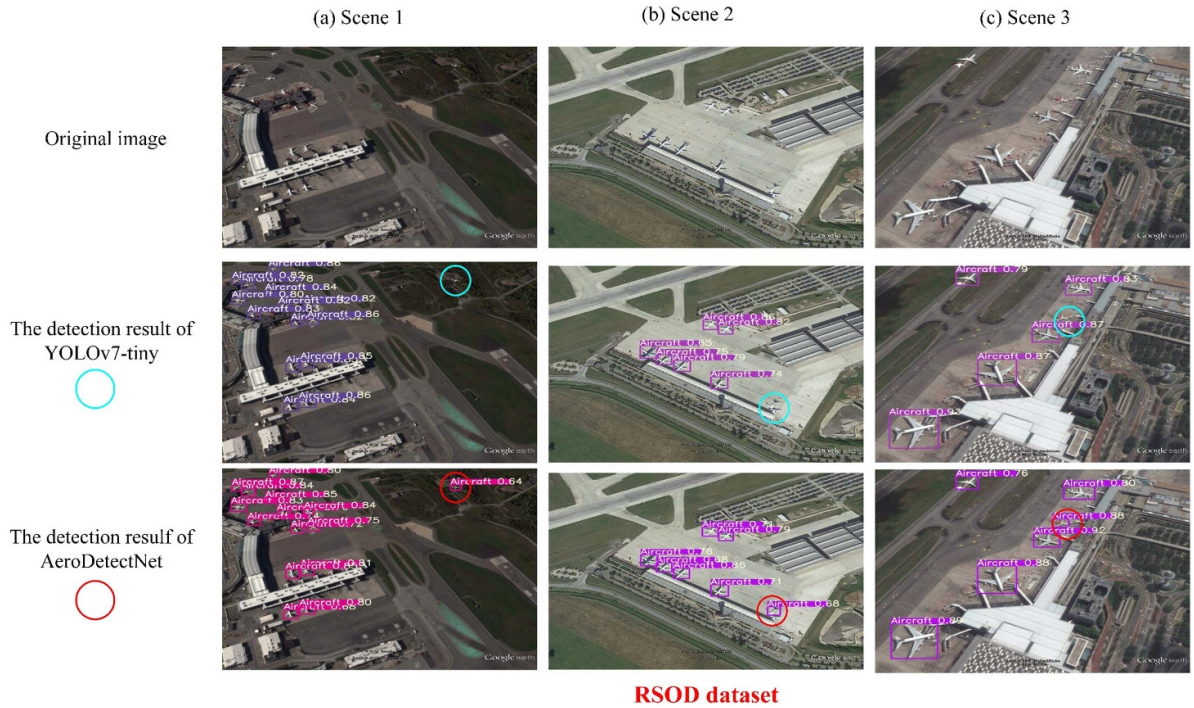

Figure 13. The detection results of the original YOLOv7-tiny and AeroDetectNet (RSOD).

experimental fairness. The detailed comparative results are presented in table 7.

As can be seen from table 7, Cascade RCNN (NWD) and Faster RCNN (NWD) models have high mAP values, exhibiting excellent object detection accuracy. However, the excellent performance comes at the expense of substantial computational resources, reflected in their high parameter and GFLOPs values. In contrast, the proposed AeroDetectNet maintains a high mAP of 0.895 while significantly reducing the parameter and computational complexity, with values of only 4.96 M and 10.3 GFLOPs, respectively. YOLOv8 has a slightly lower mAP than improved YOLOv7-tiny, but its lower parameters and GFLOPs make it suitable for scenarios with strict computational resource constraints. Nonetheless, AeroDetectNet

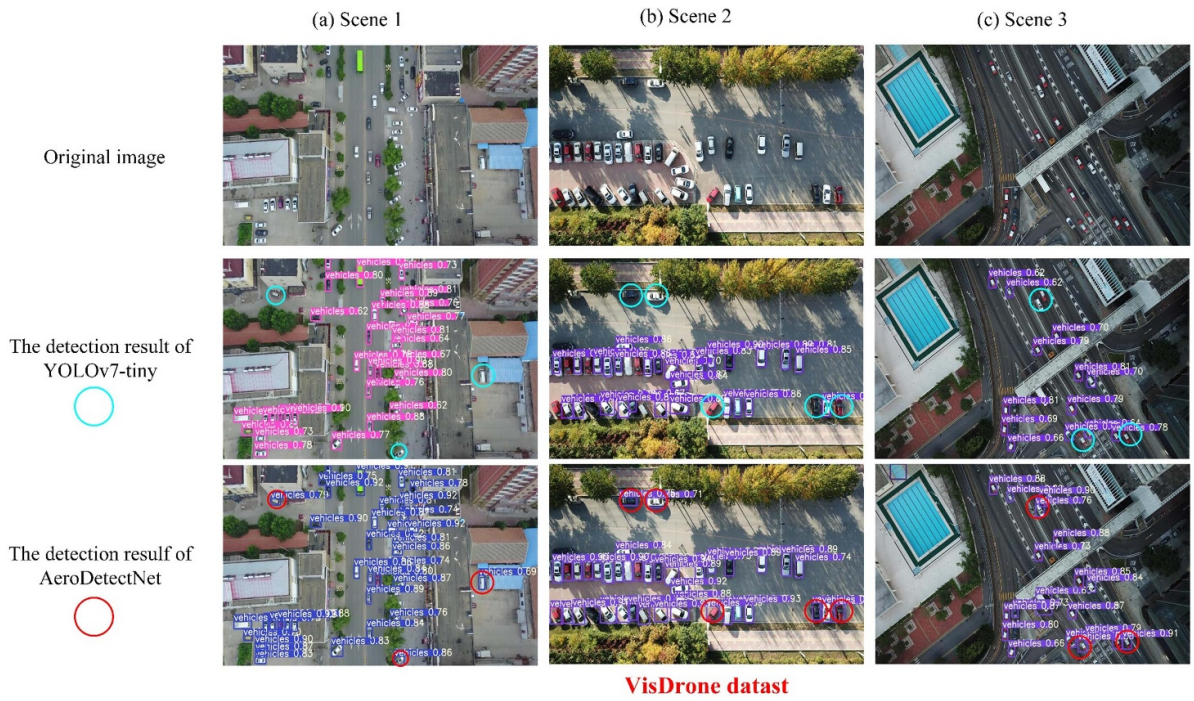

Figure 14. The detection results of the original YOLOv7-tiny and AeroDetectNet (VisDrone).

Table 5. The performance of the AeroDetectNet model on different data (mAP@0.5).

Dataset	Category	YOLOv7-tiny	AeroDetectNet
NWPU VHR-10	Aircraft	0.993	0.995
	Ship	0.846	0.877
	Storage tank	0.826	0.905
	Baseball field	0.994	0.989
	Tennis court	0.841	0.930
	Basketball court	0.579	0.939
	Surface track	0.995	0.952
	Port	0.922	0.921
	Bridge	0.741	0.789
	Vehicle	0.707	0.772
	mAP	0.844	0.907
RSOD	Aircraft	0.967	0.962
	Oiltank	0.975	0.981
	Overpass	0.765	0.853
	Playground	0.968	0.984
	mAP	0.919	0.945
VisDrone	Vehicle (mAP)	0.526	0.705

Table 6. The average pixel size in three open-source datasets that the AeroDetectNet model can detect (p).

Dataset	Category	YOLOv7-tiny	AeroDetectNet
NWPU VHR-10	Aircraft	1.98%	1.71%
RSOD	Oiltank	3.28%	2.97%
VisDrone	Vehicle	2.39%	2.12%

Table 7. Comparison with other models.

Object detection algorithm	mAP@0.5	Parameter (M)	GFLOPs
YOLOv5	0.847	7.04	15.9
YOLOX	0.879	8.04	21.6
YOLOv8	0.876	3.01	8.1
SuperYOLO [23]	0.857	4.83	10.03
PP-YOLOE-SOD [21]	0.885	7.93	17.36
Faster RCNN (NWD) [18]	0.901	41.17	206.71
Cascade RCNN (NWD) [18]	0.913	68.95	234.49
AeroDetectNet	0.895	4.96	10.3

stands out as a more attractive option for application scenarios where accuracy is significant. While PP-YOLO-SOD and YOLOX present competitive mAP, their high parameter counts could be a drawback. SuperYOLO, despite its advantages in terms of model complexity, does not perform as well in terms of detection accuracy, potentially impacting its practicality in real-world applications. Last, the YOLOv5 falls short in accuracy and the number of model parameters and does not compare favorably with other models.

Scatter plots with an x-axis representing mAP@0.5 and a y-axis representing model parameters can be plotted to visually compare the performance of various object detection algorithms. As shown in figure 15, it is found that the AeroDetectNet model demonstrates a balance between model performance and complexity.

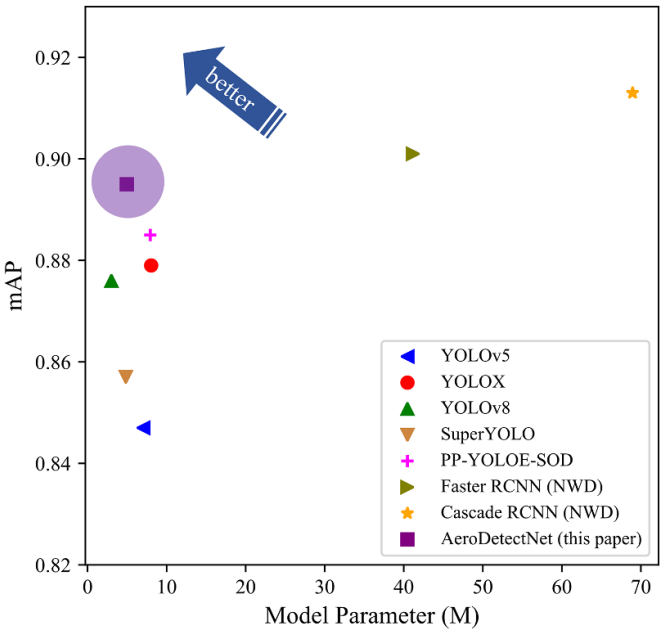

Figure 15. Comparison with other models.

5. Conclusion

Detecting objects in remote sensing images presents considerable challenges due to varied object sizes, intricate spatial context relationships, and complex background information. This paper proposes a model named AeroDetectNet that is highly accurate and lightweight. The model's feature extraction and understanding of spatial contextual relationships are enhanced. The Grad-CAM visualization demonstrated that the AeroDetectNet model effectively emphasizes key regions within objects. Tests on three open-source datasets revealed that the AeroDetectNet model performs well in detecting small-size objects. Compared with other relevant studies, the AeroDetectNet model achieves a competitive mAP while maintaining lower computational complexity and fewer model parameters, showing an outstanding balance between performance and efficiency. Overall, the study provides a novel object detection model named AeroDetectNet with enhanced performance and efficiency for applications such as real-time object detection and resource-limited environments.

Data availability statement

No new data were created or analysed in this study.

Acknowledgments

This study was supported by the Sichuan Province Science and Technology Achievement Transformation Demonstration Project (Grant No. 2022ZHCG0060) and Meteorological Disaster Prediction, Warning and Emergency Management Research Center, Chengdu University of Information Technology (Grant No. ZHYJ23-YB07).

Conflict of interest

The authors declare that they have no conflicts of interest to report regarding the present study.

ORCID iDs

Ruihan Bai https://orcid.org/0000-0003-2518-8268
Jiahui Lu https://orcid.org/0009-0005-1190-5280
Zhiping Zhang https://orcid.org/0009-0008-0452-3147
Mingkang Wang https://orcid.org/0009-0006-5597-2458
Qiang Wang https://orcid.org/0009-0009-2801-378X

References

[1] Yang J, Liu Z, Du W and Zhang S 2023 A PCB defect detector based on coordinate feature refinement *IEEE Trans. Instrum. Meas.* **72** 1–10
[2] Zhang Y, Shu S, Lang X, Liang H, Yu Z and Yang Z 2023 A real-time method for detecting bottom defects of lithium batteries based on an improved YOLOv5 model *Meas. Sci. Technol.* **34** 125149
[3] Liu Y, Han D, Cao R, Guo J and Deng L 2023 Automated vehicle wheelbase measurement using computer vision and view geometry *Meas. Sci. Technol.* **34** 125051
[4] Zhang Q and Hu X 2023 MSFFA-YOLO network: multi-class object detection for traffic investigations in foggy weather *IEEE Trans. Instrum. Meas.* **72** 1–12
[5] Wu Y, Chen P, Qin Y, Qian Y, Xu F and Jia L 2023 Automatic railroad track components inspection using hybrid deep learning framework *IEEE Trans. Instrum. Meas.* **72** 1–15
[6] Guan L, Jia L, Xie Z and Yin C 2022 A lightweight framework for obstacle detection in the railway image based on fast region proposal and improved YOLO-tiny network *IEEE Trans. Instrum. Meas.* **71** 1–16
[7] Xu S, Wang J, Shou W, Ngo T, Sadick A M and Wang X 2021 Computer vision techniques in construction: a critical review *Arch. Comput. Methods Eng.* **28** 3383–97
[8] Martinez P, Al-Hussein M and Ahmad R 2019 A scientometric analysis and critical review of computer vision applications for construction *Autom. Constr.* **107** 102947
[9] Kisantal M, Wojna Z, Murawski J, Naruniec J and Cho K 2019 Augmentation for small object detection (arXiv:1902.07296)
[10] Chen Y, Zhang P, Li Z, Li Y, Zhang X, Qi L, Sun J and Jia J 2020 Dynamic scale training for object detection (arXiv:2004.12432)
[11] Romano Y, Isidoro J and Milanfar P 2016 RAISR: rapid and accurate image super resolution *IEEE Trans. Comput. Imaging* **3** 110–25
[12] Bai Y, Zhang Y, Ding M and Ghanem B 2018 Sod-mtgan: small object detection via multi-task generative adversarial network *Proc. European Conf. on Computer Vision (ECCV)* pp 206–21
[13] Li J, Liang X, Wei Y, Xu T, Feng J and Yan S 2017 Perceptual generative adversarial networks for small object detection *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* pp 1222–30
[14] Wang J, Xu C, Yang W and Yu L 2021 A normalized Gaussian Wasserstein distance for tiny object detection (arXiv:2110.13389)
[15] Ren Y, Zhu C and Xiao S 2018 Small object detection in optical remote sensing images via modified faster R-CNN *Appl. Sci.* **8** 813

[16] Cui L, Lv P, Jiang X, Gao Z, Zhou B, Zhang L, Shao L and Xu M 2020 Context-aware block net for small object detection *IEEE Trans. Cybern.* **52** 2300–13
[17] Xu S *et al* 2022 PP-YOLOE: an evolved version of YOLO (arXiv:2203.16250)
[18] Zhang J, Lei J, Xie W, Fang Z, Li Y and Du Q 2023 SuperYOLO: super resolution assisted object detection in multimodal remote sensing imagery *IEEE Trans. Geosci. Remote Sens.* **61** 1–15
[19] Tan S, Yan J, Jiang Z and Huang L 2021 Approach for improving YOLOv5 network with application to remote sensing target detection *J. Appl. Remote Sens.* **15** 036512
[20] Zhou Q, Zhang W, Li R, Wang J, Zhen S and Niu F 2022 Improved YOLOv5-S object detection method for optical remote sensing images based on contextual transformer *J. Electron. Imaging* **31** 043049
[21] Shen L, Lang B and Song Z 2023 CA-YOLO: model optimization for remote sensing image object detection *IEEE Access* **11** 125122–37
[22] Li J, Wei Y, Liang X, Dong J, Xu T, Feng J and Yan S 2016 Attentive contexts for object detection *IEEE Trans. Multimedia* **19** 944–54
[23] Wu W, Liu H, Li L, Long Y, Wang X, Wang Z, Li J and Chang Y 2021 Application of local fully convolutional neural network combined with YOLO v5 algorithm in small target detection of remote sensing image *PLoS One* **16** e0259283
[24] Ji S J, Ling Q H and Han F 2023 An improved algorithm for small object detection based on YOLO v4 and multi-scale contextual information *Comput. Electr. Eng.* **105** 108490
[25] Liang Z, Shao J, Zhang D and Gao L 2018 Small object detection using deep feature pyramid networks *Advances in Multimedia Information Processing–PCM 2018: 19th Pacific-Rim Conf. on Multimedia* (*Hefei, China*, *21–22 September 2018*) (*Proc. Part III 19*) pp 554–64
[26] Benjumea A, Teeti I, Cuzzolin F and Bradley A 2021 YOLO-Z: improving small object detection in YOLOv5 for autonomous vehicles (arXiv:2112.11798)
[27] Li C, Cong R, Hou J, Zhang S, Qian Y and Kwong S 2019 Nested network with two-stream pyramid for salient object detection in optical remote sensing images *IEEE Trans. Geosci. Remote Sens.* **57** 9156–66
[28] Chen G, Wang H, Chen K, Li Z, Song Z, Liu Y, Chen W and Knoll A 2020 A survey of the four pillars for small object detection: multiscale representation, contextual information, super-resolution, and region proposal *IEEE Trans. Syst. Man Cybern.* A **52** 936–53
[29] Ma J, Huang S, Jin D, Wang X, Li L and Guo Y 2024 LA-YOLO: an effective detection model for multi-UAV under low altitude background *Meas. Sci. Technol.* **35** 055401
[30] Wang Z, Guo J, Huang W and Zhang S 2021 High-resolution remote sensing image semantic segmentation based on a deep feature aggregation network *Meas. Sci. Technol.* **32** 095002
[31] Lu X, Ji J, Xing Z and Miao Q 2021 Attention and feature fusion SSD for remote sensing object detection *IEEE Trans. Instrum. Meas.* **70** 1–9
[32] Gu Q, Huang H, Han Z, Fan Q and Li Y 2024 GLFE-YOLOX: global and local feature enhanced YOLOX for remote sensing images *IEEE Trans. Instrum.* **73** 1–12
[33] Hui Y, Wang J and Li B 2024 STF-YOLO: a small target detection algorithm for UAV remote sensing images based on improved SwinTransformer and class weighted classification decoupling head *Measurement* **224** 113936
[34] Chen Z, Yang J, Feng Z, Chen L and Li L 2023 BiShuffleNeXt: a lightweight bi-path network for remote sensing scene classification *Measurement* **209** 112537
[35] Wang C Y, Bochkovskiy A and Liao H Y M 2023 YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* pp 7464–75
[36] Li D, Hu J, Wang C, Li X, She Q, Zhu L, Zhang T and Chen Q 2021 Involution: inverting the inherence of convolution for visual recognition *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* pp 12321–30
[37] Zhu L, Wang X, Ke Z, Zhang W and Lau R W 2023 BiFormer: vision transformer with Bi-level routing attention *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* pp 10323–33
[38] Li Y, Yao T, Pan Y and Mei T 2022 Contextual transformer networks for visual recognition *IEEE Trans. Pattern Anal.* **45** 1489–500