Artificial Intelligence Review (2025) 58:162
https://doi.org/10.1007/s10462-025-11150-9

A survey of small object detection based on deep learning in aerial images

Wei Hua[1] · **Qili Chen**[1]

Accepted: 11 February 2025 / Published online: 15 March 2025

Abstract
Small object detection poses a formidable challenge in the field of computer vision, particularly when it comes to analyzing aerial remote sensing images. Despite the rapid development of deep learning and significant progress in detection techniques in natural scenes, the migration of these algorithms to aerial images has not met expectations. This is primarily due to limitations in imaging acquisition conditions, including small target size, viewpoint specificity, background complexity, as well as scale and orientation diversity. Although the increasing application of deep learning-based algorithms to overcome these problems, few studies have summarized the optimization of different deep learning strategies used for small target detection in aerial images. Therefore, this paper aims to explore the application of deep learning methods for small object detection in aerial images. The primary challenges in small object detection in aerial images will be summarized. Next, a meticulous analysis and categorization of the prevailing deep learning optimization strategies employed to surmount the challenges encountered in aerial image detection is undertaken. Following that, we provide a comprehensive presentation of the object detection datasets utilized in aerial remote sensing images, along with the evaluation metrics employed. Additionally, we furnish experimental data pertaining to the currently proposed detection algorithms. Finally, the advantages and disadvantages of various optimization strategies and potential development trends are discussed. Hopefully, it can provide a reference for researchers in this field.

Keywords Deep learning · Object detection · Small object · Aerial images · Remote sensing

✉ Qili Chen
qilichen@hotmail.com

Wei Hua
h13797060207@163.com

[1] Beijing Information Science and Technology University, 12 Xiaoyingdonglu, Beijing 100192, China

1 Introduction

Unmanned Aerial Vehicles (UAVs) and remote sensing technologies have been evolving constantly. They enable the execution of a wider range of tasks by leveraging the combined power of their visual perception capabilities. Intelligent analysis and processing of aerial images allow for the quick and efficient capture of target feature information, which can enhance the scene understanding capabilities of UAVs. Target detection technology automatically identifies and locates targets in images, which can improve the perception function of UAVs under weak human–machine interaction and provide basic technical support for autonomous detection and flight. Target detection on aerial images has garnered extensive research attention in civilian and military fields, such as UAV reconnaissance (Zhou et al. 2022a), traffic supervision (Khosravi et al. 2023), precision agriculture (Roy and Bhaduri 2022), wildlife tracking (Feng and Xiao 2022), and personnel rescue (Ren et al. 2022).

Recently, deep learning methods have rapidly developed in target detection and have achieved good results in many fields. The deep convolutional neural network(CNN) possesses the capability to extract intricate and comprehensive image features, encompassing both fine-grained details and higher-level semantic information. By employing the gradient descent optimization method, it is capable of iteratively refining its parameters to identify the optimal solution for detecting the target. Unlike traditional target detection algorithms, CNN-based detection algorithms efficiently learn image features in data and possess good robustness and detection performance.

Currently, target detection mainly focuses on natural scene images, which are relatively mature in corresponding application problems, such as face recognition and pedestrian detection. However, due to different imaging perspectives and a lack of effective samples for training, directly applying existing algorithms to aerial remote sensing images results in poor performance. Therefore, it is of great significance to study target detection algorithms applicable to aerial images for their applications.Therefore, it is of paramount importance to investigate and propose optimization algorithms that can effectively address the challenges associated with target detection in aerial imagery. Such endeavors hold significant implications for their widespread and efficient application in this domain.

Aerial images are different from natural scene images, and usually the targets to be detected are small in size, large in scale variation, occlusion, arbitrary orientation, and dense distribution of targets, which are the challenges for aerial image detection. Figure 1 provides a comparative analysis illustrating the disparities in target characteristics between aerial imagery and general detection images. In large-field-of-view aerial images with substantial scale, the targets often occupy only a small fraction of the image's pixel values. Even state-of-the-art general detection algorithms struggle to accurately locate these targets amidst complex backgrounds. This paper focuses on an overview of the challenges encountered in small target detection in aerial images, along with a categorical overview and summary of detection algorithms to address these challenges, and finally an evaluation and comparison of detection performance.

The main contributions of this paper are as follows.

- Firstly, this study presents a comprehensive summary of the primary challenges encountered by detection algorithms in the realm of small target detection in aerial images, based on an in-depth analysis of the data characteristics. Subsequently, these challenges

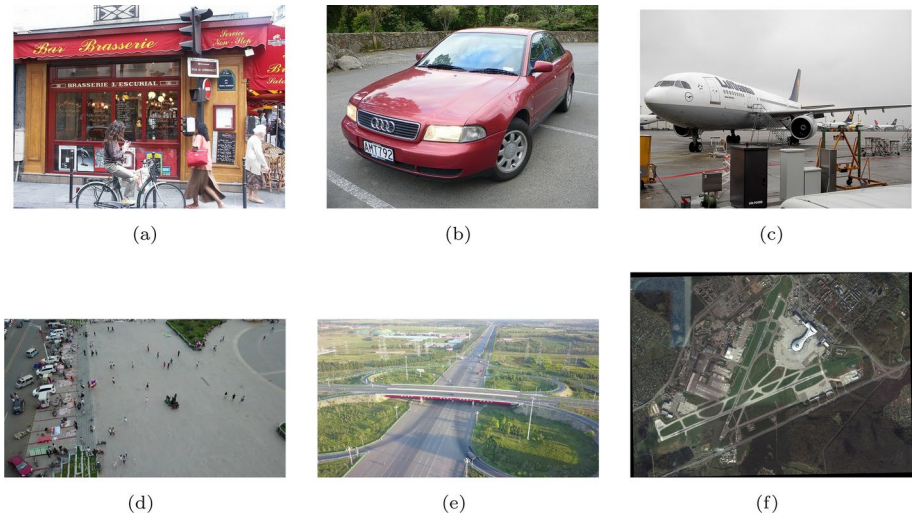

Fig. 1 While the targets in ground-based photographs **a**–**c** take up a relatively large area of the image and are distributed in a certain direction, in aerial images, **d–f** the targets are generally much smaller and denser and show an arbitrary directional distribution

serve as crucial guidelines for directing future research efforts aimed at enhancing the performance and efficacy of these algorithms.

- Drawing upon an extensive range of references, this research delves into the domain of small target detection in aerial remote sensing images, focusing specifically on the application of deep learning methods. By thoroughly examining the primary challenges encountered by deep learning techniques in this context, the literature review analysis is categorized into five distinct areas. Furthermore, this study consolidates and presents the innovative concepts derived from various enhanced algorithms, while elucidating the distinguishing characteristics of each algorithm.
- This study meticulously collects and organizes the presently prevalent datasets utilized for small target detection in aerial images. Additionally, a comprehensive comparison of experimental results obtained by various algorithms on these primary datasets is presented, facilitating a concise and prompt comprehension of the latest advancements and progress achieved in the field of aerial image target detection across each dataset.
- Through a comprehensive analysis of the existing literature, this study synthesizes the strengths and weaknesses of current algorithms. Furthermore, it identifies potential areas for future research and suggests directions for further investigation in the field.

The remainder of this paper is organized as follows. Section II explains the scope of this paper review and the challenges faced in small target detection in aerial images. Section III describes the progress of research work on small target detection using deep learning methods for these challenges. Section IV provides statistical data and evaluation metrics for the currently dominantly used datasets and analyzes and compares the experimental results of the latest network models based on deep learning methods. Section V summarizes the work of this paper and suggests future research directions.

2 Research methodology

The articles selected in this paper are mostly published within the last three years, and a few articles are based on classic articles published between 2010 and 2015. Relevant literature was searched mainly from databases such as Elsevier, IEEE Xplore, and Science direct. The method of literature collection in this paper involves keyword searches such as remote sensing images, object detection, drones, and small object detection. Subsequently, by using the collected benchmarks, the latest research citing these references was identified. In terms of research content, this paper focuses on two aspects to review the detection of small targets in air-to-ground images. On the one hand, we summarize the challenges of target detection in air-to-ground remote sensing images compared with natural images, i.e., the main reasons why detection algorithms in natural images are not effective when transposed to air-to-ground images; on the other hand, we analyze them from the perspective of solving the difficulties faced by detection tasks in aerial images. The architecture of this review is shown in Fig. 2.

Currently, the majority of reviews on deep learning-based detection algorithms (Zou et al. 2023) primarily summarize general datasets. In recent years, with the deepening of research, an increasing number of fields are employing deep learning methods to tackle challenging issues within their domains (Er et al. 2023). Cheng et al. (2023b) analyzes and compares scenarios and algorithms for small object detection. In the remote sensing

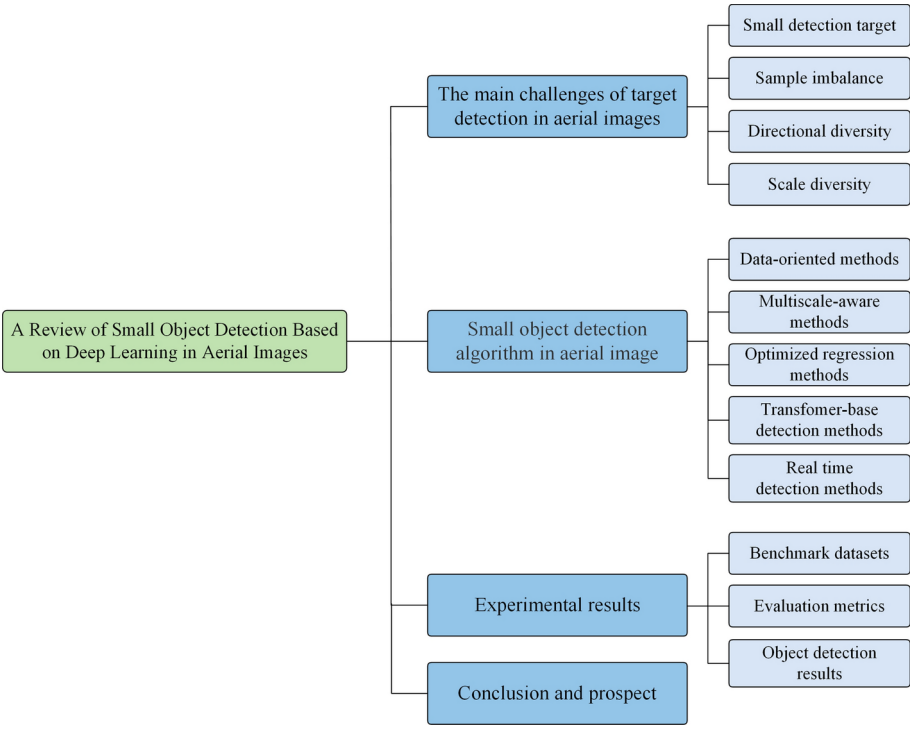

Fig. 2 Architecture of the survey. In five sections, the challenges of target detection in aerial imagery, targeted improvement strategies in existing frameworks, statistics of datasets, algorithm performance comparison, difficulties, and potential directions for future development are presented

image domain, DOTAv1 (Xia et al. 2018) reviews the object detection algorithms for remote sensing images, and Ding et al. (2022); Xia et al. (2018) provides high-quality datasets to advance this field. On the other hand, Fu et al. (2023); Jain et al. (2021) conducts a comparative analysis of target detection algorithms for drones, listing the algorithms for addressing various issues and their practical effects. However, these reviews independently discuss and analyze the target detection methods for remote sensing images and drone aerial images. Nonetheless, the characteristics of detection targets and the challenges posed by the image data in both remote sensing images and drone aerial images are significantly related. This paper consolidates the common challenges faced in both fields, with a focus on small object detection from an aerial perspective. It summarizes the research advancements on these issues in both domains, offering broader guidance to researchers on the target problems and potential solutions.

3 The main challenges of target detection in aerial images

The features of targets contained in aerial images due to different shooting angles are apparently different from typical images taken by ground-based cameras, as shown in the following figure comparing the two shooting angles. Generally, the targets in aerial images are tiny, so they do not provide enough visual features for the target detection algorithm to learn, which reduces the detection performance. Moreover, targets in aerial images usually exhibit rotational multi-angle characteristics, which are significantly different from the distribution of the target's pose in natural images in the image.

Compared with natural images, the air-to-ground images have a wider field of view, which means that the detection task is exposed to more complex background interference. In addition, the lack of contextual information for small targets leads to the fact that the features of a single convolutional layer do not contain enough information for detection. The target detection model developed so far performs well on the COCO (Lin et al. 2014) and PASCAL VOC (Everingham et al. 2010) datasets, which contain mainly medium and large targets and few small targets. As a result, the general detectors designed based on this dataset exhibit a bias towards detecting larger targets. Even advanced detection algorithms, when directly applied to the task of detecting small objects in air-to-ground imagery, fail to achieve comparable performance as observed in previous tasks.

3.1 Small detection target

The definition of a small target lacks a standardized criterion to precisely specify the desired scale or size of an object within an image. The definition of metric evaluation given for small targets starts with the fact that in the COCO (Lin et al. 2014) dataset, objects with an area smaller than 32×32 pixels are classified as small objects. The detection of small targets in natural images usually continues with this size limitation. However, the pixel values of small targets vary with different image resolutions. Therefore, RCNN (Chen et al. 2017) introduced the mean relative overlap metric as a means to define small targets. This metric quantifies the overlap area between the bounding box and the image, and it has gained widespread adoption within the realm of small target detection research.

In the aerial image dataset DOTA (Xia et al. 2018), targets encompassing pixel values ranging from 10 to 50 pixels are specifically classified as small targets. For the pedestrian datasets Tiny Person (Yu et al. 2020) and Manipal-UAV (Akshatha et al. 2023) datasets in aerial images, targets with resolutions between 20 pixels and 32 pixels are defined as small targets, and targets with pixel values between 2 pixels and 20 pixels are defined as tiny targets.The MOHR (Zhang et al. 2021a) dataset calculates the percentage of all bounding boxes, relative to the ratio of covered pixels to the original image size. Objects covering less than or equal to 0.05% of the pixels can be considered as tiny objects and objects less than 0.5% as small objects.

Small target objects cover a smaller area in the image with pixel values between tens of pixels or even a few pixels, their resolution is lower, feature information covers less, and they lack feature expression capability. After research, the main reasons for the lower accuracy of small target objects in the detection process are as follows:

Less feature information. In the commonly used small target dataset, small target samples have low resolution, small percentage of labeled area, contain inconspicuous feature information, and are susceptible to interference from noise points, which in turn leads to the model's inability to accurately locate small targets. Small objects often exhibit characteristics such as low resolution and sparse spatial distribution, making it challenging to extract effective semantic features for accurate localization. Furthermore, the regional features of small targets are prone to being overwhelmed by complex backgrounds, introducing additional noise into the learned features. In summary, the feature representations of small targets lack robustness and are highly susceptible to noise, thereby impeding subsequent detection processes.

As the CNN deepens and widens, resulting in a larger downsampling rate, the problem of information loss arises for small targets. During the detection process, the scale of the output feature map undergoes a progressive reduction due to consecutive downsampling operations and feature extraction. This can potentially result in a downsampling step that exceeds the size of the small target, causing the feature information associated with the small target to be lost within the propagated feature map.

The cross-merge ratio threshold is not set reasonably. The current matching strategy of most detectors is to use the Intersection over Union (IoU) between the generated Bounding Box and Ground Truth to divide the positive and negative samples. Generally, the target in the Anchor Box corresponding to IoU ≥ 0.5 between Bounding Box and Ground Truth is set as positive samples, and the rest are negative samples. The self-defined threshold will have a great impact on the selection of positive and negative samples, and this matching method is more suitable for large and medium target samples, which is prone to the problem of less matching for small target samples and more matching for large and medium target samples.

3.2 Sample imbalance

The distribution of positive and negative samples in the dataset is uneven. The majority of target detection datasets contain a relatively small number of small target samples, while the number of large and medium target samples is predominant. The detection datasets PASCAL VOCVOC (Everingham et al. 2010) and MSCOCO (Lin et al. 2014) in the generic natural images have only 14% and 43% of small targets with pixel values between 10 and 50. In contrast, small targets in the detection dataset DOTA (Xia et al. 2018) of aerial images

occupy 57% of all instances, as shown in the Table 1 of statistical results. Therefore, during the training process, the general detection model tends to prioritize large and medium target samples while neglecting small target samples. Consequently, the small target samples are sparsely represented in the training set, impeding the network's ability to effectively learn and adapt to datasets containing a significant number of small targets.

There are two main reasons why unevenly distributed objects make detection inefficient: (1) Small targets occupy too few pixel values compared to large areas of background and noise, and the proportion of effective features extracted by the network is too small. (2) The detection objects are not evenly distributed in the image, leading to inefficient detection. These two factors also lead to the network tends to be more biased to learn the large background area, while the regions that account for a low proportion of the image area, which have a significant impact on the final detection effect, are not sufficiently extracted features. At the same time, the large area of complex backgrounds leads to a lack of effective contextual information for sparsely distributed small targets.

3.3 Directional diversity

Aerial remote sensing images are captured from elevated vantage points, whereas natural images are typically acquired from ground-level perspectives. Due to the difference in the shooting angle, the object in the aerial remote sensing image appears in a different form in the image than the object in the natural image taken from the ground. And, the shooting angle of the UAV or satellite may change, causing the target to appear in a rotated form in the image. Targets in an image are usually distributed in various directions close to each other. The utilization of horizontal bounding boxes(HBB) in general detection algorithms is suitable for detecting horizontally distributed objects in natural images captured from ground-level perspectives. However, employing HBB as regression boxes for targets in aerial images may lead to the inclusion of unnecessary redundant background feature information during the extraction of features for objects with arbitrary orientations. Therefore, new bounding boxes need to be designed for target detection algorithms in aerial remote sensing images to accommodate the rotational multi-angle nature of the targets. Oriented bounding boxes (OBB) annotation is provided in DOTA (Xia et al. 2018) dataset, OBB can better enclose objects and distinguish crowded objects. The Fig. 3 shows the differences between two annotation methods in the DOTA dataset (Xia et al. 2018). It can be seen that the HBB annotation method includes a significant amount of background when annotating aerial images. Additionally, for dense targets, there is overlapping, which can affect feature extraction and lead to decreased detection accuracy.

Table 1 Target size distribution statistics in natural image dataset and aerial remote sensing dataset

Dataset	10–50 Pixel	50–300 Pixel	Above 300 pixel
PASCAL VOC (Everingham et al. 2010)	0.14	0.61	0.25
MSCOCO (Lin et al. 2014)	0.43	0.49	0.08
DOTA (Xia et al. 2018)	0.57	0.41	0.02

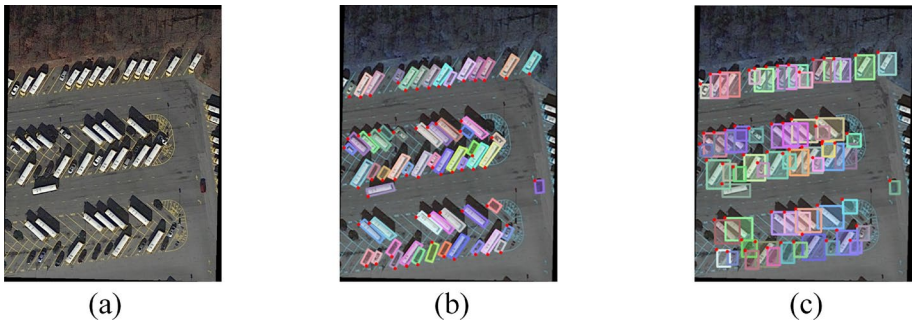

Fig. 3 **a** Is the original remote sensing image, **b** is the labeling method using OBB, **c** is the labeling method using HBB

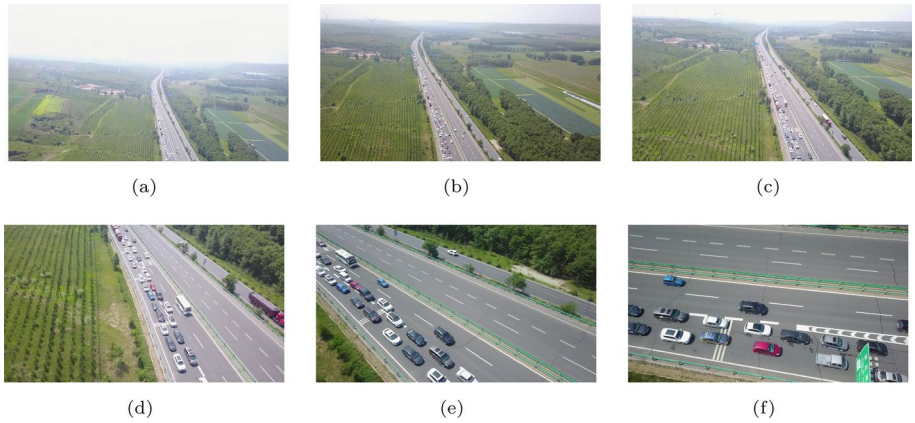

Fig. 4 Images **a**–**f** taken from various altitudes (from far to near)

3.4 Scale diversity

Aerial images are typically captured from a top-down perspective using satellites or unmanned aerial vehicles (UAVs). At the same time, since the imaging devices collect information about the ground target from different altitudes, the target appears to have multi-scale variation in the images captured at different altitudes.As shown in Fig. 4.The AU-AIR (Bozcan and Kayacan 2020) dataset comprises a collection of images obtained from various flight altitudes ranging from 5 to 30 m and different camera angles spanning from 45 to 90 degrees. The MOHR (Zhang et al. 2021a) dataset was acquired at altitudes of 200 m and 300 m above ground level, respectively. With images having a resolution of 8688 × 5792, the identification of small objects necessitates zooming in to twice the size of the original image. In the context of the detection task, the features pertaining to the target are characterized by scale and exhibit a greater range of inter-class similarity.

4 Small object detection algorithm in aerial image

The current state-of-the-art CNN-based target detection algorithms can be broadly classified into two categories: (1) regression-based single-stage target detection algorithms. (2) Two-stage target detection algorithms based on candidate region recommendation. The two-stage target detection algorithm exhibits several advantages. Firstly, the network generates multiple candidate boxes to provide coarse predictions of target positions, which are then refined through training to achieve high detection accuracy and precise localization. However, it is important to consider the trade-offs associated with this approach. Due to the complexity of the algorithm model and the requirement of performing the detection in two sequential steps, the detection speed is relatively slower. The R-CNN (Chen et al. 2017) algorithm is the first to apply deep learning methods to target detection. It uses a selective search algorithm to generate candidate frames, and then uses deep convolutional networks to extract features from Region Proposals, and then performs SVM classification and regression of bounding boxes. Based on this, algorithms such as SPP-Net (He et al. 2015), Fast R-CNN (Girshick 2015) and Faster R-CNN (Ren et al. 2015) have emerged to further improve R-CNN.

The one-stage algorithm, exemplified by the YOLO series and SSD series, eliminates the time-consuming steps of region proposal and recommendation. This significant improvement in algorithmic efficiency enables real-time detection and meets the demand for rapid detection. In 2016, Redmon et al. (2016) proposed the YOLOv1 algorithm to directly use regression for object classification and candidate frame prediction, which simplifies the network structure and significantly enhances the detection speed. Since then, optimization methods based on the YOLO framework have emerged, laying the foundation framework for real-time detection. Relatively, Liu et al. (2016) introduced the SSD algorithm, which incorporates the concept of multi-scale detection. This innovative approach successfully addresses the limitations of the YOLO algorithm, specifically the challenges related to low localization accuracy and the detection of small targets. Through the integration of multi-scale detection, the SSD algorithm overcomes these difficulties and achieves significant improvements in performance. Subsequently, Fu et al. (2017) proposed the DSSD algorithm, which further improved the detection accuracy of small targets. However, the detection speed of the algorithm decreases as the complexity of the network increases. The FSSD (Li and Zhou 2017) algorithm, inspired by the FPN (Lin et al. 2017a) architecture, adopts a similar approach of integrating multi-scale features with information fusion. This integration enhances the detection capabilities of the algorithm. Although there may be a slight decrease in the detection accuracy of small targets, the FSSD algorithm compensates by achieving a notable improvement in detection speed. By leveraging the advantages of the FPN algorithm and incorporating multi-scale feature fusion, the FSSD algorithm strikes a favorable balance between accuracy and efficiency.

In addition to the two types of detection algorithms mentioned above, several anchor-free methods have emerged in recent years, which, in turn, completely abandon the anchor-based mechanism and propose a new method based on key point detection. The CornerNet (Law and Deng 2018) algorithm, introduced by Law et al., and the CenterNet (Duan et al. 2019a) algorithm, proposed by Duan et al., employ the detection of corner or center points of objects to generate bounding boxes, thereby significantly enhancing the speed of detection. These network architectures have garnered considerable attention and adoption within the

remote sensing domain, where they have been extensively researched and applied.Due to the differences between natural images and aerial images, to address the challenges summarized in Chapter 3, existing detection algorithms in aerial images will add specificity to the current robust generic detection models designed. Next, we will outline these approaches in categories based on the challenges addressed, as shown in Fig. 5

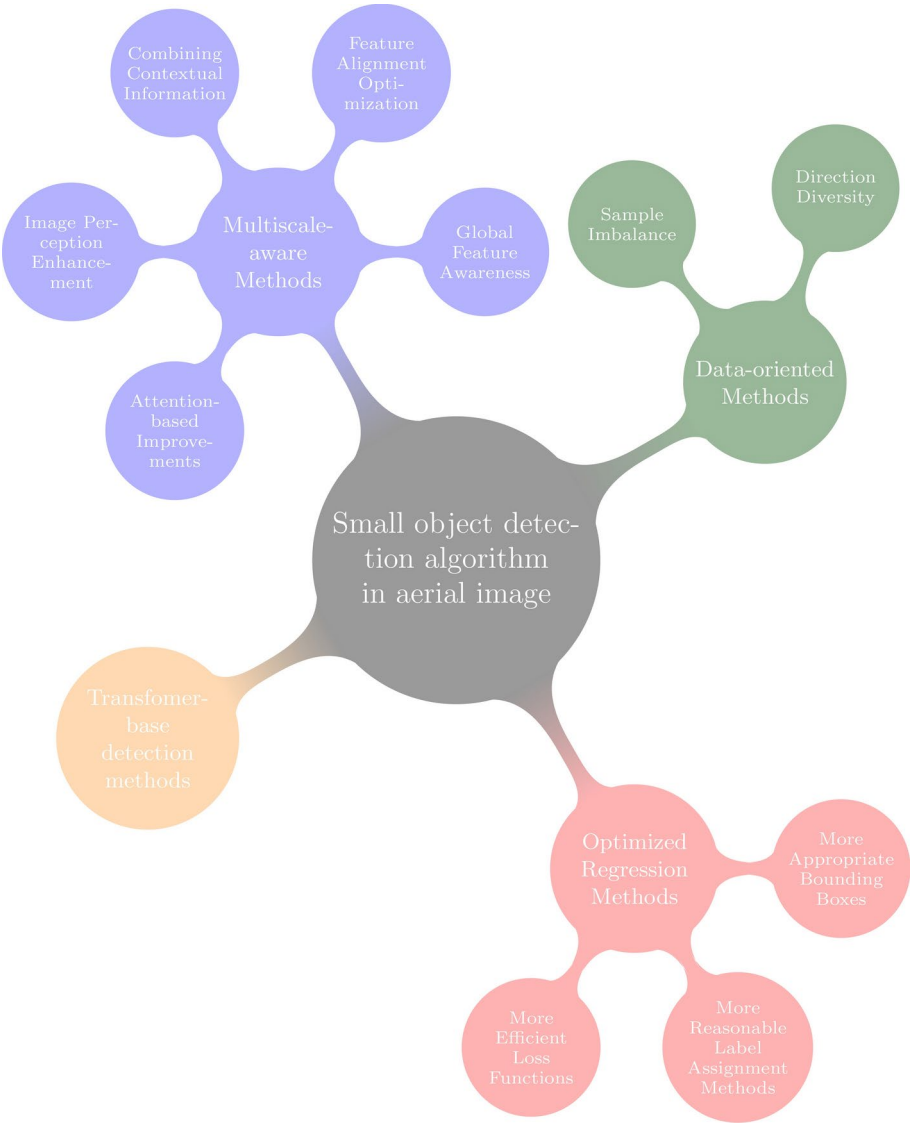

Fig. 5 Optimization strategies for detection networks around the previously summarized challenges

4.1 Data-oriented methods

The performance of deep learning-based detectors is closely linked to the characteristics of the image data itself. A fundamental reason why general-purpose detectors fall short in aerial images is due to the dissimilarities between natural and aerial images. Targets in aerial images typically occupy fewer pixels, which poses challenges for representing effective features, and they can appear at multiple angles, further complicating detection. Moreover, the unbalanced distribution of samples between targets and background in complex scenes may lead to noise and background features overwhelming effective features. To address these challenges, it is necessary to develop methods that account for these data characteristics and improve detection accuracy.

Currently, most deep learning research is still data-driven, and the quantity and quality of the data directly relate to the optimal performance of network learning. In object detection training, using data augmentation methods is a common and effective way to enrich datasets. However, in aerial image datasets, most existing data augmentation methods fail to accurately extract domain-specific features. LGNet (Liu et al. 2024a) introduced a language model that utilizes pre-trained multimodal representations as structural references to guide UAV detection network training methods. Moreover, data augmentation can lead to significant loss of original information. Wang et al. (2023) proposed a loss-based sample selection mechanism and an evolutionary auxiliary feature detection method, which improve the detection performance of low-proportion classes through the sample selection mechanism and enhance robustness to background clutter through evolutionary auxiliary feature detection. Additionally, enhancing small object detection using high-resolution images requires substantial computing resources. STNet (Fang et al. 2023) proposed a novel reinforcement learning framework that employs a Spatial Transformation Network to enhance state representation learning, avoiding unnecessary computation in background areas. However, this method cannot mitigate the resolution degradation of small targets caused by transformation. DSAA-YOLO (Hui et al. 2024) proposed a novel data augmentation strategy to address this problem, called Super-Resolution Data Augmentation. This strategy integrates the concept of image super-resolution into the dataset, introducing a super-resolution module based on dense residuals.

In contrast to general detection images, geospatial targets in aerial images are often densely arranged with significant scale variations, making manual annotation highly challenging. Training accurate object detectors demands a substantial amount of annotated data, which can be both costly and time-consuming. DUA (Yamani et al. 2024) introduced an active learning framework for single-stage object detectors targeting UAV images, designed to select images with a greater diversity of object categories and higher uncertainty. Additionally, manually labeled aerial images in model training often present a long-tail class distribution problem. MUSCDB (Liang et al. 2023) proposed a user-friendly class balancing criterion advantageous to minority objects. Considering the substantial limitations in label quantity and quality in manually labeled training samples, WDPL (Shen et al. 2023a) introduced a novel framework based on the teacher-student paradigm, which weights dense proposals using box confidence and ranks student proposals via consistency analysis to select discriminative and consistent boxes.

Most detection algorithms in aerial images employ fully supervised training methods, necessitating extensive, time-consuming, and labor-intensive instance-level annotations.

Weakly supervised object detection can significantly reduce annotation costs. Typically, weakly supervised object detection is defined as a multiple instance learning problem. However, it often only detects distinguishable parts of objects and might miss entire object instances. AEIS (Xie et al. 2023b) applies an attention erasure scheme to hide the most distinguishable regions and uses an IoU-balanced sampling component to discover more object instances. Additionally, weakly supervised object localization suffers from significant discriminative region issues. Bai et al. (2023a) utilizes GradCAM++ to precisely extract implicit spatial location information within classification models, directing the learning process. Although these optimization methods address some issues in weakly supervised methods, there may still be out-of-distribution samples mixed in the unlabeled datasets during training. Liu et al. (2024b) proposed open set semi-supervised object detection, which dynamically constructs a class feature library using labeled in-distribution data to capture class-specific features, by comparing the predicted object bounding box features with the corresponding entries in the feature library.

Aerial images also present the unique challenge of targets in arbitrary directions. Research on using weakly supervised detectors to learn rotated boxes from horizontal boxes is increasingly gaining attention. To mitigate the unreliability of pseudo-labels in localization, scale, and orientation, PST (Wu et al. 2024a) introduced Pseudo Siamese Teacher, a novel semi-supervised learning framework for object detection, employing scale-adaptive knowledge distillation to tackle this problem. However, this method still suffers from low annotation efficiency. To tackle this problem, Point2RBox (Yu et al. 2024a) introduced the first end-to-end point-supervised OOD solution, sampling around each labeled point in the image to extend object features into synthetic visual patterns with known boxes, offering knowledge for box regression. However, inconsistencies in the labeled points can introduce semantic variation. CPR++ (Yu et al. 2024b) introduced Coarse Point Refinement, marking the first attempt to alleviate semantic differences from an algorithmic standpoint. Since targets in aerial images are usually in arbitrary directions without axis alignment, mainstream weakly supervised object detection methods can only predict HBB from proposals generated by offline algorithms. WSODet (Tan et al. 2023) designed a method using layered relevance propagation and point set representation to predict OBB for aerial targets. To address the ambiguity in describing oriented targets with the aforementioned methods, RMRGM (Wu et al. 2024b) proposed a Rotation-Modulated Relation Graph Matching method to establish proposal matching relations centered on annotation points between teacher and student models.

Transfer learning has attained significant advanced achievements across numerous research fields. Presently, target detection datasets in the general domain have matured considerably, enabling the transfer of knowledge from large models to small models, thereby achieving low parameter counts and high precision. Mainstream approaches directly replicate the teacher model's features to boost the student model's performance, yet they neglect the process of using teacher features to guide the generation of high-level features in the student feature map. ACFG (Zhang et al. 2024c) utilizes the Adaptive Composite Feature Generation(ACFG) strategy to achieve end-to-end trainability for remote sensing image target detection. In this process, the source domain model generates pseudo-labels for target domain data, but variations in different remote sensing scenes are directly injected into the pseudo-labels, causing domain shifts. DualDANet (Zhu et al. 2023) introduced a dual-head correction domain adaptation network to alleviate the induced domain shifts. Furthermore,

aerial images, compared to ground images, exhibit issues of scale variation and viewpoint diversity. Ma et al. (2024a) conducts domain-adaptive object detection by allocating adaptive weights to domain classifiers at various levels to balance the transferability and discriminability of the adaptive detector.

4.1.1 Object detection with sample imbalance

Due to the influence of imaging angles in aerial images, the collected data often consists of wide-field high-resolution images, where valid targets are unevenly distributed and there is a large amount of background interference. The problems addressed by the models reviewed in this section and their advantages are listed in Table 2. This leads to an imbalance in the distribution of positive and negative samples, and small-scale targets are easily overwhelmed by background noise, resulting in low detection efficiency. To address this issue, Deng et al. (2021) proposed an end-to-end global–local adaptive network (GLSAN). An self-adaptive region selection algorithm (SARSA) was used for image data to perform adaptive dynamic cropping out of the target crowded regions. Subsequently, the cropped dense small-target area images are magnified, enhancing the network's ability to extract finer-scale features while suppressing the interference of background data in the images on the detection task, aiding the detector in distinguishing foreground from background more easily. Pareto refocus Detection (Leng et al. 2023) (PRDet) uses a reverse attention mechanism to distinguish areas with crowded targets from ordinary areas. Similarly, the Density-Map Network (Li et al. 2020a) (DMNet) addresses these challenges based on the strategy of image cropping, where DMNet determines the presence or absence of an object in a region by the change in pixel intensity, thus providing statistical guidance for cropping the image. However, these image cropping methods add other learnable components, making the network's training and inference more complex. Meethal et al. (2023) proposed an efficient cascaded amplification detector that repurposes the detector itself for density-guided training and inference. Meanwhile, large images are divided into small patches, all of which must be traversed, severely hindering inference speed. OAN (Xie et al. 2024a) is a lightweight fully convolutional network used to determine whether each patch contains an object, reducing network complexity and improving inference speed (Table 2).

Although cropping methods mitigate the issue of uneven target space distribution, they overlook the crucial problem of class imbalance among target sample classes.Yu proposed the Dual Sampling Head Network (Yu et al. 2021) (DSHNet) to handle the challenge of handling both tail and head classes in a dual-path manner, resulting in a substantial enhancement in the performance of tail classes across various detection scenarios. For the imbalance in the number of categories in the image data. Chen et al. (2023) proposed an Online Continuous Object Detector (OCOD), which leverages entropy to quantify the differences in the number of images of different categories in memory.

Due to variations in the altitude and angle of unmanned aerial vehicle (UAV) flights, occlusion phenomena in UAV images occur more frequently compared to those in natural scenes. In contrast to occlusions in natural scene images, UAV image occlusions exhibit issues of feature confusion and local clustering characteristics. The occlusion features in UAV images can be summarized as follows: (1)Local Clustering. Occlusions in UAV images tend to be clustered rather than sporadically distributed as in natural images (Kortylewski et al. 2019). The uneven distribution of objects and local crowding are unique challenges

Table 2 A briefly overview optimization algorithms for handling the problem of sample imbalance

Methods	Problem solved	Optimization strategies
LGNet (Liu et al. 2024a)	Data Augmentation	Pre-trained Multimodal
GP (Wang et al. 2023d)	Data Augmentation	Sample Selection Mechanism
STNet (Fang et al. 2023)	Data Augmentation	Spatial Transformation Network
DSAA-YOLO (Hui et al. 2024)	Data Augmentation	Super-Resolution Data Augmentation
DUA (Yamani et al. 2024)	Annotated data	Active Learning Framework
MUSCDB (Liang et al. 2023)	Annotated data	Class balancing criterion
WDPL (Shen et al. 2023a)	Annotated data	Teacher-student paradigm
AEIS (Xie et al. 2023b)	Weakly supervised	IoU-balanced sampling component
SDWSS (Bai et al. 2023a)	Weakly supervised	GradCAM++
OSSOD (Liu et al. 2024b)	Weakly supervised	Open set semi-supervised
PST (Wu et al. 2024a)	Arbitrary directions data	Pseudo Siamese Teacher
Point2RBox (Yu et al. 2024a)	Arbitrary directions data	End-to-end point-supervised OOD solution
CPR++ (Yu et al. 2024b)	Arbitrary directions data	Coarse Point Refinement
WSODet (Tan et al. 2023)	Arbitrary directions data	Layered relevance propagation
RMRGM (Wu et al. 2024b)	Arbitrary directions data	Rotation-Modulated Relation Graph Matching
ACFG (Zhang et al. 2024c)	Transfer learning	Adaptive Composite Feature Generation
DualDANet (Zhu et al. 2023)	Transfer learning	Dual-head correction domain adaptation
DAOD (Ma et al. 2024a)	Transfer learning	Domain classifiers
GLSAN (Deng et al. 2021)	Uneven spatial distribution	SARSA+LSRN
PRDet (Leng et al. 2023)	Uneven spatial distribution	Reverse attention mechanism
DMNet (Li et al. 2020a)	Uneven spatial distribution	Guidance for cropping the image
Meethal et al. (2023)	Uneven spatial distribution	Efficient cascaded amplification detector
Xie et al. (2024a)	Uneven spatial distribution	Determine whether each patch contains an object
DSHNet (Yu et al. 2021)	Imbalance between categories	Dual-path manner
OCOD (Chen et al. 2023d)	Imbalance between categories	Entropy Reservoir Sampling(ERS)
SCFNet (Yue et al. 2023)	Imbalance in samples	LCM+NLFM
DTSSNet (Chen et al. 2024a)	Imbalances in samples	Enhanced attention feature module
SARDNN (Liu et al. 2024c)	Imbalances in samples	Evidential learning
MgD (Ge et al. 2024a)	Imbalances in samples	Multiple evaluation functions
TOYOLOX (Chen et al. 2023c)	Imbalances in samples	Multi-point Single feature fusion structure
Dual neural network review (Tian et al. 2021)	Interference from background noise	Classifying secondary features
OGMN (Li et al. 2023a)	Interference from background noise	Occlusion-Guided Matching Network
MOL (Wang et al. 2023e)	Interference from background noise	Multi-view object mining strategy
HSOD-Net (Wu and Xu 2021)	Interference from background noise	Key point prediction
AFA-FPN (Wang et al. 2022a)	Interference from background noise	Polarized Hybrid Domain Attention(PHDA)
Focus-and-Detect (Koyun et al. 2022)	Less feature information	Gaussian mixture model+IBS

Table 2 (continued)

Methods	Problem solved	Optimization strategies
MRDet (Qin et al. 2022b)	Less feature information	Arbitrarily Oriented-RPN(AO-RPN)
DenseUGE (Yan et al. 2024)	Less feature information	Dense Universal Generalization Environment
SGG (Stateczny et al. 2022)	Less feature information	Improved utilization of feature selection
DCDet (Shen et al. 2024)	Less feature information	Dynamic Contextual Detection
CoF-Net (Zhang et al. 2023a)	Insufficient feature information	Enhanced feature representation
SME-Net (Ma et al. 2022)	Insufficient feature information	Feature splitting and merging(FSM)
FSANet (Wu et al. 2022b)	Insufficient feature information	Alignment mechanism and progressive optimization
FADA (Xu et al. 2022b)	Insufficient feature information	Enhancing cross-domain adaptation ability

in UAV imagery (Deng et al. 2020). (2) Feature Confusion. Conventional detectors typically aim to learn discriminative features (Zhou et al. 2019a); however, the discriminative features of occluded objects may be spatially obscured by other occluded objects. From the perspective of camera capture, natural scene images are mostly acquired from the side view of objects, whereas UAV images encompass both the side and top views of targets, presenting a rich and complex set of features. All these factors make it difficult for existing detectors to extract semantic feature information from these occluded regions of the objects. Similar to the issue of feature confusion, this problem is referred to as feature confusion of occluded objects (Li et al. 2019). Based on these occlusion characteristics, the Occlusion-Guided Multi-task Network (OGMN) (Li et al. 2023a) proposes a novel approach by introducing an occlusion localization task, which, together with the object detection task, forms the Occlusion-Guided Multi-task Network (OGMN).

Besides optimizing category imbalance, there are also imbalances in size distribution, orientation distribution, and background distribution in aerial images. DTSSNet (Chen et al. 2024a) proposed an enhanced attention feature module to enhance basic features by focusing on target-related channels and semantic information. The dynamic training sample selection module ensures a more balanced representation of positive and negative anchor boxes. However, these methods lack reliable descriptors during the training process. SARDNN (Liu et al. 2024c) uses evidential learning to obtain cognitive uncertainty as descriptors for sample bias learning. It further uses contrastive learning to utilize the uncertainty labels of samples to correct bias learning under class imbalance conditions. As for the inherent imbalance problem between positive and negative samples, Ge et al. (2024a) relies on multiple evaluation functions and two dynamic thresholds, applied strategically layer by layer, to provide detailed analysis for objects of different sizes. Although the above methods alleviate the distribution imbalance of image data features, they do not further optimize the imbalance between positive and negative samples within individual samples at the global and local levels. SCFNet (Yue et al. 2023) introduces the Semantic Correction and Focus Network (SCFNet), which employs a Local Correction Module (LCM) to correct local features. Additionally, it utilizes a Non-local Focus Module (NLFM) to enhance target feature recognition by leveraging non-local dependencies and corrected local features obtained from the LCM. TOYOLOX (Chen et al. 2023c) features a Multi-point Single feature fusion struc-

ture with spatial displacement architecture, which balances positive and negative samples, thereby improving the information interaction of local patches in remote sensing images.

The problem of missed detection is common for target regions containing small and dense targets in UAV images. Tian et al. (2021) proposes a dual neural network review method aimed at efficiently filtering out missed targets in single-stage detection. This approach involves classifying secondary features of suspected target regions to identify and capture targets that may have been initially overlooked. Moreover, occlusion in drone imagery is associated with feature confusion and local aggregation characteristics. The Occlusion-Guided Matching Network(OGMN) (Li et al. 2023a) incorporates an occlusion localization task in conjunction with the object detection task to address these issues. To tackle feature loss resulting from local occlusions, HSOD-Net (Wu and Xu 2021) is a point-to-region detection paradigm that uses key point prediction to obtain location hypotheses of targets, and then detects candidate regions near these hypothesized key points after super-resolution recovery of the image. This layered approach is more stable and also saves time. To address missed detections caused by background interference and densely arranged objects, the Multi-View Object Localization(MOL) (Wang et al. 2023e) approach introduces an innovative multi-view object mining strategy. This method progressively uncovers overlooked targets from multiple distinct perspectives, thereby alleviating the issue of missed detections. AFA-FPN (Wang et al. 2022a) introduces a novel approach that defines the correspondence between feature mappings, addresses the issue of feature mismatch between adjacent layers, and enhances the recognition of small targets. This is achieved through the alignment and fusion of shallow spatial features and deep semantic features, resulting in improved performance.

The Focus-and-Detect (Koyun et al. 2022) framework solves the small target detection problem through two phases. Additionally, to address the truncation effect of the region search method, the incomplete box suppression (IBS) method is utilized. On the other hand, two-stage network algorithms usually solve the problem by estimating additional orientation parameters and placing dense anchor points, which leads to higher model complexity and computational cost. MRDet (Qin et al. 2022b) uses an arbitrarily oriented region proposal network (AO-RPN) to generate orientation suggestions converted from horizontal anchor points. Compared with the original RPN, the AO-RPN is highly efficient with only a small number of additional parameters. However, these methods primarily focus on recognizing individual targets, which can result in a large number of algorithm parameters. The Dense Universal Generalization Environment(DenseUGE) (Yan et al. 2024) addresses this by clustering targets with similar distances and semantic similarities for detection, thereby enhancing efficiency.

Facing the problem of large background noise in image data and lack of feature representation capability in target regions. Stateczny et al. (2022) proposed spiral search grasshopper (SSG) optimization technique to improve the utilization of feature selection. However, due to significant variations in data distribution within aerial images, static models often yield unstable results. The Dynamic Contextual Detection(DCDet) (Shen et al. 2024) approach introduces a novel dynamic sensing and correlation loss detector. This method employs an object focusing module within dynamic sensing to consistently focus on small objects in each video frame. Zhang et al. (2023a) proposed a novel coarse-to-fine framework for detection in remote sensing images, called CoF-Net. The objective of CoF-Net is to progressively enhance feature representations and select stronger training samples. Nonetheless,

the extensive background areas can result in greater interference for the targets. The feature splitting and merging enhancement network (SME-Net) (Ma et al. 2022), which uses the feature splitting and merging module (FSM) to eliminate salient information of large objects and highlight features of small objects in shallow feature mapping. Although this method directs the networks focus more toward target features, issues with feature and spatial misalignment still persist during the feature extraction process. The Feature and Spatial Alignment Network (FSANet) (Wu et al. 2022b) is proposed to obtain more discriminatory features and accurate localization results using an alignment mechanism and an incremental optimization strategy. In the context of feature alignment methods, there can be conflicts in feature information between general object detectors and aerial image algorithms due to domain discrepancies. FADA (Xu et al. 2022b) is proposed as a method to enhance the domain adaptive capability between remote sensing images from different domains, thereby improving the robustness of detection in complex scenes.

4.1.2 Object detection with direction diversity

Compared to images from ground perspectives, the orientation of targets in aerial images is random, which may prevent detectors from effectively distinguishing between positive and negative samples. The problems addressed by the models reviewed in this section are listed in Table 3, along with their advantages Additionally, applying conventional detection annotation methods to aerial images unavoidably introduces background into the ground truth boxes, negatively impacting the extraction of effective target features (Table 3).

To address the boundary-arbitrary discontinuity problem of regression-based detectors, FCOSF (Rao et al. 2023) resolves the issue of discontinuous boundaries by associating regression outputs with regression target directions through a contour function. This function is designed to integrate direction information and produce corresponding distance predictions. Furthermore, difficulties are encountered due to discontinuities during the feature extraction process. A novel single-stage detector (Yuan et al. 2022) has been developed. It includes a feature alignment block (FAB), dual regression branches (DRBs), and a circular rotation box (CRB). This detector helps to extract feature information with strong recognition ability, which is beneficial for improving the positional and classification accuracy of the target. Similarly, DA-Net (Li et al. 2022) has the ability to adjust neuron receptive fields by using the Rotation Feature Selection (RFS) module. Furthermore, the Rotation Feature Alignment (RFA) module is used to achieve instance-level feature alignment by adaptively aligning features based on the size, shape, and direction of the corresponding anchor points. Additionally, the angle loss during the training process can be discontinuous. AF2Det (Yu et al. 2023c) addresses this by using bounding box projection information instead of angle information to represent and reconstruct the rotated bounding boxes of targets, thereby avoiding boundary discontinuities. However, these methods do not fully account for the different features required by classification and localization tasks. In TSH (Yu et al. 2023a), learnable points effectively adapt to targets of varying shapes and orientations, and the dynamic information aggregation module enhances relationships between scattered points, thereby improving localization accuracy. In solving rotation and aspect ratio issues, these methods often inadvertently introduce decoding discontinuities. COBB (Xiao et al. 2024) presents a new representation approach that ensures the continuity of bounding box regression.

Table 3 A briefly overview optimization algorithms for handling the problem of direction diversity

Methods	Problem solved	Optimization strategies
FCOSF (Rao et al. 2023)	Boundary-arbitrary discontinuity	Contour function
single-stage detector (Yuan et al. 2022)	Boundary-arbitrary discontinuity	FAB+DRBs+CRB
DA-Net (Li et al. 2022)	Boundary-arbitrary discontinuity	RFS+RFA
AF2Det (Yu et al. 2023a)	Boundary-arbitrary discontinuity	Using bounding box projection information
TSH (Yu et al. 2023b)	Boundary-arbitrary discontinuity	Learnable points effectively adapt
COBB (Xiao et al. 2024)	Boundary-arbitrary discontinuity	New representation approach
MGAR (Wang et al. 2022c)	Vague angle representation	CAC+FAR+IFL
OrtDet (Zhao et al. 2022)	Angular periodicity	Mean rotational accuracy (mRP)
AProNet (Zheng et al. 2021a)	Angular periodicity	Axis-based Angle Learning
FANet (Zhang et al. 2023b)	Angular periodicity	CSL
Arbitrary orientation regression (Dong et al. 2022d)	Arbitrary angle	Adaptive target orientation regression
RPG (Wang et al. 2024)	Arbitrary angle	Rotation-Equivariant CNN
DFDet (Xie et al. 2024b)	Angle sensitivity	Dual-focus detector
CPCCL (Mei et al. 2023)	Angle sensitivity	Polar coordinate
ARC (Pu et al. 2023)	Angle sensitivity	Adaptive Rotational Convolution
SFFD (Zheng et al. 2023)	Angle sensitivity	Layered Frequency Domain Analysis
CFC-Net (Ming et al. 2022)	Arbitrary angle	Rotation anchor refinement module
Angle encoding mechanism (Xiao et al. 2022a)	Arbitrary angle	Aspect ratio-based bidirectional coding label
RH-RCNN (Yang et al. 2022b)	Arbitrary angle	Distinguish tilted targets
AOPDet (Zhu et al. 2022)	Rotation object representation	Non-sequential angular representation
ACE (Dai et al. 2022a)	Rotation object representation	Directed quadrilateral box
Faster R-CNN-based (Ji et al. 2021)	Rotated Region Proposal	Majority voting strategy
RiDOP (Wei et al. 2023)	Rotated Region Proposal	Sliding only two vertices
R-RCNN (Fu et al. 2020)	Rotated Region Proposal	Directional RoI pooling operation
Point RCNN (Zhou and Yu 2022)	Rotated Region Proposal	PointRPN module generates RRoI
NGC (Wang et al. 2023a)	Rotated Region Proposal	Naive geometric calculations
New anchor-free detector (Liu et al. 2022a)	Arbitrary angle	Center Boundary Dual-Attention (CBDA)
AOPG (Cheng et al. 2022)	Arbitrary angle	Generates orientation boxes in an anchor-free manner
AOPG+FRIoU (Qian et al. 2023b)	Arbitrary angle	Focal Rotated Intersection over Union(FRIoU)
R2YOLOX (Liu et al. 2022b)	Arbitrary angle	Refined Rotation Module (RRM)
DARDet (Zhang et al. 2022a)	Arbitrary angle	ACM+PIoU
ADT-Det (Zheng et al. 2021b)	Inadequate expression of features	Feature Pyramid Transformer (FPT)
AFA-FPN (Wang et al. 2022a)	Inadequate expression of features	Employs RROI to rotate the horizontal frames
RINet (Feng et al. 2022)	Inadequate expression of features	Flexible multi-branch online detector improvement

Table 3 (continued)

Methods	Problem solved	Optimization strategies
FoRDet (Zhang et al. 2022b)	Inadequate expression of features	Foreground Relationship module (FRL)
FFN (Zhang et al. 2023c)	Inadequate expression of features	Fountain Feature Enhancement Module

In the process of directional prediction for targets, angular periodicity issues can arise. The network may exhibit instability when learning directional angles during training. AProNet (Zheng et al. 2021a) is a novel angle learning network that utilizes axis projection to obtain the orientation angle of an object without the issue of angular periodicity. However, this method can lead to abrupt changes in the boundaries of rotated frames. FANet (Zhang et al. 2023b) introduces an angle prediction branch and utilizes the Circular Smoothing Label (CSL) method to convert the angle regression problem into a classification problem. Ambiguity in angle prediction is avoided by using coarse-grained angle classification (CAC) (Wang et al. 2022c). Based on CAC, fine-grained angle regression (FAR) was developed to improve angle prediction with fine-grained angle regression at a lower cost.

Direction encoding in the aforementioned networks is frequently unstable and noisy. Wang et al. (2024) addresses this by proposing a Rotation-Equivariant CNN to extract rotation-equivariant features. Dong et al. (2022d) proposed a method based on adaptive target orientation regression with five coordinate parameters to obtain target regions in any direction. However, this method inadequately accounts for sensitivity to contextual priors and angle regression. DFDet (Xie et al. 2024b) introduces a dual-focus detector that addresses both the extraction of contextual knowledge and the reduction of angle sensitivity. Regarding the issue of rotation sensitivity in traditional CNNs, CPCCL (Mei et al. 2023) uses polar coordinate transformations to convert rotational changes into translational changes, making them easier for CNNs to handle. Additionally, the intrinsic nature of rotational information makes it challenging for standard backbone networks to capture high-quality features of objects in arbitrary orientations. ARC (Pu et al. 2023) introduces the Adaptive Rotational Convolution(ARC) module, which adaptively adjusts rotation to extract features of targets from different directions in diverse images. During the process of implicit modeling of directional features, CNNs struggle to perceive objects under different transformations. SFFD (Zheng et al. 2023) constructs a Layered Frequency Domain Analysis module to extract frequency features and compensate for the directional information overlooked by instance-level CNN features. From the perspective of anchor point optimization, CFC-Net (Ming et al. 2022) proposed a Key Feature Capture Network (CFC-Net). The network's Rotational Anchor Refinement Module (R-ARM) utilizes extracted discriminative regression features to refine the positioning of preset horizontal anchors, resulting in improved rotational anchors. Xiao et al. (2022a) designed an aspect ratio-based angle encoding mechanism to address the problem of inconsistent angles of square objects in current classification methods. This mechanism employs an aspect ratio-based bidirectional coding label to replace the circular smoothing label (CSL) for angle coding of square-like objects. By doing so, it enhances the detection accuracy specifically for square-like objects.

Existing rotational object detectors mostly suffer from the issue of supervision ambiguity due to inadequate rotational object representations. The Automatic Organization Point Detector (Zhu et al. 2022) (AOPDet) uses a novel rotating object representation called

non-sequential angular representation to obtain accurate localization results. This method addresses the problem of blurred supervision due to inappropriate rotating target representation. ACE (Dai et al. 2022a) detection method evolves the axial bounding box into a directed quadrilateral box with the assistance of dynamically collected contour information. Key points are set based on the contour points of the sampled axial bounding boxes. Finally, the network estimates the offsets between the key points of the axial bounding box and the corner points of the directed quadrilateral box.

Several approaches have been proposed to enhance the region proposal mechanism for detecting rotating objects. One such approach involves incorporating angle information of the object into the region proposal. For instance,a Faster R-CNN-based target detection model is proposed in (Ji et al. 2021) that combines multi-angle feature driving with a majority voting strategy.In another approach, a novel, lightweight rotated region proposal network is proposed in RiDOP (Wei et al. 2023) that generates arbitrarily-oriented recommendations by sliding only two vertices on adjacent edges and uses a simple and efficient representation to describe oriented objects. RH-RCNN (Yang et al. 2022b) is proposed for target detection in arbitrary directions. It distinguishes between near-horizontal and tilted targets by using RH-head networks and uses a rotational envelope only for locating objects that are significantly tilted. R-RCNN (Fu et al. 2020) is another approach that is built on orientation boxes to locate targets in remotely sensed images. In the initial stage of network training, the region proposal method provides proposal boxes with multiple default angles to cover the detection targets. It uses orientation proposal boxes to encapsulate the objects instead of roughly locating the horizontal proposals of the oriented objects.The region proposal mechanism based on rectangular boxes has certain limitations when it comes to regressing the orientation angle of the targets. However, by utilizing corner points for localization and regression of rotated objects, this issue can be effectively alleviated. Corresponding to this, Point RCNN (Zhou and Yu 2022) proposes to generate accurate rotating regions of interest (RRoI) through the PointRPN module. The PointReg module builds upon the learned Region of Interest (RoI) features obtained from PointRPN and focuses on regressing and refining the corner points of each Rotated RoI (RRoI), enabling more precise detection of rotating objects. These approaches learn to predict object orientation directly under individual supervision from one or several base truths. By adding extra constraints for joint supervision during training in the areas of proposals and rotational information regression, directional object detection can become more accurate and robust. NGC (Wang et al. 2023a) applies naive geometric calculations to consistently learn horizontal and directional proposals as well as object rotation angles simultaneously, offering an additional stabilizing constraint.

When dealing with rotation problems, the mechanism for generating anchor boxes cannot directly adapt to the varying angles of objects in an image, and using anchor-free can avoid this issue. The Center Boundary Dual-Attention Network (Liu et al. 2022a) (CBDA-net) is a new anchor-free detector that constructs a CBDA module to learn more essential features of rotated objects and reduce interference from complex backgrounds. An innovative anchor-free orientation proposal generator (Cheng et al. 2022) is proposed, which eliminates the need for horizontal frame-related operations in the network architecture. A coarse localization module (CLM) generates coarse orientation boxes in an anchor-free manner, which are then refined into high-quality directional proposals. Based on the AOPG architecture, Qian et al. (2023b) designs a boundary box regression loss called Rotated Intersection over Union

(RIoU) for oriented object detection. In addition, the Focal RIoU (FRIoU) loss is proposed based on the RIoU loss to give more weight to hard samples in boundary box regression. However, current anchor-free rotation object detection methods that achieve high performance are often accompanied by increased complexity, resulting in slower inference speeds. The anchor-free rotation detector R2YOLOX (Liu et al. 2022b) uses a refined rotation module (RRM) and a new allocation method, namely the Gaussian distribution sampling optimal transport allocation method. RRM aligns features and obtains more useful priors for the final detector head. In addition to optimizing the allocation method, optimizing the loss function is also helpful for regression tasks. The Dense Anchor-Free Rotation Object Detector directly predicts the five parameters of rotated boxes for every foreground pixel in the feature map. Additionally, the network incorporates a novel alignment convolution module (ACM) to extract aligned features and introduces the Pixel Intersection over Union (PIoU) loss to achieve precise and stable regression.These methods incorporate new loss distance metrics and efficient feature extraction modules, which not only enhance detection accuracy but also improve detection speed.

The effectiveness of detecting rotating targets is improved by using a multi-branch network structure, which allows for multi-level flexible perception of images. The Feature Pyramid Transformer (Zheng et al. 2021b) (FPT) mechanism enhances feature extraction in rotating target detection through feature interaction. The rotation branch in the Attention-based Feature Alignment FPN (Wang et al. 2022a) (AFA-FPN) method employs RROI to rotate the horizontal frames obtained by RPN, thereby preventing missed detections of small targets with dense distribution and arbitrary orientation. The Rotation-Invariant network (Feng et al. 2022) adopts a flexible multi-branch online detector improvement strategy that enhances rotation-awareness for oriented objects. However, foreground context characterization remains a challenge for multi-branch networks. FoRDet (Zhang et al. 2022b) overcomes this limitation by introducing a foreground relationship module(FRL) that captures the contextual information of foreground regions. The FRL module aggregates the foreground context representation in the coarse stage and enhances the differentiation of foreground regions on the feature map in the fine stage. These detectors employ excessively standardized feature extraction structures and lack the capacity to adaptively adjust the feature representations of detection units. This susceptibility to background information and distracting targets results in diminished feature expression abilities. FFN (Zhang et al. 2023c) introduces a novel feature enhancement module known as the Fountain Feature Enhancement Module.

4.2 Multiscale-aware methods

The scale-varying nature of aerial images presents challenges for target detection, and various methods have been proposed to address this issue. Table 4 lists the problems addressed by the models reviewed in this section and their advantages. One effective approach is the use of feature pyramid networks (Lin et al. 2017a) (FPNs), which have been shown to improve detection of multi-scale, multi-class targets by fusing adjacent low-level fine-grained features to generate high-quality feature representations at each scale.One recent advancement in this area is the end-to-end Feature Return Pyramid Network (Wang et al. 2022d), which uses a feature return pyramid structure to improve detection of multi-scale, multi-class targets. This approach fuses adjacent low-level fine-grained features to generate

Table 4 A briefly overview optimization algorithms for handling the problem of Multiscale-aware Methods

Methods	Problem solved	Optimization strategies
FRPNet (Wang et al. 2022d)	Scale-varying	Feature return pyramid structure
SBFPN (Yu et al. 2023d)	Scale-varying	Bidirectional Feature pyramid structure
BMFFN (Xiao et al. 2022b)	Scale-varying	Bidirectional multiscale feature fusion network
MFPNet (Yuan et al. 2021)	Scale-varying	Receiver field blocks(RFBs)
ARPN (Yu et al. 2023b)	Region proposal networks	Adaptive Region Proposal Network
SDPDet (Yin et al. 2024)	Region proposal networks	Scale-Divided Activation Pyramid
SSPNet (Hong et al. 2022)	Conflicting information	CAM+SSM
SMAG (Hu et al. 2022)	Conflicting information	Multi-scale supervision module
MSAN (Zhang et al. 2020a)	Scale-varying	Multi-scale activation feature fusion block (MAFB)
HA-MHGEN (Tian et al. 2022)	Scale-varying	Extract explicit and implicit relationships
RRNet (Cong et al. 2022)	Scale-varying	Parallel multi-scale attention (PMA)
A-MLFFMs (Dong et al. 2022a)	Scale-varying	Adaptively integrate the multi-level outputs
MNAN (Liu et al. 2023)	Scale-varying	Enhancing multi-scale targets
FE-CenterNet (Shi et al. 2022)	Scale-varying	FAS+AGS
CDD-Net (Wu et al. 2022a)	Scale-varying	LCFN+HAPN
DNTR (Liu et al. 2024d)	Scale-varying	Contrastive learning
AGMF-Net (Gao et al. 2023b)	Scale-varying	Multi-task enhancement structure
YOLC (Liu et al. 2024e)	Small object regions	Local scale module
AdaZoom (Xu et al. 2022a)	Small object regions	Variable magnification for adaptive multi-scale detection
ZoomInNet (Liu et al. 2021a)	Small object regions	Adaptive key distillation point(AKDP)
CDKD (Chen et al. 2023b)	Small object regions	A spatial and channel-oriented structural discriminative module
UFPMP-Det (Huang et al. 2022)	Small object regions	Unified foreground packing(UFP)
SRAF-Net (Sun et al. 2021)	Scale-varying	Context-based deformable (CBD)
GSDet (Li et al. 2021a)	Scale-varying	Converts GSD regression into a probabilistic estimation process
GFA-Net (Zhu et al. 2022)	Scale-varying	Graph Focusing Process (GFP)

high-quality feature representations at each scale.Another method, called the Simplified Bidirectional Feature Pyramid Network (Yu et al. 2023d) (SBFPN), is designed to fuse multi-scale features effectively. Jump connections are utilized in the middle layer of SBFPN to offset and reuse small-scale target information. The proposed target detection method (Xiao et al. 2022b) introduces a novel approach that enables the selection of multiple features across a multi-scale feature mapping. It utilizes a bidirectional multiscale feature fusion network to effectively combine semantic features and shallow features. This fusion process aims to enhance the detection performance of small targets in complex backgrounds. By integrating diverse features at different scales, the method improves the accuracy and robustness of target detection in challenging scenarios. Furthermore, the Multi-Feature Pyramid Network (Yuan et al. 2021) (MFPNet) has been proposed to construct local context information using receiver field blocks (RFBs), which makes the network more suitable for target detection in complex backgrounds. In order to enhance the feature characterization

capability and introduce nonlinear transformations, the proposed method incorporates an asymmetric convolution kernel within the RFB. This novel convolutional kernel modifies the traditional symmetric kernel by introducing an asymmetry that enables more expressive and flexible feature representations. These approaches demonstrate the potential of using FPNs to address scale variation in aerial target detection (Table 4).

Although most previous methods have concentrated on developing efficient feature fusion strategies within feature pyramid networks, there is limited research on improving the performance of region proposal networks. The quality of the proposal boxes generated by RPN heavily relies on the rich feature representations extracted from the FPN backbone network. Furthermore, the fixed number of generated proposal boxes restricts their adaptability to small human target distributions. ARPN (Yu et al. 2023b) proposes a novel Adaptive Region Proposal Network to enhance the quality of proposal boxes and generate particularly compact and accurate proposals. To address the issues of high model complexity and low inference efficiency caused by high input resolution, SDPDet (Yin et al. 2024) introduces a Scale-Divided Activation Pyramid that focuses on object-clustering regions at each scale, while a scale-separated learnable proposal mechanism learns proposal boxes and corresponding features for these regions.

4.2.1 Attention-based improvements

The attention-based multi-level feature fusion modules provide a flexible and adaptive framework for effectively exploiting the hierarchical representations of FPN, enabling more accurate and robust target detection across different scales. A novel detector introduces a series of attention-based multi-level feature fusion modules (Dong et al. 2022a), which serve to adaptively integrate the multi-level outputs of FPN. This integration process enhances the detector's ability to detect targets at various scales. A multi-scale non-local attention-based network (Liu et al. 2023), which designed to effectively capture and fuse information at multiple scales, allowing for comprehensive analysis of the scene and enhancing the detection performance for objects of various sizes. Shi et al. (2022) proposes an anchor-free detector that mines multi-scale contextual information using a feature enhancement module consisting of a feature aggregation structure and an attention generation structure, combined with a coordinate attention mechanism to suppress the interference of false alarms in the scene, thus improving the perception of small objects. A context-driven detection network (Wu et al. 2022a) is introduced, which employs a local contextual feature network to capture local neighboring objects and features in the region of interest.

Attention mechanisms can help address the challenge of conflicting information exchange at different levels in multi-scale perception. The Scale Selection Pyramid Network (Hong et al. 2022) (SSPNet) for minutiae detection leverages the Contextual Attention Module (CAM) to create a hierarchical attention map that considers contextual information. The Scale Selection Module (SSM) is then employed to utilize the relationship between neighboring layers, enabling effective sharing of features between shallow and deep layers while avoiding inconsistencies in gradient computation across different layers. Alongside conflicts in gradient calculation between layers, the interaction of feature information between spatial and channel dimensions must also be taken into account. A novel hyper-visual multi-scale attention-guided detection framework (Hu et al. 2022) is presented, which utilizes a attention-guided module in this approach enhances the representation of features by computing

map correlations, which allows for the aggregation of contextual information from both spatial and channel dimensions. In addition, a new hybrid attention-driven multi-stream hierarchical graph embedding network (Tian et al. 2022)(HA-MHGEN) is a novel approach that aims to enhance the performance of multi-class target detection. On the other side, starting with the image data itself, by effectively capturing and integrating relevant information from different parts of the feature maps, the attention-guided module enriches the feature representation and enables more comprehensive and discriminative feature learning. The multi-scale attention network (Zhang et al. 2020a) is proposed, a scene-adaptive remote sensing image super-resolution strategy is utilized to accurately capture the structural features of various scenes. However, noisy features may arise during the fusion of features at different scales. DNTR (Liu et al. 2024d) utilizes contrastive learning to reduce noise in features at each level of the top path of the FPN and employs self-attention mechanisms to concentrate on the representation of small objects. Due to significant variations in target scales in aerial images, a relational reasoning network (Cong et al. 2022)(RRNet) based on parallel multi-scale attention (PMA) is introduced, in which the PMA module effectively recovers detailed information using underlying features refined by multi-scale attention to solve the scale-varying problem of salient targets. In the process of multi-scale information interaction, background interference must also be considered. AGMF-Net (Gao et al. 2023b) proposes a multi-task enhancement structure and multi-task feature preprocessing to enhance the feature representation of multi-scale targets while eliminating interference from complex backgrounds.

4.2.2 Image perception enhancement

Focusing on small object regions is crucial for accurate target detection. To address the challenges posed by large-scale images and non-uniform target distributions, YOLC (Liu et al. 2024e) introduces a local scale module that adaptively searches cluster regions for amplification to enable accurate detection. Similarly, One promising approach is the development of an adaptive zoom network, known as AdaZoom (Xu et al. 2022a), which acts as a magnifier with a flexible shape and focal length. It scales the focal region and uses variable magnification for adaptive multiscale detection, depending on the region's scale. However, due to the uneven sample distribution in aerial image datasets, with some targets being sparse, using the same methods may lead to a waste of computational resources. Another innovation is the cross-scale knowledge distillation method, also known as ZoomInNet (Liu et al. 2021a), which enhances the features of small targets in a similar way to image magnification. The teacher-student network is trained using images of various scales, employing an efficient FPN structure. By incorporating images of different scales during training, the teacher-student network can effectively learn and generalize across a wide range of input scales. Subsequently, the Adaptive Key Distillation Point algorithm is used to identify key locations for knowledge distillation. Finally, cross-scale information compression is achieved using a location-aware L2 loss measure, which compares the differences between cross-scale model feature mappings. The issues of detection inference time and model size, referred to as consistency and dependency-guided knowledge distillation problems, are tackled by (Chen et al. 2023b), which introduces a spatial and channel-oriented structural discriminative module to extract the discriminative spatial locations and channels focused on by the teacher model, improving model efficiency. While knowledge distillation methods

are effective in enhancing model efficiency, this post-processing approach may potentially degrade model performance. To handle numerous instances at very small scales, a Unified Prospect Packing Multi-Agent Detection Network (UFPMP-Det) (Huang et al. 2022) has been developed. It includes a unified foreground packing (UFP) module, which initially merges subregions given by coarse detectors by clustering to suppress the background. The results are then packed into a mosaic for individual inference, significantly reducing the total time cost.

Two different approaches for object detection are anchor-based and anchor-free methods. The former requires the definition of a number of anchors with fixed shapes, whereas the latter can learn the shape of an object from the training data. Shape-Robust Anchor-Free Network (Sun et al. 2021), proposed by Sun Xian, is an anchor-free method that uses contextual information obtained from the context-based deformable module. This allows the network to prioritize objects with ambiguous appearances and efficiently extract target features using deformable convolutions. Existing algorithms tend to overlook the clues provided by ground sample distances to solve problems such as extremely variable object scales. A Ground Sample Distance detector (Li et al. 2021a) addresses this issue by using a GSD identification subnet based on a two-stage detection framework that converts GSD regression into a probabilistic estimation process. GSD information is then combined with the size of the Region of Interest (RoI) to determine the physical size of the target.Zhu proposes a Graph-Focused Aggregation Network (Zhu et al. 2022), which represents the structural features of remotely sensed ground objects by studying the invariant structural features of ground objects in remotely sensed aerial images. The Graph Focusing Process is introduced based on the idea of graph convolution, aiming to address the challenge of extracting structural features from objects of the same category that exhibit arbitrary orientations and multi-scale variations, which is difficult for traditional CNNs.

4.2.3 Combining contextual information

Although traditional multiscale detection networks have made significant strides in addressing the problem of large variations, there are still limitations to their effectiveness. Table 5 lists the problems addressed by the models reviewed in this section and the advantages of the modelsThese methods typically focus on the scale of features, while neglecting the correlation between feature levels. Additionally, each feature mapping is represented by a single backbone network layer, leading to insufficiently comprehensive feature extraction. To address these limitations, various novel techniques have been proposed. For instance, the cross-scale feature fusion pyramid network (Huang et al. 2021) incorporates a cross-scale fusion module to effectively extract comprehensive semantic information from features, enabling multi-scale fusion. This module enables the network to capture and combine information from different scales, facilitating a more holistic understanding of the scene and improving the detection performance across various object sizes. To fully leverage the multiscale complementary information in the upper scale feature aggregation framework, the end-to-end single-stage target detector HawkNet (Lin et al. 2020) is introduced. Furthermore, a gated context-aware module (Dong et al. 2022c) (G-CAM) is proposed to address the lack of critical contextual information for FPN to accurately classify and localize objects. Similarly, Dong et al. (2022b) adds a receiver field expansion block at the top of the backbone network to adaptively expand the receiver field of the FPN. An adaptive

Table 5 A briefly overview optimization algorithms for handling the problem of Contextual Information

Methods	Problem solved	Optimization strategies
CF2PN (Huang et al. 2021)	Cross-level information exchange	Cross-scale fusion module (CSFM)
HawkNet (Lin et al. 2020)	Cross-level information exchange	Gated context-aware module (G-CAM)
FPN-based (Dong et al. 2022b)	Cross-level information exchange	Receiver field expansion block
ABNet (Liu et al. 2022c)	Cross-level information exchange	Adaptively combining multiscale features
MSGN (Zhu et al. 2022)	Cross-level information exchange	Backward semantic guided filtering (BSGF)
CEASC (Du et al. 2023)	Cross-level information exchange	Adaptive multi-layer masking strategy
MFEPN (Zhang and Shen 2022)	Cross-level information exchange	CAFUS+FEM
FFAGRNet (Zhang et al. 2024f)	Inadequate expression	FFA
SCANet (Zhang et al. 2022c)	Inadequate expression	RFEM+SCFM
MVNet (Han et al. 2022)	Inadequate expression	MRBs+MRFEM
MSFC-Net (Zhang et al. 2022d)	Inadequate expression	Composite Semantic Feature Fusion (CSFF)
CFANet (Zhang et al. 2023d)	Inadequate expression	Cross-Layer Feature Aggregation
SRAF-Net (Liu et al. 2022d)	Inadequate expression	SE-FPN+SADH
VSRNet (Ge et al. 2022)	Inadequate expression	VRG+SRG
ADCG (Gao et al. 2023a)	Inadequate expression	Constructing dense connection
mSODANet (Chalavadi et al. 2022)	Inadequate expression	Dilation Convolution for multiple scales
YoloOW (Xu et al. 2024)	Inadequate expression	OaohRep
PKINet (Cai et al. 2024)	Inadequate expression	Multi-scale convolutional kernels
SNLA (Ma et al. 2023)	Scale confusion problem	Scale-decoupling module
SGFTHR (Li et al. 2022a)	Channel information loss	Structure Guided Feature Transformation (SGFT)
MFAF (Lv et al. 2022)	Channel information loss	FI+SAW+CSP
M2S (Guo 2023)	Channel information loss	Improving feature extraction and feature refinement
Info-FPN (Chen et al. 2023a)	Feature misalignment	Feature alignment module (FAM)
FDLR-Net (Xiao et al. 2023)	Feature misalignment	Localization refinement module (LRM)
SME-Net (Ma et al. 2022)	Feature misalignment	Feature splitting and merging module (FSM)
GLNet (Teng et al. 2022)	Feature misalignment	MSP+SA
HyNet (Zheng et al. 2020a)	Feature misalignment	Learning hyperscale feature representations
ASSD (Xu et al. 2022b)	Feature misalignment	Pseudo-Anchor Proposal Module (PAPM)
VistrongerDet (Wan et al. 2021)	Feature misalignment	FPN-level, ROI-level, and head-level enhancements
DIMA (Cheng et al. 2024)	Inter-class similarity	Frequency-Aware Representation Supplement mechanism
PCLDet (Ouyang et al. 2023)	Inter-class similarity	Prototype learning
HRDNet (Liu et al. 2021b)	Feature misalignment	MD-IPN+MS-FPN
GLSANet (Gao et al. 2023c)	Restricted local information	Global semantic information interaction module
AGMFNet (Gao et al. 2023b)	Restricted local information	Spatial bias module

Table 5 (continued)

Methods	Problem solved	Optimization strategies
GLGCNet (Bai et al. 2023b)	Restricted local information	Saliency enhancement modules
Hyneter (Chen et al. 2024b)	Restricted local information	Hybrid network backbone and dual-switching modules
DCL-Net (Liu et al. 2021c)	Restricted local information	RFAM+PAM
SMSR (Dong et al. 2021)	Restricted local information	Aggregating features learned at different depths
GLSAN (Deng et al. 2021)	Crowded targets	GLDN+SARSA+LSRN
GDF-Net (Zhang et al. 2020b)	Crowded targets	Global density model (GDM)
GSNet (Shen et al. 2023b)	Crowded targets	Feature fusion refinement module (FRM)

balanced network (Liu et al. 2022c) (ABNet) is designed to capture more discriminative features by adaptively combining multiscale features at different channel and spatial locations, and a context enhancement module is introduced to exploit rich semantic information for multiscale target detection (Table 5).

Moreover, a multiscale semantic guidance network (Zhu et al. 2022)(MSGN) is proposed to utilize the features of different layers for detecting multiscale targets, and a multi-level semantic guided filtering sub-network is developed to suppress complex backgrounds. To tackle the problem of awkward control over mask ratios with foregrounds at different scales, CEASC (Du et al. 2023) introduces an adaptive multi-layer masking strategy that generates optimal mask ratios across different scales. Finally, the multi-stage feature Enhancement pyramid network (Zhang and Shen 2022)(MFEPN) is able to solve the problems of small-scale target blurring and large-scale target variation detected in optical remote sensing images. This is achieved through the use of content-aware feature upsampling and feature enhancement module that address the fusion problem of feature mapping in the neighboring stage.

Current feature fusion methods are restricted by the layer-by-layer propagation structure, which limits the effective exchange of information between feature maps at different scales. Zhang et al. (2024f) addresses this by using the FFA module to adapt scales and aggregate information across multiple feature map sets, resulting in high-quality aggregated feature maps. To fully utilize semantic context information and extract multi-scale feature representations, several models have been proposed. SCANet (Zhang et al. 2022c) is a semantic context-aware network. It uses the Received Field Enhancement Module (RFEM) to convolve multi-branch structures with different perceptual fields. Another model, MVNet (Han et al. 2022), proposes a multi-visual small target detector using Multi-Scale Residual Blocks (MRBs) with extended convolution in cascaded residual blocks. While these methods effectively leverage contextual semantic information, the multi-scale nature of the features requires careful consideration of the fusion process. MSFC-Net (Zhang et al. 2022d) is a single-level detection method using the Composite Semantic Feature Fusion (CSFF) method for generating valid semantic descriptions. For the fusion process, CFANet (Zhang et al. 2023d) designs a novel Cross-Layer Feature Aggregation module to aggregate features at different scales while avoiding semantic gaps. This module addresses the limitation of layer-by-layer feature propagation methods, which focus only on the previous layer's features and fail to fully integrate spatial and semantic information. SRAF-Net (Liu et al. 2022d) captures scene contextual features of the target with a Scene Enhanced Feature Pyramid Network, followed by Scene Assisted Detection Head (SADH) for more accurate

detection. For the optimization of the fine-grained inference detection phase, VSRNet (Ge et al. 2022) is proposed, which consists of a visual inference graph (VRG) and a semantic inference graph (SRG) to refine the feature representation of each instance.

Convolutional blocks are fundamental components of networks and directly impact the network's ability to extract features and transmit effective information. Standard convolution struggles with sparse labeled samples and imbalanced categories. The key improvement of ADCG (Gao et al. 2023a) lies in constructing dense connection modules and lightweight convolutional block attention modules. mSODANet (Chalavadi et al. 2022) enhances convolutional methods by using hierarchical dilated convolutions for multi-scale learning, enabling it to capture contextual information of various types of objects across multiple scales and fields of view. Fixed-kernel-size convolutional blocks can lead to missed and false detections due to significant differences in object appearance and size. YoloOW (Xu et al. 2024) uses the OaohRep convolutional block to extract image features at a broader spatial scale. However, in aerial image data, where there are large-scale variations and diverse backgrounds, large-kernel convolutions or dilated convolutions tend to introduce considerable background noise, while dilated convolutions may produce overly sparse feature representations. PKINet (Cai et al. 2024) addresses this by using multi-scale convolutional kernels without employing dilated convolutions to extract features of objects at different scales and capture local background information.

4.2.4 Feature alignment optimization

Existing methods utilize multi-level features to address scale variation issues, but they overlook the scale confusion problem in shallow features. SNLA (Ma et al. 2023) designs a scale-decoupling module to emphasize small object features by eliminating large object features in shallow layers. The remote sensing image target detector based on Feature Pyramid Network (FPN) encounters several challenges that hinder its performance. These challenges include the loss of channel information, feature misalignment, and the need for additional computational resources to mitigate the blending effect. These factors collectively lead to insufficient feature extraction for detecting multi-scale targets in remote sensing images. To address these issues, advanced methodologies are required to preserve channel information, align features accurately, and minimize the computational overhead associated with eliminating the blending effect. In response, Chen et al. (2023a) introduced the information feature pyramid network (Info-FPN), which includes a feature alignment module (FAM) to mitigate the confusion caused by feature misalignment. Moreover, to eliminate the confounding effect, the model features a semantic encoder module that reduces the parameters and computation required, while still achieving the desired detection accuracy. Additionally, the Structure Guided Feature Transformation Hybrid Residual (Li et al. 2022a) (SGFTHR) network was proposed, incorporating the Structure Guided Feature Transformation (SGFT) module, which extracts differentiated structural information and directs it to the higher-level contextual feature maps, thereby avoiding the loss of important lower-level spatial and structural information as the network goes deeper. The SGFTHR network presents a novel approach to address the challenge of achieving accurate detection performance across various scales, especially for small and densely packed objects, without relying on anchor-based mechanisms. Finally, the feature decoupling and localization refinement network (Xiao et al. 2023) (FDLR-Net) was designed with a localization refinement module aimed

at automatically optimizing the anchor box parameters. This module facilitates the spatial alignment between the anchor box and the target regression features, thereby enhancing the accuracy of object localization.

In aerial images captured from a large field of view, it is often necessary to have feature maps with a larger receptive field. As a result, network architectures typically incorporate deeper layers for feature extraction. However, this approach can pose a challenge as downsampling during the network's processing can lead to the loss of information related to small objects. To address this challenge, the multi-scale feature adaptive fusion (Lv et al. 2022) method is proposed, which incorporates a feature fusion module using a Feature Integration module and a spatial attention rights module to enable adaptive fusion of multi-scale features. In addition, the cross-stage partial block is used to reduce parameter numbers and feature loss. To account for feature scale variation and information loss, the Many-to-Single (Guo 2023) (M2S) module is introduced to enhance specific layers and improve feature extraction and refinement. The SME-Net (Ma et al. 2022) utilizes a Feature Splitting and Merging Enhancement Network (FSM) to strengthen features of small objects and transfer effective detailed features of large scale objects to depth feature maps, thereby mitigating feature confusion among multi-scale objects. A novel remote sensing image target detection network, GLNet (Teng et al. 2022), is proposed, which incorporates a Multi-Scale Perception (MSP) module to extract rich semantic features, and an adaptive anchor module to address the scale differences in semantics during sampling. A Hyperscale Target Detection Framework (Zheng et al. 2020a) is proposed to address the extreme scale variation problem by learning hyperscale feature representations. To address the feature misalignment problem, an Feature-Aligned Single-Shot Detector (Xu et al. 2022b) is proposed, which includes a novel pseudo-anchor proposal module to solve the spatial misalignment problem. In order to address the detrimental effects caused by large scale span in object detection, the VistrongerDet (Wan et al. 2021) method has been developed. VistrongerDet is specifically designed to enhance the performance of FPN-based two-level detectors by integrating new components at different levels: FPN-level, ROI-level, and head-level. Finally, the High Resolution Detection Network (Liu et al. 2021b) is proposed, which uses a multi-depth backbone network, a multi-depth image pyramid network, and a multi-scale feature pyramid network to take full advantage of multiple features and maintain high resolution images without introducing new problems. The aforementioned methods have not yet considered the high inter-class similarity in samples and the intrinsic relationships between comprehensive visual patterns of objects and multi-level features. DIMA (Cheng et al. 2024) designs a simple yet effective Frequency-Aware Representation Supplement mechanism, constructing a hierarchical relationship between coarse-grained and fine-grained representations. Additionally, PCLDet (Ouyang et al. 2023) introduces prototype learning to capture features of fine-grained objects and then uses contrastive learning to compare targets with the learned features, thereby improving the differentiability of fine-grained objects.

4.2.5 Global feature awareness

To tackle the challenges posed by complex backgrounds and small targets, GLSANet (Gao et al. 2023c) introduces a global semantic information interaction module to explore and strengthen high-level semantic information in deep feature maps, aiding in mitigating the obstruction of foreground targets by complex backgrounds. AGMFNet (Gao et al. 2023)

proposes a spatial bias module, which is part of our global information extraction module, designed to capture long-range dependencies and effectively gather global information. Various methods have been proposed to optimize small target detection in aerial imagery, but many fall into the local–global model. To solve this problem, GLGCNet (Bai et al. 2023b) investigates the collaborative effects of global context awareness and local context awareness modeling, creating an enhanced Global–Local-Global Context-Aware Network (GLGCNet). Moreover, there is a gap between local information and global dependencies in feature extraction and propagation, causing CNN and transformer methods to perform unevenly when dealing with objects of varying sizes. In contrast to the divide-and-conquer approach in previous methods, Hyneter (Chen et al. 2024b) consists of a hybrid network backbone and dual-switching modules that integrate local information and global dependencies while transmitting both simultaneously. DCL-Net (Liu et al. 2021c) introduces a Decoupled Classification and Localization Network (DCL-Net) that takes into account different features between the two branches. This network can greatly enhance the independence of the classification and localization branches, which helps improve the target detection performance in aerial remote sensing images. Because the detection results from the global branch can impede the enhancement of detail performance when merging detection results from branches of different resolutions, CGL (Chen et al. 2023e) proposes a Coupled Global–Local (CGL) network that uses a multi-scale feature fusion module to facilitate information sharing between global and local branches. SMSR (Dong et al. 2021) develops a Second-Order Multi-Scale Super-Resolution Network (SMSR) that skillfully captures multi-scale feature information by aggregating features learned at different depths in a single path.

To solve the issue of detection difficulties in areas occupied by crowded targets, Deng proposes an end-to-end Global–Local Self-Adaptive Network (Deng et al. 2021)(GLSAN). This network incorporates a global–local fusion strategy into a progressively scaled network for more accurate detection. To optimize the fusion algorithm for local and global features, Zhang introduces a new Global Density Fusion Convolutional Network (Zhang et al. 2020b)(GDF-Net). GDF-Net refines density features by employing an expanded convolutional network, aiming to provide a larger receptive field and generate global density fusion features. Besides optimizing the convolutional modules, there is also an issue of insufficient information in the neck network. Shen proposes a neck network consisting of a Global Semantic Network (Shen et al. 2023b)(GSNet) and a Feature Refinement Module (FRM). As a bridge between the backbone and head networks, GSNet is designed to perceive the context environment and propagate discriminative knowledge through bidirectional global patterns. On the other hand, FRM is developed to utilize features from different levels to obtain comprehensive positional information.

4.3 Optimized regression methods

In general object detectors, the deviation between IoU and ground truth is commonly used as the input for the loss function, which is then used to update the weight parameters through backpropagation. However, the choice of distance metric and label matching strategy based on this design is not suitable for datasets that predominantly contain small objects.Table 6 lists the problems addressed by the models reviewed in this section and the advantages of the models. This is because small objects occupy fewer pixels, resulting in smaller absolute

Table 6 A briefly overview optimization algorithms for handling the problem of Optimized regression

Methods	Problem solved	Optimization Strategies
KLDet (Zhou and Zhu 2024)	Label assignment	Localization metric
MENet (Zhang et al. 2024e)	Label assignment	Center region
RSADet (Yu and Ji 2022)	Label assignment	New bounding box confidence prediction
TCD (Zhang et al. 2023e)	Label assignment	Learning joint features
TSConv (Huang et al. 2024)	Label assignment	Adaptively samples task-specific features
ARG (Zhang et al. 2022e)	Label assignment	Dynamically change the weights of angle regression
RD-Net (Zhang et al. 2021b)	Label assignment	Adjustable sample selection module
ARSDETR (Zeng et al. 2024)	Label assignment	Aspect Ratio-Aware Circular Smooth Labels
PETDet (Li et al. 2023d)	Label assignment	Quality-Oriented Proposal Network
HAANet (Deng et al. 2023)	Label assignment	Latent label assignment modules
ARUBA (Sairam et al. 2023)	Label assignment	Neighborhood-driven approach
GOBB (Li et al. 2023b)	Label assignment	Gaussian OBB
DILA (Chen et al. 2024c)	Label assignment	Dynamic Gaussian distribution fitting
EMO2DETR (Hu et al. 2023)	Label assignment	Reallocating bipartite graph matching
FSME-Net (Ma et al. 2022)	Label assignment	Offset error correction(OER)
PointOBB (Luo et al. 2024)	Unreliable bounding boxes	Self-supervised learning
New BBR loss	Unreliable bounding boxes	Generalized intersection over union(GIoU)
UTS (Qian et al. 2023b)	Unreliable bounding boxes	Facilitate the transfer of BBR loss from HOD to OOD
LSKNet (Li et al. 2023c)	Unreliable bounding boxes	Dynamically adjusts
AOPG (Cheng et al. 2022)	Unreliable bounding boxes	Coarse orientation module (CLM)
B2V (Wang et al. 2023b)	Unreliable bounding boxes	Tanimoto coefficient
TIOEDet (Ming et al. 2023)	Unreliable bounding boxes	Posteriori hierarchical alignment(PHA)
CAF2ENet (Ge et al. 2024b)	Unreliable bounding boxes	Adjusts sample weights
New BBR (Yao et al. 2023)	Unreliable bounding boxes	Refine the final localization results
SESANet (Ma et al. 2024b)	Mismatch of evaluation indicators	generating samples using ADM
MIOU (Shen et al. 2022)	Mismatch of evaluation indicators	Manhattan distance
New evaluation metric (Xu et al. 2022b)	Mismatch of evaluation indicators	NWD+RKA
HIOU (Wang et al. 2022e)	Mismatch of evaluation indicators	Hausdorff distance
SOOD (Hua et al. 2023)	Mismatch of loss function	Designs two loss functions
GALoss Doloriel and Cajote 2023)	Mismatch of loss function	Designs an attention network composed of two types of losses
GFL (Wang et al. 2022f)	Mismatch of loss function	Adaptive Gaussian decay at negative positions
GCL (Ming et al. 2024)	Mismatch of loss function	Gradient correction loss

differences between generated anchors and ground truth. Moreover, inappropriate distance metrics during anchor regression can lead to unstable modifications of the deviations within a small range. Additionally, the dense distribution of targets requires a more reasonable label matching strategy to achieve a higher recall rate (Table 6).

4.3.1 More reasonable label assignment methods

Current solutions mainly focus on the aggregation of contextual information at different levels, rarely addressing label assignment. The mainstream IoU-based label assignment strategy fails to accurately measure the localization of tiny bounding boxes. In contrast, Zhou and Zhu (2024) localization metric precisely reflects slight displacements in small bounding boxes. Furthermore, IoU-based label assignment significantly worsens positive samples for small targets. MENet (Zhang et al. 2024e) proposes a center region based label assignment, treating anchors that fall into the center region of the ground truth box as positive samples, thereby providing more positive samples for small targets. The methods based on this idea aim to alleviate the reliance on overlap-based matching strategies and design more reasonable sampling methods based on the characteristics of objects in aerial imagery. One such approach is the Remote Sensing Spatially Adapted Detector (RSADet) (Yu and Ji 2022), which introduces a new bounding box confidence prediction method using the IoU score as a constraint to eliminate unreliable boxes and improve performance.

These fixed strategies undermine the consistency between classification and localization predictions. TCD (Zhang et al. 2023e) can flexibly adjust the spatial feature distribution of classification and localization tasks by learning joint features of the aggregation layer. This method ignores the inconsistency between localization and classification task features. TSConv (Huang et al. 2024) adaptively samples task-specific features from their respective sensitive regions and maps these features together to guide dynamic label assignment for better predictions. To tackle the misalignment issue between tasks, TIOEDet (Ming et al. 2023) proposes a posterior hierarchical alignment label to optimize the detection process. Another approach is the Visual Image Guidance (ARG) method proposed by (Zhang et al. 2022e), which considers the a priori information of object aspect ratio to adjust the label assignment criteria and dynamically change the weights of angle regression using the ARG label assignment and IoU loss. RD-Net (Zhang et al. 2021b) by Huijie Zhang employs an adaptable sample selection module to alleviate the dependence on IoU threshold hyperparameters and determine positive and negative training samples by considering the statistical features between anchor points and bounding boxes. However, classification-based methods for angle prediction overlook the sensitivity of targets with different aspect ratios to angles. ARSDETR (Zeng et al. 2024) proposes a new angle classification method called Aspect Ratio-Aware Circular Smooth Labels, which more reasonably smooths angle labels and removes the hyperparameters introduced in previous work.

Region proposal methods commonly used in object detection networks ignore some proposal-related processes inherited from general detection that are not as applicable to aerial image detection tasks. PETDet (Li et al. 2023d) introduces a Quality-Oriented Proposal Network (QOPN) based on dynamic label assignment and attention decomposition to generate high-quality proposals. To address inconsistencies in candidate region selection, target feature extraction, and preset box label assignment, HAANet (Deng et al. 2023) designs region refinement, feature alignment, and latent label assignment modules to alleviate misalignment at the region, feature, and label levels, respectively. Regarding the alignment of features and labels, ARUBA (Sairam et al. 2023) introduces a new balanced loss that follows a neighborhood-driven method inspired by the commonality of object sizes. Lastly, the Feature Splitting and Merging Enhancement Network (FSME-Net) (Ma et al. 2022) with

Offset Error Correction (OER) corrects inconsistencies in multi-layer feature mappings by using the proposed offset loss, enabling supervised elimination and transmission in FSM.

Most OBB detectors used in aerial images adopt a one-to-many label assignment strategy, where the angle discontinuity problem leads to boundary shifts. GOBB (Li et al. 2023b) uses Gaussian OBB to handle angle discontinuity, thus eliminating the shifts caused by direct synthesis. Additionally, DILA (Chen et al. 2024c) proposes a label assignment strategy based on dynamic Gaussian distribution fitting and simulated learning for small object detection, addressing the positional bias of effective receptive fields in different network layers during Gaussian modeling. When setting a fixed number of object queries and using bipartite graph matching for one-to-one label assignment, the matching can lead to relative redundancy in object queries. EMO2DETR (Hu et al. 2023) proposes reallocating bipartite graph matching to extract high-quality negative samples from negative samples.

4.3.2 More appropriate bounding boxes

Existing methods primarily focus on generating HBB, neglecting OBB commonly used in aerial images. PointOBB (Luo et al. 2024) achieves accurate object angle prediction by combining self-supervised learning to predict angles and employing a scale-guided dense-to-sparse matching strategy to aggregate dense angles corresponding to sparse objects. Compared to horizontal target detection, directional target detection can more accurately locate targets in any direction in remote sensing images. The most commonly used bounding box regression(BBR) loss in out-of-domain (OOD) methods is the smoothed L1 loss. Qian et al. (2023a) proposed a new BBR loss, the smooth Generalized Intersection over Union (GIoU) loss, which can solve the problem of the gradient value not being dynamically adjusted with IoU by using a more appropriate learning intensity in different ranges of GIoU values. So far, most of the BBR losses in OOD have been transferred from horizontal target detection (HOD) methods. A unified transfer strategy (Qian et al. 2023b) (UTS) is proposed to facilitate the transfer of BBR loss from HOD to OOD.

While improving bounding box representations, existing methods often overlook the unique prior knowledge in remote sensing scenes. LSKNet (Li et al. 2023c) dynamically adjusts its large spatial receptive field to better simulate the distance measurement environment of various objects in remote sensing scenes. A new anchor-free orientation proposal generator (Cheng et al. 2022) (AOPG) is proposed to generate coarse orientation boxes in an anchor-free manner using a coarse orientation module (CLM), and then refine them into high-quality orientation proposals. In the process of oriented bounding box regression, inconsistencies can arise. B2V (Wang et al. 2023b) introduces the tanimoto coefficient to assess the similarity of bounding box vectors in terms of shape and orientation perception. To address the conflicts between classification, location regression, and angle regression tasks during training. A task interleaving and orientation estimation detector (Ming et al. 2023) (TIOEDet) is proposed, with a posteriori hierarchical alignment (PHA) labeling to optimize the detection pipeline and solve the problem of mismatch between the classification sub-task and localization sub-task. A simple and efficient bounding box representation (Yao et al. 2023) is proposed to refine the final localization results using the proposed new bounding box representation, thus fully releasing the capability of the orientation detector. To solve the mismatch between mainstream detector architectures and model optimization strategies in small object detection, CAF2ENet (Ge et al. 2024b), differing from previously

introduced optimization approaches, adjusts sample weights to guide the optimizer to prioritize higher-quality detection boxes.

4.3.3 More efficient loss functions

BBR plays a pivotal role in many target detection algorithms, as it directly impacts the localization accuracy and regression speed of CNNs. However, existing bounding box regression losses have two main drawbacks. First, non-vanilla losses do not match the evaluation metric IOU (Intersection over Union), resulting in poor regression performance. SESANet (Ma et al. 2024b) overcomes the limitations of the IoU loss function by generating high-quality positive samples using ADM, offering a strategy for evaluating different positive samples. In response to this, the latest algorithms propose some optimization ideas. For example, the Manhattan Distance IOU (Shen et al. 2022) loss function has been proposed to alleviate large and unstable gradients in the early stage of regression. Additionally, IoU-based metrics are very sensitive to position deviations of small objects, which can significantly degrade detection performance when used in anchor-based detectors. To address this issue, Xu et al. (2022b) proposed a new evaluation metric called Normalized Wasserstein Distance (NWD) and a new Ranking-based Assignment (RKA) strategy for tiny target detection.This method significantly improves the label assignment and provides sufficient supervised information for network training. Based on the same idea, Wang et al. (2021) proposes a new evaluation metric for minutiae target detection using Wasserstein distance.Wang et al. (2022e) introduces Hausdorff distance and combines it with IoU as a new evaluation metric (HIOU). And. Contextual information of position confidence is considered and the Contextual Maximum Selection NMS(Wang et al. 2022) (Cms-NMS) algorithm is proposed.

These improvements mainly focus on horizontal targets and do not address the multi-directional targets commonly found in aerial images. SOOD (Hua et al. 2023) designs two loss functions that focus on target direction and image layout, applying consistency regularization to directional differences and regularizing similarity. However, this approach gives less consideration to object size in orientation modeling. GALoss (Doloriel and Cajote 2023) uses instance segmentation masks as ground truth to learn attention features needed for improved small object detection and designs an attention network composed of two types of losses. The other drawback of existing bounding box regression losses is that circular smooth labels (CSL) used for angle prediction can overwhelm the target angle category when summing over all negative angle categories, preventing the network from predicting accurate angle information. To solve this problem, Wang et al. (2022f) proposed a new loss function called GFL, which is a more effective alternative to classification-based rotation detectors. Additionally, there is a negative correlation between loss gradients and angle errors. The aforementioned methods do not consider the scale sensitivity of the optimization process for rotated IoU loss. GCL (Ming et al. 2024) proposes a gradient correction loss that optimizes rotated IoU loss through gradient analysis and correction.

4.4 Transfomer-base detection methods

Over the past decade, deep learning-based algorithms have been extensively applied to various fields of remote sensing image analysis. In recent years, transformer-based architectures, initially introduced in natural language processing, have gained widespread popular-

ity in computer vision. Self-attentive mechanisms have replaced the prevalent convolutional operators to capture long-term dependencies. Table 7 lists the problems addressed by the models reviewed in this section and the advantages of the models. Among these, Vision Transformers (ViTs) (Dosovitskiy et al. 2020) have exhibited impressive performance in several computer vision tasks. ViTs utilize a self-attentive mechanism that effectively captures global interactions by learning the relationships between sequence elements. It extracts global features, enhancing the semantic representation of detected objects and thereby improving detection accuracy (Table 7).

In recent years, transformer-based methods have shown promising progress in target detection by eliminating post-processing steps such as NMS and enhancing depth char-

Table 7 A briefly overview optimization algorithms for Transfomer-base detection methods

Methods	Problem solved	Optimization Strategies
Scene text detection (Wang et al. 2023c)	Complex background	Based on a few representative features
GANsformer (Zhang et al. 2022f)	Complex background	Combining GAN and transformer
SFRNet (Cheng et al. 2023a)	Complex background	Spatial and channel transformer
STD (Yu et al. 2024c)	Complex background	Divide-and-conquer approach
Hyneter (Chen et al. 2024b)	Complex background	HNB+DS
TPH-YOLO (Zhu et al. 2021a)	Complex background	Transformer prediction head
ETAM (Zhang et al. 2023f)	Complex background	Ensemble learning
SASS (Zhang et al. 2024d)	Complex background	Occlusion image modeling
RODFormer (Dai et al. 2022b)	Boundary-arbitrary discontinuity	A spatial-FFN feedforward network
RingMoLite (Wang et al. 2024c)	Boundary-arbitrary discontinuity	Dual-branch structure with Transformer modules
TransConvNet (Liu et al. 2022g)	Boundary-arbitrary discontinuity	Combines CNN and self-attention
AO2-DETR (Dai et al. 2022)	Boundary-arbitrary discontinuity	Generates explicit orientation suggestions
O2DETR (Ma et al. 2021)	Boundary-arbitrary discontinuity	Apply Transformer to locate objects
EFNet (Liu et al. 2022e)	Unreliable bounding boxes	FCF+RCF
TRD (Li et al. 2022b)	Inadequate expression	Aggregate multiple scales of features
LPSW (Xu et al. 2021)	Inadequate expression	Combining transformers and CNNs
TransMIN (Xu et al. 2023)	Inadequate expression	LGFI+CVFI
ViT-YOLO (Zhang et al. 2021c)	Inadequate expression	MHSA-Darknet+BiFPN
AST (He et al. 2023)	Inadequate expression	Collaboratively learn global dependencies
PCViT (Li et al. 2024a)	Inadequate expression	Feature Flow Pyramid Network

acterization. However, these methods have limitations in handling scene text due to the wide variation in scale and aspect ratio. Scene text detection (Wang et al. 2023c) based on a few representative features reduces computational costs and avoids interference from the background. To address these limitations, The transformer architecture introduced in GANsformer (Zhang et al. 2022f) is incorporated as a branching structure within the network, allowing for the extraction of region-wide feature information through attention mechanisms. The main difference between transformer-based and CNN-based detection methods is in the feature extraction method and the gap between local and global propagation of the feature map, which leads to poorer detection of small targets. To emphasize discriminative information advantageous for fine-grained classification, SFRNet (Cheng et al. 2023a) introduces a spatial and channel transformer designed to capture long-range spatial interactions and essential correlations embedded in feature channels. STD (Yu et al. 2024c) effectively leverages the spatial transformation potential of ViT using a divide-and-conquer approach.

Unlike previous divide-and-conquer strategies, Hyneter (Chen et al. 2024b), unlike the previous divide-and-conquer strategy, consists of a hybrid network backbone (HNB) and a dual switching module (DS) that integrates local and global dependent information and transmits them simultaneously. Based on a balanced strategy, the HNB extends the scope of local information by embedding convolutional layers in the Transformer block, while the DS adjusts the over-reliance on global dependencies outside the patch. Unlike adding an optimization mechanism to the backbone network, TPH-YOLO (Zhu et al. 2021a) explores prediction potential using a self-attentive mechanism by replacing the prediction head with a transformer prediction head. However, the above methods sacrifice some detection capability for large objects. To balance the detection of both small and large objects, ETAM (Zhang et al. 2023f) employs ensemble learning in its encoder, which has two branches. SASS (Zhang et al. 2024d) explores the inherent robustness of ViT and uses recent occlusion image modeling as a pretext task, further refining network pre-training into a streamlined end-to-end framework.

To tackle the challenges posed by the lack of local spatial perceptual field and discontinuous boundary loss in transformer-based rotating target detection, Dai et al. (2022b) proposes RODFormer, a high-precision model that uses a structured transformer architecture to collect feature information at different resolutions and improve the range of feature information collection. By integrating the strengths of both transformers and CNNs, RingMoLite (Wang et al. 2024c) employs a dual-branch structure with Transformer modules serving as low-pass filters to extract global features from remote sensing images, and CNN modules as stacked high-pass filters to efficiently capture fine-grained details. Moreover, the hybrid network TransConvNet (Liu et al. 2022g) combines CNN and self-attention based network advantages to focus more on global and local information aggregation while making up for CNN's deficiency with strong contextual focus on rotation invariance. It adapts to arbitrary orientation of detection targets in remote sensing images. Dai et al. (2022) proposes AO2-DETR, an arbitrary object-oriented detection converter framework that generates explicit orientation suggestions to provide better location prior for feature aggregation in the transformer decoder through cross-attention modulation. Transformer-based O2DETR (Ma et al. 2021) is an end-to-end network that directly and efficiently locates objects using the transformer, eliminating the tedious process of traditional rotating anchor points. Directional localization tends to focus on rotation-sensitive features.

Inspired by the cascaded attentional hawk-eye central concave mechanism, Liu et al. (2022e) proposes a novel attentional mechanism network (EFNet) with two central concaves. The front central fovea(FCF) module learns candidate object knowledge based on channel attention and spatial attention, while the rear central fovea predicts refined objects without anchors with two subnets. The detection performance is enhanced through the combination of attentional information and the generation of adaptive anchor boxes using the attentional map. This approach allows for the modulation of the feature adaptation module, which transforms the feature mapping to effectively accommodate different anchor boxes. By incorporating attentional mechanisms and adapting the feature mapping accordingly, the model can better capture relevant information and optimize the detection process, leading to improved performance.

To improve the semantic relationships of remote sensing images at a distance, attention mechanism based transformers are used rather than CNNs. This paper proposes TRD (Li et al. 2022b), a combination of a CNN and a multilayer transformer with encoder and decoder, where an improved Transformer is designed to model interactions between pairs of instances and aggregate global spatial location features at multiple scales. Additionally, the Swin transformer is enhanced by combining the advantages of transformers and CNNs, resulting in a locally sensing Swin transformer (Xu et al. 2021) to improve the detection accuracy of small-scale targets. For local–global feature interaction (LGFI), a transformer-guided multi-interaction network (Xu et al. 2023) (TransMIN) is proposed to learn complementary features using convolution and transformers in the residual blocks in the backbone network. Cross-view feature interaction (CVFI) is also implemented by transformers in the FPN pyramid layer to capture the correlation between reference features and pyramid features.

Integrating self-attention mechanisms into the FPN structure significantly enhances the efficiency of contextual information extraction. Furthermore, the ViT-YOLO (Zhang et al. 2021c) architecture is modified by adding a multi-headed self-attention mechanism to the MHSA-Darknet backbone to retain sufficient global context information and extract more differentiated features for target detection. In this architecture, a modified version of Bidirectional Feature Pyramid Network(BiFPN) is utilized, building upon the traditional FPN. The enhanced structure facilitates effective cross-scale feature fusion in the neck network, where the paths are aggregated. Based on the design principles of cross-scale Transformer architectures, AST (He et al. 2023) introduces a pyramid structure to collaboratively learn global dependencies and local details, facilitating multi-level feature interactions and generating scale-invariant representations. PCViT (Li et al. 2024a) overcomes the limitations of existing Transformer methods on FPN structures by proposing a Feature Flow Pyramid Network to enhance contextual information exchange and further improve the network's ability to handle multi-scale information.

4.5 Real time detection methods

Real-time object detection is a crucial component of various practical applications. While deep learning methods have improved many issues in aerial image analysis, their high computational demands may limit their practical use in resource-constrained edge devices such as smart satellites and drones. Lightweight detection models are receiving increasing attention, but their performance often falls short compared to deep models. Knowledge

distillation is an effective technique for model compression without additional parameters. Existing fixed-threshold methods struggle to select optimal distillation samples due to the variation in object sizes and the complexity of features in remote sensing images. Table 8 lists the problems addressed by the models reviewed in this section and the advantages of the models. Yang et al. (2022c) proposes a statistical sample selection and multi-modal knowledge mining framework. The statistical sample selection module frames the task as modeling and segmenting the cost probability distribution of sample selection, better suited for the dynamic selection of multi-scale samples in remote sensing images, eliminating the distortions of previous static distillation selections. However, previous distillation methods are plagued by misleading background information in remote sensing images and neglect the study of relationships between different instances. InsDist (Li et al. 2023e) combines feature-based and relationship-based knowledge distillation to maximize the utilization of instance-related information during the knowledge transfer process from teacher to student (Table 8).

On one hand, knowledge transfer does not consider the attributes of instances. On the other hand, offline distillation methods often overlook the interactive learning between the student and teacher models. Such attribute-agnostic offline distillation strategies may not be suitable for instances in multi-scale, diverse, and complex remote sensing images, leading to suboptimal training states. To address these issues, DIL (Yang et al. 2022) proposes

Table 8 A briefly overview optimization algorithms for Real time detection methods

Methods	Problem solved	Optimization strategies
SSSMKM (Yang et al. 2022c)	Knowledge distillation	A statistical sample selection and multi-modal knowledge mining framework
InsDist (Li et al. 2023e)	Knowledge distillation	Combines feature-based and relationship-based
DIL (Yang et al. 2022a)	Interactive learning	Dynamic interactive learning framework
APDNet (Zhang et al. 2024b)	Interactive learning	Different depths of the backbone network
GHOST (Zhang et al. 2023g)	Interactive learning	One-to-one self-supervised guided mixed-quantization framework
CDFD (Gu et al. 2023)	Interactive learning	A novel context-aware dense feature distillation method
ALP (Zhang et al. 2024a)	Boundary discontinuity	Directional distillation
JDBNet (Xie et al. 2023a)	Boundary discontinuity	Joint-guided distillation
RoMP (Moon et al. 2024)	Boundary discontinuity	Multi-level feature pyramid transformer
PVT (Zhang et al. 2023h)	Computational complexity	Adaptive multi-granularity routing mechanism
A2Net (Li et al. 2023f)	Computational complexity	Progressive feature aggregation and supervised attention
SOCDet (Pang et al. 2023)	Computational complexity	Single-kernel full-dimensional dynamic convolution
AMFLWYOLO (Peng et al. 2023)	Computational complexity	Depthwise separable convolutions, inverted residuals, and linear bottleneck structures

a dynamic interactive learning framework to optimize lightweight detectors. Regarding the selection of distillation positions within the network structure, APDNet (Zhang et al. 2024b) employs distillation networks between different depths of the backbone network to reduce the floating-point operations of network parameters. GHOST (Zhang et al. 2023g) introduces a one-to-one self-supervised guided mixed-quantization framework, achieving a lightweight model through the synergy of quantization and distillation. Existing methods attempt to eliminate redundant parameters, but this inevitably leads to a decrease in detection accuracy. CDFD (Gu et al. 2023) proposes a novel context-aware dense feature distillation method that guides small student networks to integrate features extracted from multiple teacher networks, training lightweight and high-performance airborne remote sensing object detection detectors.

Compared to general detectors, directional detectors for remote sensing images require additional prediction branches to estimate orientation angles, making the network more complex. Due to the boundary discontinuity caused by the periodic nature of the rotation angle in rotated bounding boxes, ALP (Zhang et al. 2024a) proposes an effective knowledge distillation method called directional distillation, which addresses this issue through reverse blurring spatial transformations. However, traditional binary quantization limits the representational capacity of network parameters. JDBNet (Xie et al. 2023a) introduces a joint-guided distillation binary neural network, which alleviates the impact of quantization errors through dynamic channel diversity enhancement. On the other hand, RoMP (Moon et al. 2024) proposes a multi-level feature pyramid transformer for rotated bounding boxes, utilizing rotated bounding boxes to minimize the impact of background on feature map construction. It optimizes real-time performance through Bayesian optimization and semi-tensor weight reduction.

The optimization methods introduced in previous sections undoubtedly increase computational complexity. To address the problem of low training and sample efficiency caused by a lack of inductive bias, PVT (Zhang et al. 2023h) proposes an adaptive multi-granularity routing mechanism that promotes token sparsity in transformers, significantly reducing computational costs without affecting accuracy. The high memory and computational costs of CNN methods hinder their successful application in practical scenarios. A2Net (Li et al. 2023f) introduces a new lightweight network based on mobile networks, which uses progressive feature aggregation and supervised attention to extract features for change detection. To maintain generalization capabilities for remote sensing image interpretation after using lightweight solutions, RingMoLite (Wang et al. 2024c) employs a dual-branch structure combining a Transformer module as a low-pass filter to extract global features and a CNN module as stacked high-pass filters to effectively capture fine-grained details. Although previous optimization methods enhance convolutional operators, they often make the network too cumbersome, making it challenging to adapt to resource-constrained edge devices. SOCDet (Pang et al. 2023) designs efficient computing units and uses single-kernel full-dimensional dynamic convolution to build the network. AMFLWYOLO (Peng et al. 2023) replaces standard convolutional layers with depthwise separable convolutions, inverted residuals, and linear bottleneck structures to reduce model parameters in the backbone network.

5 Experimental results

To facilitate the evaluation of small target detection algorithms in aerial remote sensing images, this section provides an overview of commonly used public datasets and standard performance evaluation metrics for detection algorithms. We will then select representative datasets and compare the performance of detection algorithms using these datasets.

5.1 Benchmark datasets

GLH-Bridge Li et al. (2024b) is a large-scale dataset collected from multiple satellite sensing platforms (such as Google Earth and MapBox). The images cover various sizes ranging from 2048 $\times$ 2048 to 16,384 $\times$ 16,384 pixels, with spatial resolutions from 0.3 to 1.0 m. These bridges span various backgrounds, totaling 59,737 bridges. Each bridge is manually annotated using OBB and HBB.

NVD Mokayed et al. (2023) primarily consists of aerial videos captured in snowy weather conditions in northern Sweden. The collection altitude ranges from 120 to 250 ms, with varying snow cover and cloud conditions. The annotated videos include a total of 8450 labeled frames, containing 26,313 labeled cars. The ground sampling distance or pixel size ranges from 11.1 to 22.2 cms.

LH-UAV-Vehicle Ying et al. (2024) consists of 18,663 optical images. The collection process involved determining the flight altitude and GPS location of the drone, capturing multiple remote sensing images from different directions at specified locations, and using a synthesis algorithm to generate panoramic images. The images were taken at altitudes ranging from 250 to 400 ms. Specific instance categories include sedans, vans, buses, and trucks. The dataset contains 180,358 manually annotated bounding boxes.

DTOD Zhao et al. (2024a) contains 11,600 images and 1,019,800 instances, with the average absolute size of the targets being less than 13 pixels. Each image contains an average of 88 targets.

VisDrone2018 Zhu et al. (2018) consists of 263 video clips and 10,209 images with rich annotations, including object bounding boxes, object classes, occlusion, truncation ratios, and more. To facilitate the evaluation and investigation of vision analysis algorithms on UAV platforms, this dataset contains over 2.5 million annotated instances in 179,264 image and video frames. The images were captured by the camera on the UAV in 14 different cities and villages and contain 10 categories of targets, including pedestrians, people, cars, vans, buses, trucks, bicycles, awning tricycles, and tricycles. The biggest challenge presented by this dataset is the difficulty in distinguishing between similar classes, such as dividing people into pedestrians and people based on pose variation, or distinguishing between canopy tricycles and regular tricycles based on morphological similarity.

UAV-DT Du et al. (2018) dataset was curated from a comprehensive collection of 10 h of raw aerial video footage, comprising approximately 80,000 representative frames. Each frame was meticulously annotated with bounding boxes, encompassing diverse attributes including but not limited to weather conditions, flight altitude, camera view, vehicle class, and occlusion. The flight altitude denotes the elevation at which the unmanned aerial vehicle operates, thereby influencing the scale variation of the detected targets. Notably, three distinct altitude levels, namely low-altitude, medium-altitude, and high-altitude, were meticu-

lously annotated to provide a comprehensive understanding of the target dynamics across different altitude ranges.

UAV-BD Wang et al. (2018) collected 25,407 UAV images with different backgrounds. Unlike traditional horizontal bounding box-based annotation methods, the dataset used OBB for accurate and compact bottle annotation that can be used for rotational target detection. The fully annotated images contain 34,791 bottles, each annotated by an arbitrary quadrilateral. To collect bottle images covering a wide range of scales and aspect ratios, images were collected for different flight heights from 10 to 30 ms. Most of the bottles were between 500 and 3000 pixels in size and the bottle scales ranged from 1.0 to 4.0.

DIOR Li et al. (2020b) dataset comprises a comprehensive collection of 23,463 images, featuring a total of 192,472 instances, spanning across 20 distinct object classes. These images were meticulously acquired under diverse imaging conditions, encompassing variations in weather, season, image quality, and other pertinent factors. Notably, the dataset exhibits a remarkable balance between inter-class similarity and intra-class diversity, enabling robust evaluation of object detection algorithms. The images within the dataset are standardized to a spatial resolution of 800 $\times$ 800 pixels, with a corresponding ground sampling distance ranging from 0.5 to 30 m. This resolution is consistent with the prevailing standards in existing benchmark datasets, facilitating seamless comparison and benchmarking against state-of-the-art methods.

AU-AIR Bozcan and Kayacan (2020) caters to UAV vision and robotics, with multimodal data from different on-board sensors. The dataset consists of 8 video streams for traffic monitoring. The videos were captured at different flight heights from 5 to 30 ms and different camera angles from 45 to 90 degrees. The entire dataset consists of 32,823 tagged video frames with object annotations and the corresponding flight data. Eight object classes are annotated, including people, cars, vans, trucks, motorcycles, bicycles, buses, and trailers. The total number of annotated instances is 132,034.

DOTAv2 Ding et al. (2022) is an expanded dataset based on DOTAv1 Xia et al. (2018). The dataset consists of a vast collection of 11,268 aerial images, with a total of 1,793,658 object instances annotated using OBB. These object instances span across 18 diverse categories, encompassing a wide range of classes. Notably, the dataset presents several challenging aspects, including the presence of object instances with arbitrary orientations, varying density distributions, and a diverse set of aerial scenes sourced from multiple image sources. The sheer number of object instances within the dataset adds to its complexity, providing a rich and comprehensive testbed for evaluating object detection algorithms.

RSOC dataset Gao et al. (2021) dataset comprises a total of 3057 images, accompanied by 286,539 annotations, which collectively represent four major categories: buildings, ships, large vehicles, and small vehicles. Notably, the dataset exhibits significant variation in terms of object scales, with object sizes ranging from small objects spanning tens of pixels to large objects encompassing thousands of pixels. This wide range of scales presents a diverse set of challenges for object detection and recognition algorithms, necessitating the development of robust and scalable methods to accurately detect and classify objects across different size ranges within the dataset.

MOHR Zhang et al. (2021a) aims to perform high-resolution multi-scale object detection on UAV images. A total of 9014 object instances with labels and bounding boxes were annotated. A total of 10,631 images were collected, including 90,014 instances and five categories. MOHR includes five categories such as cars, trucks, buildings, collapses, and

flood damage. The diversity of resolution is also a key attribute of our dataset. The drones are available at heights of 200, 300 and 400 ms, and the highest resolution images were taken at an altitude of 400 ms above the ground.

UAV-ROD Zhou et al. (2022b) dataset is composed of 1577 aerial images that were captured using UAVs. The dataset includes a total of 30,090 annotated vehicle instances, where the annotations are provided using the directed bounding box method. The UAV-ROD dataset serves as a valuable resource for tasks related to rotating target detection, vehicle orientation recognition, and target counting. Researchers can utilize this dataset to develop and evaluate algorithms that aim to address these specific challenges in aerial imagery analysis.

FAIR1M-High Sun et al. (2022) Resolution Remote Sensing Image Dataset has over 1 million instances and over 40,000 images of high resolution remote sensing images for fine-grained target identification. We collected remote sensing images with resolutions from 0.3 to 0.8 m from different platforms and labeled 5 categories and 37 subcategories by directed bounding boxes.

AI-TOD-v2 Xu et al. (2022b) contains 8 classes with 28,036 images and 752,754 instances. AI-TOD-v2 is specifically designed for the detection of tiny targets in aerial images. the average absolute object size of AI-TOD-v2 is only 12.7 pixels, and about 86% of the instances are smaller than 16 pixels, which is much smaller than the existing dataset.

Manipal-UAV dataset Akshatha et al. (2023) dataset consists of a curated collection of 13,462 sampled images extracted from 33 videos. These videos were recorded in an unconstrained environment, capturing complex scenes with a diverse range of human objects. In total, the dataset contains 153,112 instances of human objects, providing a substantial amount of annotated data for research and development purposes. The dataset encompasses various challenges encountered in real-world scenarios, including small object sizes, variations in scale, diverse poses, challenging lighting conditions, and occlusions. Researchers can leverage the Manipal-UAV dataset to advance the field of object detection and tracking, specifically in the context of human objects in aerial imagery.

TinyPerson Yu et al. (2020) dataset contains 1610 labeled images and 759 unlabeled images (both mainly from the same video set), with a total of 72,651 instances.The TinyPerson dataset categorizes people as either a 'sea person' or an 'earth person'. earth person'. Four rules are defined to determine which label a person belongs to: 1) a person on a boat is considered as a 'sea person'; 2) a person lying in the water is considered as a 'sea person'; 3) a person with more than half of his/her body in the water is considered as a 'sea person'; 4) a person with more than half of his/her body in the water is considered as a 'earth person'. (3) a person whose body is more than half in the water is regarded as 'sea person'; (4) other people are regarded as 'earth person'.

HazyDet dataset Feng et al. (2024) is a large-scale dataset designed for UAV-based object detection in hazy scenes. It encompasses 383,000 real-world instances collected from both naturally hazy environments and normal scenes, with synthetic haze effects applied to simulate adverse weather conditions. The dataset consists of 11,000 synthetic images, each with a resolution of 1333×800 pixels, totaling 365,000 object instances. It is partitioned into training, validation, and testing subsets in an 8:1:2 ratio, featuring object categories such as cars, trucks, and buses. In addition to the synthetic data, the dataset also includes 600 images captured under foggy conditions. The integration of synthetic and real-world data, characterized by high object density, ensures that our dataset serves as a high-quality resource suitable for the comprehensive evaluation of various detection models.

Table 9 lists the current datasets and their statistics that are mainly used for target detection tasks in remotely sensed aerial imagery.

5.2 Evaluation metrics

The standard evaluation metrics commonly used in target detection tasks are recall (R), precise (P), average precision (AP), mean average precision (mAP), and frame per second (FPS) for detection speed evaluation. Among them, the intersection of union (IOU) is measured by calculating the ratio of the intersection area between the truth region (GT) and the proposed result (PR) region to the area of the concatenation, defined as follows:

$$IoU = \frac{area(GT \cap PR)}{area(GT \cup PR)} \tag{1}$$

By setting the threshold value of IOU to determine whether the target is detected or not, the number of true positives (TP), false negatives (FN), false positives (FP) and true negatives (TN) are classified according to the different results of the judgment. Recall and precise are defined as follows:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

where TP+FN denotes the number of generated bounding boxes and TP+FP denotes the total number of objects to be detected. Thus, P calculates the number of correctly detected objects out of the total number of objects, while R calculates the number of detected objects. AP considers both accuracy and recall, and is defined as follows:

$$AP = \int_0^1 P(R)dR \tag{4}$$

It represents the average precision between 0 and 1 for the recall of detecting the same class of objects. mAP represents the average precision of all classes, expressed as follows:

$$mAP = \frac{1}{D}\sum_{i=1}^{D} AP(i) \tag{5}$$

where N denotes the number of detected object classes and i represents one of them. mAP metric takes into account the object detection accuracy in a comprehensive way. Therefore, the higher the mAP value, the more accurate the detector is.

In addition to evaluating the accuracy of object detection, the speed of the detection process plays a crucial role in comprehensive performance evaluation. The detection speed is typically measured using two metrics. Firstly, the number of frames processed per second

Table 9 Statistical data comparison of aerial remote sensing datasets, where HBB is horizontal bounding box and OBB is oriented bounding box

Dataset	Year	Images	Instances	Resolution	Categories	Annotation	Acquisition method
NVD (Mokayed et al. 2023)	2023	8450	26.3k	1920 × 1080	1	HBB	UAV
LH-UAV-Vehicle (Ying et al. 2024)	2024	18,663	180.3k	640 1024	4	HBB	UAV
DTOD (Zhao et al. 2024a)	2024	11,600	1019.8k	4700 × 2700	2	HBB	UAV
VisDrone (Zhu et al. 2018)	2018	10,209	54.2k	2000 × 1500	10	HBB	UAV
UAV-DT (Du et al. 2018)	2018	80,000	841.5K	1000	3	HBB	UAV
UAV-BD (Wang et al. 2018)	2018	25,407	34,791	500 3000	8	OBB	UAV
AU-AIR (Bozcan and Kayacan 2020)	2020	32,823	132,034	1920 × 1080	8	HBB	UAV
TinyPerson (Yu et al. 2020)	2020	1610	72,651	–	5	HBB	UAV
UAV-ROD (Zhou et al. 2022b)	2022	1,577	30.0K	1920	1	OBB	UAV
Manipal-UAV (Akshatha et al. 2023)	2023	13,462	153,112	1280 × 720	5	HBB	UAV
HazyDet-Dataset (Feng et al. 2024)	2024	11,000	383?000	1333 × 800	3	HBB	UAV
DOTAv2 (Ding et al. 2022)	2021	11,268	1,793,658	800-20,000	17	OBB	Satellite
RSOC (Gao et al. 2021)	2021	3057	286,539	800–2000	4	OBB/Center Point	Satellite
MOHR (Zhang et al. 2021a)	2021	10,631	90,014	5472,7360,8688	5	HBB	Satellite
DIOR (Li et al. 2020b)	2018	23,463	192,472	800	20	HBB	Satellite
FAIR1M (Sun et al. 2022)	2022	42,796	1020k	600–10000	37	OBB	Satellite
AI-TOD-v2 (Xu et al. 2022b)	2022	28,036	752,754	800	8	HBB	Satellite
GLH-Bridge (Li et al. 2024b)	2024	6000	59.7k	2,048 16,384	1	HBB/OBB	Satellite

(FPS), which represents the rate at which images can be detected. A higher FPS indicates a faster detection process. Secondly, the time taken to process individual images can be used as a measure of speed. This metric quantifies the computational time required to perform object detection on a single image. By considering both accuracy and speed, researchers can assess the overall efficiency and effectiveness of a detection algorithm in real-world applications.

5.3 Object detection results

In this section, we use the benchmark dataset to illustrate the detection performance of representative algorithms in remotely sensed aerial images.The VisDrone statistics are shown in the Table 10. The numbers bolded in Table 10 represent the highest performance effects Specifically, AP is computed by averaging over all 10 Intersection over Union (IoU) thresholds with the uniform step size 0.05) of all categories, which is used as the primary metric for ranking.AP50 and AP75 are computed at the single IoU thresholds 0.5 and 0.75 over all categories. The AR1,AR10,AR100 and AR500 scores are the maximum recalls given 1,10, 100 and 500 detections per image respectively, averaged over all categories and IoU thresholds.

In the Table 10, boldface indicates the highest performance. Hyneters (Chen et al. 2024b) is a hybrid network Transformer composed of a hybrid network backbone and a dual exchange

Table 10 Object detection results on the VisDrone-DET test-dev set * indicates that the detection algorithm is submitted by the VisDrone committee

Methods	Dataset year	AP	AP50	AP75	AR1	AR10	AR100	AR500
Light-RCNN* (Li et al. 2017)	2019	16.53	32.78	15.13	0.35	3.16	23.09	25.07
FPN* (Lin et al. 2017a)	2019	16.51	32.20	14.91	0.33	3.03	20.72	24.93
Cascade R-CNN* (Cai and Vasconcelos 2018)	2019	16.09	31.91	15.01	0.28	2.79	21.37	28.43
DetNet59* (Li et al. 2018)	2019	15.26	29.23	14.34	0.26	2.57	20.87	22.28
RefineDet* (Zhang et al. 2018)	2019	14.90	28.76	14.08	0.24	2.41	18.13	25.69
RetinaNet* (Lin et al. 2017b)	2019	11.81	21.37	11.62	0.21	1.21	5.31	19.29
Cascade GDF (Zhang et al. 2020b)	2018	18.7	31.7	19.4	8.4	24.2	28.8	–
HRDNet (Liu et al. 2021b)	2019	35.51	62.00	35.13	0.39	3.38	30.91	46.62
ViT-YOLO (Zhang et al. 2021c)	2019	38.5	63.15	40.48	2.33	**14.93**	**48.04**	55.47
Hyneter-Max (Chen et al. 2024b)	2021	**46.1**	**73.9**	**47.0**	–	–	–	–
Focus and Detect (Koyun et al. 2022)	2018	42.06	66.12	44.64	–	–	–	–
AdaZoom(ResNeXt-101) (Xu et al. 2022a)	2019	40.33	66.94	41.77	–	–	–	–
GLSAN(ResNet50) (Deng et al. 2021)	2019	32.5	55.8	33.0	–	–	–	–
VSRNet(ResNext-101) (Ge et al. 2022)	2021	39.0	66.5	39.7	–	–	–	–
DMNet(ResNext-101) (Li et al. 2020a)	2018	29.4	49.3	30.6	–	–	–	–
DMNet cropping+DSHNet (Yu et al. 2021)	2020	30.3	51.8	30.9	–	–	–	–
VistrongerDet (Wan et al. 2021)	2018	38.77	64.28	40.24	0.77	8.10	43.23	55.12
TPH-YOLOv5 (Zhu et al. 2021a)	2021	39.18	62.83	41.34	**2.61**	13.63	45.62	**56.88**
mSODANet-E6, E7 (Chalavadi et al. 2022)	2019	36.89	55.92	37.41	1.15	11.36	37.25	48.92

module. By integrating local information with global dependencies and transmitting them simultaneously, it achieves superior performance metrics in AP, AP50, and AP75 compared to other network models. TPH-YOLOv5 (Zhu et al. 2021a) enhances its capability to detect objects at various scales by incorporating an additional prediction head. Subsequently, replacing the original prediction heads with Transformer Prediction Heads (TPH), which leverage the predictive potential of self-attention mechanisms to capture objects, leads to superior performance in recall metrics AR1 and AR500. The hybrid detector ViT-YOLO (Zhang et al. 2021c) integrates an efficient convolutional backbone network with multi-head self-attention for feature extraction. The introduction of multi-head self-attention facilitates the extraction of more discriminative features, thereby improving recall metrics AR10 and AR100 compared to other models.

The Table 11 presents the performance of algorithms proposed in aerial remote sensing image datasets for detecting objects of different sizes. The definitions of small, medium, and large detection targets adhere to the evaluation metrics introduced in MS COCO. In aerial remote sensing

Table 11 The performance of both the baseline and the proposed method in multi-scale object detection on the aerial remote sensing dataset

Methods	Backbone	AP_S	AP_M	AP_L	Dataset
PRDet (Leng et al. 2023)	ResNeXt-101	25.6	40.8	52.9	VisDrone2021-DET (Cao et al. 2021)
CDMNet (Duan et al. 2021)	ResNeXt-101	22.2	42.4	44.7	VisDrone2021-DET (Cao et al. 2021)
Clusterdet (Yang et al. 2019a)	ResNeXt-101	19.1	40.8	54.4	VisDrone2021-DET (Cao et al. 2021)
Focus-and-Detect (Koyun et al. 2022)	ResNeXt-101	32.0	47.9	54.5	VisDrone2021-DET (Cao et al. 2021)
CRENet (Wang et al. 2020)	Hourglass-104	25.6	45.3	58.7	VisDrone2021-DET (Cao et al. 2021)
HawkNet (Lin et al. 2020)	ResNet-50	34.1	59.6	41.6	UAVDT (Du et al. 2018)
RetinaNet (Lin et al. 2017b)	Resnet-50	32.0	62.1	50.9	UAVDT (Du et al. 2018)
Faster-RCNN (Ren et al. 2015)	Resnet-50	31.9	60.4	49.3	UAVDT (Du et al. 2018)
Cascade-RCNN (Cai and Vasconcelos 2018)	Resnet-101	32.6	59.5	48.7	UAVDT (Du et al. 2018)
DMNet (Li et al. 2020a)	ResNeXt-101	21.6	41.0	56.9	VisDrone2018-DET (Zhu et al. 2018)
HawkNet (Lin et al. 2020)	ResNet-50	19.9	36.0	39.1	VisDrone2018-DET (Zhu et al. 2018)
RetinaNet (Lin et al. 2017b)	Resnet-50	14.1	29.5	33.7	VisDrone2018-DET (Zhu et al. 2018)
Faster-RCNN (Ren et al. 2015)	Resnet-50	15.4	34.6	37.1	VisDrone2018-DET (Zhu et al. 2018)
Cascade-RCNN (Cai and Vasconcelos 2018)	Resnet-101	16.5	36.8	39.4	VisDrone2018-DET (Zhu et al. 2018)
RetinaNet (Lin et al. 2017b)	Resnet-50	2.5	20.8	42.7	DIOR (Li et al. 2020b)
Faster-RCNN (Ren et al. 2015)	Resnet-50	7.1	26.8	54.4	DIOR (Li et al. 2020b)
SSD (Liu et al. 2016)	VGG16	6.6	32.6	62.2	DIOR (Li et al. 2020b)
YOLOv5-X (Jocher et al. 2021)	Darknet53	6.2	23.9	35.6	DIOR (Li et al. 2020b)
YOLOv5-X+MFDF (Lv et al. 2022)	Darknet53	6.7	26.1	40.7	DIOR (Li et al. 2020b)
EFNet (Liu et al. 2022e)	Resnet-50	3.6	23.9	44.2	DIOR (Li et al. 2020b)
Clusterdet (Yang et al. 2019a)	ResNet-50	16.6	32.0	50.0	DOTA (Xia et al. 2018)
AFA-FPN (Wang et al. 2022a)	ResNet-101	78.21	84.77	80.43	RSSO (Wang et al. 2022a)

datasets, the size variation of objects is significant, and conventional baseline detectors struggle to adapt well to such high-scale variations, resulting in poor detection performance. To address this issue, networks with Feature Pyramid Network (FPN) structures are relatively more effective in localizing objects across different scales. Moreover, the high proportion of small objects poses a challenge for detection tasks. Due to their limited pixel coverage, extracting meaningful features from small objects is challenging, leading to significant differences in detection accuracy compared to medium or large objects. Experimental statistics indicate that methods incorporating attention mechanisms provide the network with additional semantic information about the objects, thereby improving the network's performance and achieving substantial progress.

As illustrated in the Table 12, models based on the YOLO series (Jocher et al. 2023; Wang and Liao 2024; Wang et al. 2024b) achieve an admirable balance between performance and real-time capability on the VisDrone dataset (Zhu et al. 2018). However, these detectors typically necessitate Non-Maximum Suppression (NMS) for post-processing, which not only decelerates inference speed but also introduces hyperparameters that can lead to instability in terms of both speed and accuracy. The end-to-end model, DETR (Carion et al. 2020), leverages a Transformer architecture to construct an end-to-end detector, enhancing small object detection capabilities.

Table 12 The computational efficiency and object detection results of network models on VisDrone-DET test-dev

Model	Input size	Parameters(M)	GFLOPs	AP	AP50
YOLOv8m (Jocher et al. 2023)	640 × 640	25.9	78.9	24.6	40.7
YOLOv8l (Jocher et al. 2023)	640 × 640	43.7	165.2	26.1	42.7
YOLOv9m (Wang and Liao 2024a)	640 × 640	20.1	76.8	25.2	42.0
YOLOv10m (Wang et al. 2024b)	640 × 640	15.4	59.1	24.5	40.5
PP-YOLOE-P2-Alpha-l (Authors 2019)	640 × 640	54.1	111.4	30.1	48.9
ClusDet (Yang et al. 2019b)	1000 × 640	30.2	207	26.7	50.6
HIC-YOLOv5 (Tang et al. 2024)	640 × 640	9.4	31.2	20.8	36.1
DETR (Carion et al. 2020)	1333 × 750	60	187	24.1	40.1
Deformable DETR (Zhu et al. 2021b)	1333 × 800	40.0	173	27.1	42.2
Sparse DETR (Roh et al. 2022)	1333 × 800	40.9	121	27.3	42.5
RT-DETR (Zhao et al. 2024b)	640 × 640	20.0	60.0	26.7	44.6
UAV-DETR (Zhang et al. 2025)	640 × 640	20.0	77.0	29.8	48.8

Nevertheless, its high computational cost and inferior real-time performance render it unsuitable for real-time applications. To address these limitations, RT-DETR (Zhao et al. 2024b) surpasses the popular YOLO framework (Jocher et al. 2023; Wang and Liao 2024; Wang et al. 2024b) in both accuracy and speed. Despite this, existing DETR (Carion et al. 2020) models are primarily designed for natural images, posing challenges when applied to UAV image detection tasks. UAV-DETR (Zhang et al. 2025), adopting a single-stage model akin to the DETR structure, executes multi-scale feature fusion across spatial and frequency domains. By employing learned offsets to align features across different fusion pathways, UAV-DETR (Zhang et al. 2025) resolves misalignment issues and enhances detection performance. Achieving superior accuracy at a computational cost of less than 100 GFLOPs, UAV-DETR (Zhang et al. 2025) outperforms all other methods within its category. Compared to other models, UAV-DETR (Zhang et al. 2025) distinguishes itself through two key aspects: Firstly, it eliminates the need for NMS and anchor settings, significantly reducing the complexity of model deployment. Secondly, by leveraging dual-domain information during the feature fusion process, UAV-DETR (Zhang et al. 2025) achieves remarkable detection performance.The Fig. 6 illustrates the detection performance of the highly efficient YOLOv8 (Jocher et al. 2023) model on sample images from the VisDrone dataset (Zhu et al. 2018). The results demonstrate its robust capability to effectively capture small objects while maintaining exceptional detection efficiency.

In order to gain a deeper understanding of the impact of loss functions designed for aerial remote sensing object detection on detection performance, the Table 13 presents the detection results using different improved loss functions in various baseline networks. The spatial distribution of objects in aerial remote sensing images differs significantly from that in natural scenes, and the dense distribution of objects leads to numerous occlusions and connections in the

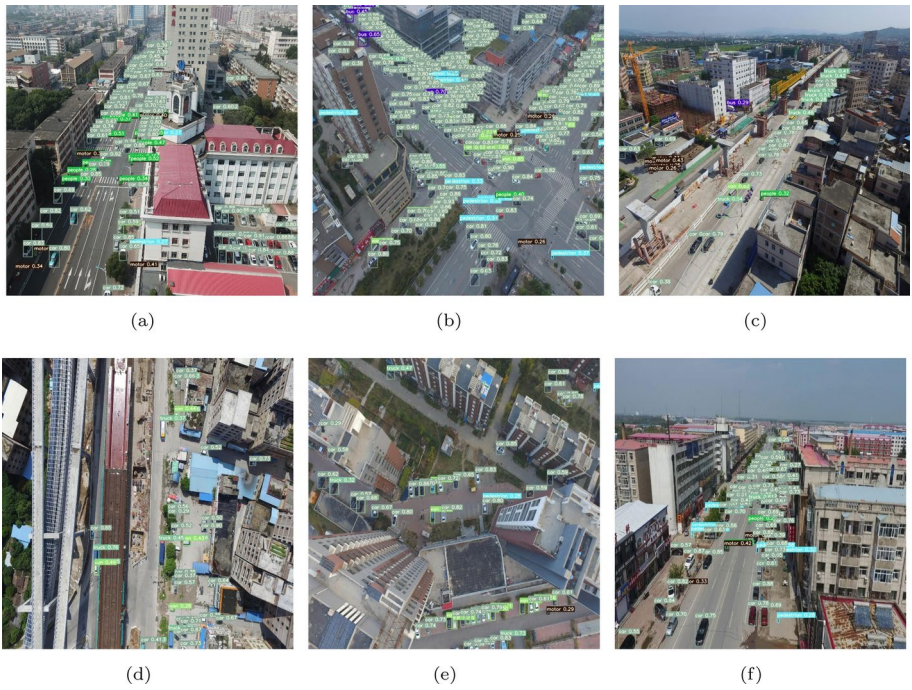

Fig. 6 The Detection Performance of YOLOv8 on the VisDrone Dataset

Table 13 The detection performance of the baseline network with the incorporation of the proposed improved loss function on the aerial remote sensing dataset

Methods	Backbone	Loss	mAP	AP50	AP75	Dataset
Faster-RCNN (Ren et al. 2015)	ResNet-50	Baseline	33.9	53.4	37.3	DOTA (Xia et al. 2018)
		MIoU-C (Shen et al. 2022)	34.7	55.7	37.4	DOTA (Xia et al. 2018)
		CIoU (Zheng et al. 2020b)	34.3	54.9	37.4	DOTA (Xia et al. 2018)
		EIoU (Zhang et al. 2022g)	34.2	54.8	37.4	DOTA (Xia et al. 2018)
	ResNet-101	HIOU (Wang et al. 2022e)	85.24	–	–	RSOD (Gao et al. 2021)
Mask-RCNN (He et al. 2017)	ResNet-50	Baseline	33.9	53.6	36.7	DOTA (Xia et al. 2018)
		MIoU-C (Shen et al. 2022)	34.8	55.6	37.7	DOTA (Xia et al. 2018)
		CIoU (Zheng et al. 2020b)	34.2	54.3	37.2	DOTA (Xia et al. 2018)
		EIoU (Zhang et al. 2022g)	34.4	55.1	37.8	DOTA (Xia et al. 2018)
RetinaNet (Lin et al. 2017b)	ResNet-50	Baseline	34.4	56.9	35.8	DOTA (Xia et al. 2018)
		MIoU-C (Shen et al. 2022)	36.6	58.2	38.8	DOTA (Xia et al. 2018)
		CIoU (Zheng et al. 2020b)	34.5	56.9	36.2	DOTA (Xia et al. 2018)
		EIoU (Zhang et al. 2022g)	35.1	57.5	37.3	DOTA (Xia et al. 2018)
S^2ANet (Han et al. 2021)	ResNet-50-FPN	PIoU (Chen et al. 2020)	72.42	–	–	DOTA (Xia et al. 2018)
		Smooth L1 (Girshick 2015)	74.12	–	–	DOTA (Xia et al. 2018)
		IoU (Zhou et al. 2019b)	74.64	–	–	DOTA (Xia et al. 2018)
		Smooth IoU (Rezatofighi et al. 2019)	75.69	–	–	DOTA (Xia et al. 2018)
		GIoU (Rezatofighi et al. 2019)	74.80	–	–	DOTA (Xia et al. 2018)
		Smooth GIoU (Qian et al. 2023a)	75.87	–	–	DOTA (Xia et al. 2018)
AOPG (Cheng et al. 2022)	ResNet-50-FPN	PIoU (Chen et al. 2020)	73.73	–	–	DOTA (Xia et al. 2018)
		Smooth L1 (Girshick 2015)	75.24	–	–	DOTA (Xia et al. 2018)
		IoU (Zhou et al. 2019b)	75.51	–	–	DOTA (Xia et al. 2018)
		GIoU (Qian et al. 2023a)	75.56	–	–	DOTA (Xia et al. 2018)
		RIoU (Qian et al. 2023b)	76.00	–	–	DOTA (Xia et al. 2018)

Table 13 (continued)

Methods	Backbone	Loss	mAP	AP50	AP75	Dataset
		FRIoU (Qian et al. 2023b)	76.45	–	–	DOTA (Xia et al. 2018)
R-CNN (Chen et al. 2017)	ResNet-50	NWD-Loss (Xu et al. 2022b)	12.4	–	–	AI-TOD (Xu et al. 2022b)

images. The label matching strategy and loss function design in traditional general detectors fail to effectively address the aforementioned issues. Additionally, the presence of a large number of small objects in the data and the differences between the objects in aerial remote sensing datasets and natural images necessitate the design of loss functions that better align with the target characteristic distribution in the aerial remote sensing dataset. Smooth GIoU (Qian et al. 2023a) can adapt to the spatial distribution of objects in the data and dynamically adjust the learning intensity, thereby improving localization accuracy. Designing more suitable distance metrics (Shen et al. 2022) also benefits the convergence speed and detection performance.

As shown in Tables 14 and 15, all the categories in DOTA are shown: Plane (PL), Baseball diamond(BD),Bridge(BR), Ground track field (GTF), Small vehicle (SV),Large vehicle (LV), Ship (SH), Tennis court (TC), Basketball court(BC), Storage tank (ST), Soccer-ball field(SBF),Roundabout(RA), Harbor (HA), Swimming pool (SP), and Helicopter(HC). Two annotation methods, HBB and OBB, were employed in the aerial remote sensing dataset DOTA. Due to variations in imaging angles, the detected objects are typically distributed in arbitrary orientations, and rotated bounding boxes enable the network to extract more informative features compared to HBB. However, when objects are in close proximity, angle regression poses new challenges to detection. To address this issue, an intuitive approach is to incorporate the object's rotation angle into the learning process and design a metric loss for angle regression, thereby enhancing the network's adaptive capability to handle arbitrary object orientations. This idea has been applied to region proposal-based network architectures and has shown promising regression performance based on experimental results. Among these methods, the ones designed based on anchor-free mechanisms effectively improve the detection performance of objects with multi-orientation characteristics. By learning the regression boxes based on the object's orientation features, the anchor-free approach avoids the issue of challenging angle regression caused by the generation of anchor boxes.

A total of twelve leading object detection models, encompassing single-stage, two-stage, and end-to-end methodologies, were evaluated on the HazyDet dataset (Feng et al. 2024). To ensure a fair comparison, all models were trained using the default schedule of 12 epochs, with the exception of Deformable DETR (Zhu et al. 2021b), which were trained for 50 epochs, and YOLOv3 (Redmon and Farhadi 2018) and YOLOX (Ge et al. 2021), which underwent 300 epochs of training. Test-time augmentation and multi-scale training were excluded, except for Deformable DETR Zhu et al. (2021b), which required enhanced data augmentation. All models were trained on the synthetic data from the HazyDet (Feng et al. 2024) training set and evaluated on its test set and Real-hazy Drone Detection Testing Set(RDDTS), using accuracy and efficiency as metrics. Detailed results can be found in the table.

Table 16 reveals consistent performance trends of the detectors on both the test set and RDDTS, demonstrating that the simulated environment effectively mirrors real-world hazy scenarios. The bolded numbers in Table 16 represent the highest performance effects While each detector exhibits strengths under foggy conditions, they also possess inherent limitations. Single-

Table 14 Performance Comparisons of HBB Task on DOTA

Methods	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
ACE(DLA34) (Dai et al. 2022a)	89.5	76.3	45.1	60.0	77.8	77.1	86.5	90.8	79.5	85.7	47.0	59.4	65.7	7137	63.9	71.7
AOPG(R101FPN) (Cheng et al. 2022)	89.98	96.14	60.20	79.55	78.47	84.93	88.79	90.88	87.32	87.07	71.50	71.22	83.57	72.47	70.77	80.19
AOPG+FRIoU (Qian et al. 2023b)	90.13	85.85	60.04	80.90	79.21	85.34	88.73	90.88	86.35	87.55	70.22	72.09	83.40	82.12	72.55	81.02
AOPDet(Resnet101) (Zhu et al. 2022)	90.2	77.9	47.5	74.3	73.6	78.7	89.5	90.9	66.4	85.7	57.1	71.9	75.6	72.8	60.9	74.3
YOLOv5m+AR-BCL (Xiao et al. 2022a)	89.41	85.12	49.22	65.92	80.62	84.58	88.71	90.64	88.12	87.97	64.95	68.3	74	81.42	70.98	78
ASSD+(VGG16) (Xu et al. 2022b)	89.2	82.1	53.1	72.8	79.3	82.7	88.3	90.8	84.5	84.9	64.9	68.3	80.5	80.6	64.6	77.8
FDLR-Net(ResNet152) (Xiao et al. 2023)	89.04	79.16	52.10	68.60	72.12	75.08	77.91	89.42	86.73	86.34	64.84	61.40	66.91	68.65	57.93	73.08
FoRDet (Zhang et al. 2022b)	89.50	85.50	40.53	72.37	69.44	76.92	79.55	89.45	72.74	79.98	76.44	67.15	64.25	60.76	60.75	72.35
GSDet(ResNet101) (Li et al. 2021a)	81.12	76.78	40.78	75.89	64.50	58.37	74.21	89.92	79.40	78.83	64.54	63.67	66.04	58.01	52.13	68.28
GFA-Net(ResNet101) (Zhu et al. 2022)	88.9	58.9	23.5	63.3	38.5	63.3	73.2	90.7	63.7	74.8	62.1	59.3	41.5	31.0	32.7	57.7
MRDet(ResNet101) (Qin et al. 2022)	89.49	84.29	55.40	66.68	76.27	82.13	87.86	90.81	86.92	85.00	52.34	65.98	76.22	76.78	67.49	76.24
R2YOLOX-X (Liu et al. 2022b)	88.12	85.93	55.38	76.61	80.46	85.21	88.45	90.88	88.51	81.44	67.92	61.13	78.62	77.69	77.72	79.33
CGL (Chen et al. 2023e)	89.50	84.19	55.86	76.24	78.68	82.52	88.03	90.90	86.67	85.03	62.96	68.95	76.69	70.69	68.18	77.12

stage detectors excel in speed and resource efficiency but often compromise on accuracy and generalization capabilities. Two-stage detectors offer superior detection accuracy at the cost of computational efficiency. End-to-end detectors streamline the process workflow but face challenges with complex training procedures. Current algorithms still hold significant potential for improving accuracy, particularly under real-world hazy conditions. Variations in detection accuracy among target types (e.g., cars, buses, trucks) highlight the challenges associated with the long-tail distribution in the dataset, necessitating further algorithmic enhancements. DeCoDet (Feng et al. 2024) outperforms most single-stage and two-stage detectors but is surpassed by the state-of-the-art end-to-end detector, Deformable DETR (Zhu et al. 2021b). However, these advanced detectors heavily rely on extensive data augmentation and longer training times, limiting their practical applicability. The DeCoDet detector (Feng et al. 2024) offers a notable performance advantage with fewer parameters.

6 Conclusion and prospect

The paper categorizes and analyzes various deep learning strategies designed to improve the detection of small targets in aerial images. These strategies include multi-scale aware methods, attention-based improvements, and techniques for enhancing image perception, combining contextual information, and optimizing feature alignment.The paper discusses the strengths and weaknesses of current detection algorithms and suggests potential directions for future research, including addressing issues like sample imbalance, scale diversity, and the need for more sophisticated feature extraction techniques.A comprehensive review of existing datasets used for small object detection in aerial images, along with a comparison of the performance of various algorithms on these datasets, is provided. This helps in understanding the current advancements and challenges in the field.

Despite numerous challenges in object detection in aerial images, the deepening of research allows for the adoption of methods from various fields to address these issues. The pixels occupied by the targets are few, and the resolution is low. Super-resolution techniques can be used to improve the quality of input images, aiding in the better detection of small objects. To mitigate interference from complex backgrounds, improving the extraction of contextual information near positive samples and developing robust background subtraction methods can help reduce clutter and background noise. In optimizing the matching mechanism of positive and negative samples, more sophisticated and rational NMS techniques can be utilized to manage overlapping detections more effectively. Additionally, using region-based methods is also a good solution. This approach can focus on densely populated target areas, improving detection in crowded scenes. Regarding the issues of viewpoint and scale changes, more comprehensive data augmentation strategies can be developed, including rotation, scaling, and flipping, to make the model robust to viewpoint and scale variations. For datasets with data imbalance, synthetic data can be generated to augment the training dataset, helping to balance class distribution and introduce more diverse scenarios. By leveraging transfer learning and domain adaptation techniques, transfer learning from related fields can be utilized to develop domain adaptation techniques, leveraging more comprehensive datasets. In practical applications, to efficiently deploy models, exploring edge computing solutions and optimizing on-device inference models is essential. Utilizing hardware accelerators like GPU or TPU and optimizing software frameworks can achieve

Table 15 Performance Comparisons of OBB Task on DOTA

Methods	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
DA-Net(ResNet101) (Li et al. 2022)	89.70	82.41	53.28	69.55	78.24	79.54	89.04	90.68	84.76	86.33	65.03	65.70	76.16	73.37	58.86	76.11
CFC-Net(ResNet-50) (Ming et al. 2022)	89.08	80.41	52.41	70.02	76.28	78.11	87.21	90.89	84.47	85.64	60.51	61.52	67.82	68.02	50.09	73.50
CNN-AOOF(Darknet53) (Dong et al. 2022d)	88.21	81.62	58.80	72.93	71.02	78.82	77.63	89.54	85.31	86.12	62.63	60.92	80.63	77.84	63.61	75.71
DARDet(ResNet50) (Zhang et al. 2022)	89.08	84.30	56.64	77.83	81.10	83.39	88.46	90.88	85.44	87.56	62.77	66.23	77.97	82.03	67.40	78.74
GF-CSL(ResNet101) (Wang et al. 2022f)	89.87	85.20	53.24	75.48	79.13	83.59	88.45	90.88	88.28	86.97	66.37	72.64	82.11	82.19	74.63	79.94
GSNet+FRM(RetinaNet) (Shen et al. 2023b)	88.92	82.79	51.93	69.53	79.13	79.16	87.26	90.85	82.19	85.12	55.34	66.73	71.28	70.46	56.64	74.49
ADT-Det(FPN) (Zheng et al. 2021b)	89.71	84.71	59.63	80.94	80.30	83.53	88.94	90.86	87.06	87.81	70.72	70.92	78.66	79.40	65.99	79.95
AProNet(ResNet101) (Zheng et al. 2021a)	88.77	84.95	55.27	78.40	76.65	78.54	88.45	90.83	86.56	87.01	65.62	70.29	75.43	78.17	67.28	78.16
DRB+FAB+CRB(ResNet101) (Yuan et al. 2022)	87.62	80.55	50.33	71.90	74.63	82.47	88.20	90.83	84.96	78.84	54.17	56.33	74.82	63.58	62.37	73.44
S^2ANet(ResNet101-FPN) (Han et al. 2021)	88.70	81.41	54.28	69.75	78.04	80.54	88.04	90.69	84.75	86.22	65.03	65.81	76.16	73.37	58386	76.11
CBDA-Net(DLA-34) (Liu et al. 2022a)	89.17	85.92	50.28	65.02	77.72	82.32	87.89	90.48	86.47	85.90	66.85	66.48	67.41	71.33	62.89	75.74
CenterOBB(DLA34) (Wang et al. 2022b)	90.87	83.91	54.39	74.16	78.45	84.39	87.25	91.82	85.63	85.73	68.32	68.27	72.82	74.63	67.14	77.85
RH-RCNN(ResNet50) (Yang et al. 2022b)	89.64	80.19	39.52	71.21	65.91	64.68	65.07	92.08	78.43	80.14	55.27	67.55	60.62	60.93	56.44	68.51
OrtDet(Transformer) (Zhao et al. 2022)	89.02	84.82	50.68	71.45	72.06	74.98	87.64	90.85	78.76	85.81	58.97	68.95	67.29	70.49	59.61	74.09
MGAR(DarkNet53) (Wang et al. 2022c)	89.84	85.75	51.59	77.00	76.38	74.81	86.40	90.73	87.70	87.48	63.25	69.70	75.79	80.88	71.07	77.85

Table 16 Comparison of the performance of different state-of-the-art detectors on the HazyDet dataset

Model	Parameters(M)	GFLOPs	AP on Test-set				AP on RDDTS			
			Car	Truck	Bus	mAP	Car	Truck	Bus	mAP
One-stage										
YOLOv3 (Redmon and Farhadi 2018)	61.63	20.19	36.1	21.4	47.5	35.0	30.2	7.1	20.4	19.2
GFL (Li et al. 2021b)	32.26	198.65	50.3	11.5	48.5	36.8	33.5	2.4	5.9	13.9
YOLOX (Ge et al. 2021)	8.94	13.32	53.1	23.0	51.2	42.3	48.0	11.0	17.7	24.7
RepPoints (Yang et al. 2019c)	36.83	184.32	52.7	24.6	54.2	43.8	42.4	5.0	16.5	21.3
FCOS (Tian et al. 2019)	32.11	191.48	54.4	27.1	56.2	45.9	43.3	8.7	16.4	22.8
Centernet (Duan et al. 2019a)	32.11	191.49	56.7	27.9	57.0	47.2	45.6	8.6	17.3	23.8
ATTS (Zhang et al. 2020c)	32.12	195.58	58.5	32.2	60.4	50.4	48.5	8.1	18.8	25.1
DDOD (Chen et al. 2021)	32.20	173.05	59.5	32.1	60.4	50.7	48.2	9.2	20.9	26.1
VFNet (Zhang et al. 2021d)	32.89	187.39	**59.6**	32.5	61.3	51.1	48.8	8.9	19.1	25.6
TOOD (Feng et al. 2021)	32.02	192.51	58.4	33.6	62.2	51.4	48.3	9.0	20.1	25.8
Two-stage										
Sparse RCNN (Sun et al. 2023)	108.54	147.45	33.0	14.2	35.6	27.7	20.0	3.4	7.8	10.4
Dynamic RCNN (Zhang et al. 2020d)	41.35	201.72	56.8	27.3	58.7	47.6	44.3	6.1	17.0	22.5
Faster RCNN (Ren et al. 2017)	41.35	201.72	56.3	30.5	59.3	48.7	44.0	7.9	19.0	23.6
Libra RCNN (Pang et al. 2019)	41.62	209.92	57.3	30.4	59.3	49.0	45.7	8.5	16.8	23.7
Grid RCNN (Lu et al. 2019)	64.46	317.44	58.1	32.8	50.7	50.5	46.5	10.1	18.9	25.2
Cascade RCNN (Cai and Vasconcelos 2018)	69.15	230.40	59.0	**34.2**	61.7	51.6	46.5	10.6	20.9	26.0
End2End										
Conditional DETR (Meng et al. 2021)	43.55	94.17	42.1	12.6	36.8	30.5	22.2	2.3	11.2	11.7
DAB DETR (Liu et al. 2022f)	43.70	97.02	36.8	15.1	42.3	31.3	22.2	2.3	11.2	11.7
Deform DETR (Zhu et al. 2021b)	40.01	192.51	58.8	34.1	**62.9**	**51.9**	46.3	**11.2**	**21.9**	**26.5**
Plug-and-play										
FCOS-DeCoDet (Feng et al. 2024)	34.61	249.91	55.9	28.6	57.6	47.4	48.1	11.1	17.8	24.3
VFNet-DeCoDet (Feng et al. 2024)	34.62	225.37	58.3	33.7	62.5	51.5	**49.0**	9.0	19.7	25.9

real-time performance. Furthermore, model compression techniques such as pruning, quantization, and knowledge distillation can be employed to reduce model size and inference time.

Acknowledgements This research was funded by National Natural Science Foundation of China (62103056); Qin Xin Talents Cultivation Program, Beijing Information Science & Technology University (QXTCP B202403).

Declarations

Conflict of interest The authors declare that they have no Conflict of interest.

References

Akshatha KR, Karunakar AK, Shenoy S, Dhareshwar CV, Johnson DG (2023) Manipal-uav person detection dataset: a step towards benchmarking dataset and algorithms for small object detection. ISPRS J Photogramm Remote Sens 195:77–89. https://doi.org/10.1016/j.isprsjprs.2022.11.008

Authors P (2019) PaddleDetection, object detection and instance segmentation toolkit based on PaddlePaddle. https://github.com/PaddlePaddle/PaddleDetection

Bai J, Ren J, Xiao Z, Chen Z, Gao C, Ali TAA, Jiao L (2023a) Localizing from classification: self-directed weakly supervised object localization for remote sensing images. IEEE transactions on neural networks and learning systems

Bai Z, Li G, Liu Z (2023b) Global-local-global context-aware network for salient object detection in optical remote sensing images. ISPRS J Photogramm Remote Sens 198:184–196. https://doi.org/10.1016/j.isprsjprs.2023.03.013

Bozcan I, Kayacan E (2020) Au-air: a multi-modal unmanned aerial vehicle dataset for low altitude traffic surveillance. In: 2020 IEEE international conference on robotics and automation (ICRA), p 8504–8510. IEEE

Cai Z, Vasconcelos N (2018) Cascade R-CNN: delving into high quality object detection. In: 2018 IEEE conference on computer vision and pattern recognition, CVPR 2018, Salt Lake City, UT, USA, June 18–22, 2018, p 6154–6162

Cai X, Lai Q, Wang Y, Wang W, Sun Z, Yao Y (2024) Poly kernel inception network for remote sensing detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, p 27706–27716

Cao Y, He Z, Wang L, Wang W, Yuan Y, Zhang D, Zhang J, Zhu P, Van Gool L, Han J, Hoi S, Hu Q, Liu M (2021) Visdrone-det2021: the vision meets drone object detection challenge results. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV) workshops, p 2847–2854

Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S (2020) End-to-end object detection with transformers. In: Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16, p 213–229. Springer

Chalavadi V, Jeripothula P, Datla R, Ch SB, (2022) msodanet: a network for multi-scale object detection in aerial images using hierarchical dilated convolutions. Pattern Recogn 126:108548. https://doi.org/10.1016/j.patcog.2022.108548

Chen C, Liu M-Y, Tuzel O, Xiao J (2017) R-cnn for small object detection. In: Computer vision–ACCV 2016: 13th Asian conference on computer vision, Taipei, Taiwan, November 20-24, 2016, Revised selected papers, Part V 13, p 214–230. Springer

Chen Z, Chen K, Lin W, See J, Yu H, Ke Y, Yang C (2020) Piou loss: towards accurate oriented object detection in complex environments. In: Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16, pp. 195–211. Springer

Chen Z, Yang C, Li Q, Zhao F, Zha Z-J, Wu F (2021) Disentangle your dense object detector. In: Proceedings of the 29th ACM international conference on multimedia. MM '21, p 4939–4948. Association for computing machinery, New York, NY, USA. https://doi.org/10.1145/3474085.3475351

Chen X, Jiang J, Li Z, Qi H, Li Q, Liu J, Zheng L, Liu M, Deng Y (2023a) An online continual object detector on vhr remote sensing images with class imbalance. Eng Appl Artif Intell 117:105549. https://doi.org/10.1016/j.engappai.2022.105549

Chen Z, Liang Y, Yu Z, Xu K, Ji Q, Zhang X, Zhang Q, Cui Z, He Z, Chang R et al (2023b) To-yolox: a pure cnn tiny object detection model for remote sensing images. Int J Digital Earth 16(1):3882–3904

Chen Y, Lin M, He Z, Polat K, Alhudhaif A, Alenezi F (2023c) Consistency-and dependence-guided knowledge distillation for object detection in remote sensing images. Expert Syst Appl 229:120519

Chen S, Zhao J, Zhou Y, Wang H, Yao R, Zhang L, Xue Y (2023d) Info-fpn: an informative feature pyramid network for object detection in remote sensing images. Expert Syst Appl 214:119132. https://doi.org/10.1016/j.eswa.2022.119132

Chen X, Wang C, Li Z, Liu M, Li Q, Qi H, Ma D, Li Z, Wang Y (2023e) Coupled global-local object detection for large vhr aerial images. Knowl-Based Syst 260:110097. https://doi.org/10.1016/j.knosys.2022.110097

Chen L, Liu C, Li W, Xu Q, Deng H (2024a) Dtssnet: dynamic training sample selection network for uav object detection. IEEE transactions on geoscience and remote sensing

Chen D, Miao D, Zhao X (2024b) Hyneter: hybrid network transformer for multiple computer vision tasks. IEEE transactions on industrial informatics

Chen P, Wang J, Zhang Z, He C (2024c) Dila: dynamic gaussian distribution fitting and imitation learning-based label assignment for tiny object detection. Appl Soft Comput, p 111980

Cheng G, Wang J, Li K, Xie X, Lang C, Yao Y, Han J (2022) Anchor-free oriented proposal generator for object detection. IEEE Trans Geosci Remote Sens 60:1–11. https://doi.org/10.1109/TGRS.2022.3183022

Cheng G, Yuan X, Yao X, Yan K, Zeng Q, Xie X, Han J (2023a) Towards large-scale small object detection: survey and benchmarks. IEEE transactions on pattern analysis and machine intelligence

Cheng G, Li Q, Wang G, Xie X, Min L, Han J (2023b) Sfrnet: fine-grained oriented object recognition via separate feature refinement. IEEE Trans Geosci Remote Sens 61:1–10

Cheng J, Yao X, Yang X, Yuan X, Feng X, Cheng G, Huang X, Han J (2024) Dima: Digging into multigranular archetype for fine-grained object detection. IEEE transactions on geoscience and remote sensing

Cong R, Zhang Y, Fang L, Li J, Zhao Y, Kwong S (2022) Rrnet: relational reasoning network with parallel multiscale attention for salient object detection in optical remote sensing images. IEEE Trans Geosci Remote Sens 60:1–11. https://doi.org/10.1109/TGRS.2021.3123984

Dai L, Liu H, Tang H, Wu Z, Song P Ao2-detr: arbitrary-oriented object detection transformer. IEEE transactions on circuits and systems for video technology, p 1–1 https://doi.org/10.1109/TCSVT.2022.3222906

Dai P, Yao S, Li Z, Zhang S, Cao X (2022a) Ace: anchor-free corner evolution for real-time arbitrarily-oriented object detection. IEEE Trans Image Process 31:4076–4089. https://doi.org/10.1109/TIP.2022.3167919

Dai Y, Yu J, Zhang D, Hu T, Zheng X (2022b) Rodformer: high-precision design for rotating object detection with transformers. Sensors 22(7):2633. https://doi.org/10.3390/s22072633

Deng S, Li S, Xie K, Song W, Liao X, Hao A, Qin H (2020) A global-local self-adaptive network for drone-view object detection. IEEE Trans Image Process 30:1556–1569

Deng S, Li S, Xie K, Song W, Liao X, Hao A, Qin H (2021) A global-local self-adaptive network for drone-view object detection. IEEE Trans Image Proc 30:1556–1569. https://doi.org/10.1109/TIP.2020.3045636

Deng C, Jing D, Han Y, Chanussot J (2023) Towards hierarchical adaptive alignment for aerial object detection in remote sensing images. IEEE transactions on geoscience and remote sensing

Ding J, Xue N, Xia G-S, Bai X, Yang W, Yang MY, Belongie S, Luo J, Datcu M, Pelillo M, Zhang L (2022) Object detection in aerial images: a large-scale benchmark and challenges. IEEE Trans Pattern Anal Mach Intell 44(11):7778–7796. https://doi.org/10.1109/TPAMI.2021.3117983

Doloriel CTC, Cajote RD (2023) Improving the detection of small oriented objects in aerial images. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, p 176–185

Dong X, Wang L, Sun X, Jia X, Gao L, Zhang B (2021) Remote sensing image super-resolution using second-order multi-scale networks. IEEE Trans Geosci Remote Sens 59(4):3473–3485. https://doi.org/10.1109/TGRS.2020.3019660

Dong Z, Wang M, Wang Y, Liu Y, Feng Y, Xu W (2022a) Multi-oriented object detection in high-resolution remote sensing imagery based on convolutional neural networks with adaptive object orientation features. Remote Sens 14(4):950. https://doi.org/10.3390/rs14040950

Dong X, Qin Y, Gao Y, Fu R, Liu S, Ye Y (2022b) Attention-based multi-level feature fusion for object detection in remote sensing images. Remote Sens 14(15):3735. https://doi.org/10.3390/rs14153735

Dong X, Fu R, Gao Y, Qin Y, Ye Y, Li B (2022c) Remote sensing object detection based on receptive field expansion block. IEEE Geosci Remote Sens Lett 19:1–5. https://doi.org/10.1109/LGRS.2021.3110584
Dong X, Qin Y, Fu R, Gao Y, Liu S, Ye Y (2022d) Remote sensing object detection based on gated context-aware module. IEEE Geosci Remote Sens Lett 19:1–5. https://doi.org/10.1109/LGRS.2022.3223069
Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S et al (2020) An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929
Du D, Qi Y, Yu H, Yang Y, Duan K, Li G, Zhang W, Huang Q, Tian Q (2018) The unmanned aerial vehicle benchmark: object detection and tracking. In: proceedings of the european conference on computer vision (ECCV)
Du B, Huang Y, Chen J, Huang D (2023) Adaptive sparse convolutional networks with global context enhancement for faster object detection on drone images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, p 13435–13444
Duan K, Bai S, Xie L, Qi H, Huang Q, Tian Q (2019a) Centernet: keypoint triplets for object detection. In: 2019 IEEE/CVF international conference on computer vision, ICCV 2019, Seoul, Korea (South), October 27–November 2, 2019, p 6568–6577
Duan C, Wei Z, Zhang C, Qu S, Wang H (2021) Coarse-grained density map guided object detection in aerial images. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV) workshops, p 2789–2798
Er MJ, Zhang Y, Chen J, Gao W (2023) Ship detection with deep learning: a survey. Artif Intell Rev 56(10):11825–11865
Everingham M, Van Gool L, Williams CK, Winn J, Zisserman A (2010) The pascal visual object classes (voc) challenge. Int J Comput Vision 88:303–338
Fang F, Liang W, Cheng Y, Xu Q, Lim J-H (2023) Enhancing representation learning with spatial transformation and early convolution for reinforcement learning-based small object detection. IEEE Trans Circuits Syst Video Technol 34(1):315–328
Feng J, Xiao X (2022) Multiobject tracking of wildlife in videos using few-shot learning. Animals 12(9):1223
Feng C, Zhong Y, Gao Y, Scott MR, Huang W (2021) TOOD: task-aligned one-stage object detection. In: 2021 IEEE/CVF international conference on computer vision, ICCV 2021, Montreal, QC, Canada, October 10–17, 2021, p 3490–3499. https://doi.org/10.1109/ICCV48922.2021.00349
Feng X, Yao X, Cheng G, Han J (2022) Weakly supervised rotation-invariant aerial object detection network. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), p 14146–14155
Feng C, Chen Z, Kou R, Gao G, Wang C, Li X, Shu X, Dai Y, Fu Q, Yang J (2024) HazyDet: open-source benchmark for drone-view object detection with depth-cues in Hazy scenes. arXiv:2409.19833
Fu C-Y, Liu W, Ranga A, Tyagi A, Berg AC (2017) Dssd: deconvolutional single shot detector. arXiv preprint arXiv:1701.06659
Fu K, Chang Z, Zhang Y, Xu G, Zhang K, Sun X (2020) Rotation-aware and multi-scale convolutional neural network for object detection in remote sensing images. ISPRS J Photogramm Remote Sens 161:294–308. https://doi.org/10.1016/j.isprsjprs.2020.01.025
Fu C, Lu K, Zheng G, Ye J, Cao Z, Li B, Lu G (2023) Siamese object tracking for unmanned aerial vehicle: a review and comprehensive analysis. Artif Intell Rev 56(Suppl 1):1417–1477
Gao G, Liu Q, Wang Y (2021) Counting from sky: a large-scale data set for remote sensing object counting and a benchmark method. IEEE Trans Geosci Remote Sens 59(5):3642–3655. https://doi.org/10.1109/TGRS.2020.3020555
Gao T, Li Z, Wen Y, Chen T, Niu Q, Liu Z (2023a) Attention-free global multiscale fusion network for remote sensing object detection. IEEE transactions on geoscience and remote sensing
Gao G, Dai Y, Zhang X, Duan D, Guo F (2023b) Adcg: a cross-modality domain transfer learning method for synthetic aperture radar in ship automatic target recognition. IEEE transactions on geoscience and remote sensing
Gao T, Niu Q, Zhang J, Chen T, Mei S, Jubair A (2023c) Global to local: a scale-aware network for remote sensing object detection. IEEE transactions on geoscience and remote sensing
Ge Z, Liu S, Wang F, Li Z, Sun J (2021) YOLOX: exceeding YOLO series in 2021. arXiv:2107.08430
Ge Z, Qi L, Wang Y, Sun Y (2022) Zoom-and-reasoning: joint foreground zoom and visual-semantic reasoning detection network for aerial images. IEEE Signal Process Lett 29:2572–2576. https://doi.org/10.1109/LSP.2022.3229638
Ge F, Zhang Y, Wang L, Liu W, Liu Y, Coleman S, Kerr D (2024a) Multi-level feedback joint representation learning network based on adaptive area elimination for cross-view geo-localization. IEEE transactions on geoscience and remote sensing

Ge L, Wang G, Zhang T, Zhuang Y, Chen H, Dong H, Chen L (2024b) Regression-guided refocusing learning with feature alignment for remote sensing tiny object detection. IEEE transactions on geoscience and remote sensing
Girshick R (2015) Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision, p 1440–1448
Gu L, Fang Q, Wang Z, Popov E, Dong G (2023) Learning lightweight and superior detectors with feature distillation for onboard remote sensing object detection. Remote Sens 15(2):370
Guo X (2023) A novel multi to single module for small object detection. arXiv preprint arXiv:2303.14977
Han J, Ding J, Li J, Xia G-S (2021) Align deep features for oriented object detection. IEEE Trans Geosci Remote Sens 60:1–11
Han W, Kuerban A, Yang Y, Huang Z, Liu B, Gao J (2022) Multi-vision network for accurate and real-time small object detection in optical remote sensing images. IEEE Geosci Remote Sens Lett 19:1–5. https://doi.org/10.1109/LGRS.2020.3044422
He K, Zhang X, Ren S, Sun J (2015) Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans Pattern Anal Mach Intell 37(9):1904–1916
He K, Gkioxari G, Dollar P, Girshick R (2017) Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision (ICCV)
He Q, Sun X, Yan Z, Wang B, Zhu Z, Diao W, Yang MY (2023) Ast: adaptive self-supervised transformer for optical remote sensing representation. ISPRS J Photogramm Remote Sens 200:41–54
Hong M, Li S, Yang Y, Zhu F, Zhao Q, Lu L (2022) Sspnet: scale selection pyramid network for tiny person detection from uav images. IEEE Geosci Remote Sens Lett 19:1–5. https://doi.org/10.1109/LGRS.2021.3103069
Hu J, Zhi X, Jiang S, Tang H, Zhang W, Bruzzone L (2022) Supervised multi-scale attention-guided ship detection in optical remote sensing images. IEEE Trans Geosci Remote Sens 60:1–14. https://doi.org/10.1109/TGRS.2022.3206306
Hu Z, Gao K, Zhang X, Wang J, Wang H, Yang Z, Li C, Li W (2023) Emo2-detr: efficient-matching oriented object detection with transformers. IEEE transactions on geoscience and remote sensing
Hua W, Liang D, Li J, Liu X, Zou Z, Ye X, Bai X (2023) Sood: towards semi-supervised oriented object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, p 15558–15567
Huang W, Li G, Chen Q, Ju M, Qu J (2021) Cf2pn: a cross-scale feature fusion pyramid network based remote sensing target detection. Remote Sens 13(5):847. https://doi.org/10.3390/rs13050847
Huang Y, Chen J, Huang D (2022) Ufpmp-det: toward accurate and efficient object detection on drone imagery. Proc AAAI Conf Artif Intell 36:1026–1033
Huang Z, Li W, Xia X-G, Wang H, Tao R (2024) Task-wise sampling convolutions for arbitrary-oriented object detection in aerial images. IEEE transactions on neural networks and learning systems
Hui Y, Wang J, Li B (2024) Dsaa-yolo: Uav remote sensing small target recognition algorithm for yolov7 based on dense residual super-resolution and anchor frame adaptive regression strategy. J King Saud Univ-Comput Inform Sci 36(1):101863
Jain A, Ramaprasad R, Narang P, Mandal M, Chamola V, Yu FR, Guizan M (2021) Ai-enabled object detection in uavs: challenges, design choices, and research directions. IEEE Network 35(4):129–135
Ji F, Ming D, Zeng B, Yu J, Qing Y, Du T, Zhang X (2021) Aircraft detection in high spatial resolution remote sensing images combining multi-angle features driven and majority voting cnn. Remote Sens 13(11):2207. https://doi.org/10.3390/rs13112207
Jocher G, Stoken A, Borovec J, Chaurasia A, Changyu L, Hogan A, Hajek J, Diaconu L, Kwon Y, Defretin Y et al (2021) ultralytics/yolov5: v5. 0-yolov5-p6 1280 models, aws, supervise. ly and youtube integrations. Zenodo
Jocher G, Chaurasia A, Qiu J (2023) Ultralytics YOLOv8. https://github.com/ultralytics/ultralytics
Khosravi MR, Rezaee K, Moghimi MK, Wan S, Menon VG (2023) Crowd emotion prediction for human-vehicle interaction through modified transfer learning and fuzzy logic ranking. IEEE transactions on intelligent transportation systems
Kortylewski A, Liu Q, Wang H, Zhang Z, Yuille A (2019) Localizing occluders with compositional convolutional networks. In: Proceedings of the IEEE/CVF international conference on computer vision workshops, p 0–0
Koyun OC, Keser RK (2022) Akkaya: focus-and-detect: a small object detection framework for aerial images. Signal Proc: Image Commun 104:116675. https://doi.org/10.1016/j.image.2022.116675
Law H, Deng J (2018) Cornernet: detecting objects as paired keypoints. In: Proceedings of the European conference on computer vision (ECCV)
Leng J, Mo M, Zhou Y, Gao C, Li W, Gao X (2023) Pareto refocusing for drone-view object detection. IEEE Trans Circuits Syst Video Technol 33(3):1320–1334. https://doi.org/10.1109/TCSVT.2022.3210207
Li Z, Zhou F (2017) Fssd: feature fusion single shot multibox detector. arXiv preprint arXiv:1712.00960

Li Z, Peng C, Yu G, Zhang X, Deng Y, Sun J (2017) Light-head r-cnn: In defense of two-stage object detector. arXiv preprint arXiv:1711.07264
Li Z, Peng C, Yu G, Zhang X, Deng Y, Sun J (2018) Detnet: a backbone network for object detection. arXiv preprint arXiv:1804.06215
Li J, Jing M, Lu K, Zhu L, Yang Y, Huang Z (2019) Alleviating feature confusion for generative zero-shot learning. In: Proceedings of the 27th ACM international conference on multimedia, p 1587–1595
Li C, Yang T, Zhu S, Chen C, Guan S (2020a) Density map guided object detection in aerial images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR) workshops
Li K, Wan G, Cheng G, Meng L, Han J (2020b) Object detection in optical remote sensing images: a survey and a new benchmark. ISPRS J Photogramm Remote Sens 159:296–307. https://doi.org/10.1016/j.isprsjprs.2019.11.023
Li W, Wei W, Zhang L (2021a) Gsdet: object detection in aerial images based on scale reasoning. IEEE Trans Image Process 30:4599–4609. https://doi.org/10.1109/TIP.2021.3073319
Li X, Wang W, Hu X, Li J, Tang J, Yang J (2021b) Generalized focal loss v2: learning reliable localization quality estimation for dense object detection. In: 2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR), p 11627–11636
Li Y, Kong C, Dai L, Chen X (2022) Single-stage detector with dual feature alignment for remote sensing object detection. IEEE Geosci Remote Sens Lett 19:1–5. https://doi.org/10.1109/LGRS.2021.3130379
Li J, Zhang H, Song R, Xie W, Li Y, Du Q (2022a) Structure-guided feature transform hybrid residual network for remote sensing object detection. IEEE Trans Geosci Remote Sens 60:1–13. https://doi.org/10.1109/TGRS.2021.3103964
Li Q, Chen Y, Zeng Y (2022b) Transformer with transfer cnn for remote-sensing-image object detection. Remote Sens 14(4):984. https://doi.org/10.3390/rs14040984
Li X, Diao W, Mao Y, Gao P, Mao X, Li X, Sun X (2023a) Ogmn: occlusion-guided multi-task network for object detection in uav images. ISPRS J Photogramm Remote Sens 199:242–257
Li Z, Hou B, Wu Z, Ren B, Ren Z, Jiao L (2023b) Gaussian synthesis for high-precision location in oriented object detection. IEEE transactions on geoscience and remote sensing
Li Y, Hou Q, Zheng Z, Cheng M, Yang J, Li X (2023c) Large selective kernel network for remote sensing object detection. arxiv 2023. arXiv preprint arXiv:2303.09030
Li W, Zhao D, Yuan B, Gao Y, Shi Z (2023d) Petdet: proposal enhancement for two-stage fine-grained object detection. IEEE transactions on geoscience and remote sensing
Li C, Cheng G, Wang G, Zhou P, Han J (2023e) Instance-aware distillation for efficient object detection in remote sensing images. IEEE Trans Geosci Remote Sens 61:1–11
Li Z, Tang C, Liu X, Zhang W, Dou J, Wang L, Zomaya AY (2023f) Lightweight remote sensing change detection with progressive feature aggregation and supervised attention. IEEE Trans Geosci Remote Sens 61:1–12
Li J, Tian P, Song R, Xu H, Li Y, Du Q (2024a) Pcvit: a pyramid convolutional vision transformer detector for object detection in remote sensing imagery. IEEE transactions on geoscience and remote sensing
Li Y, Luo J, Zhang Y, Tan Y, Yu J-G, Bai S (2024b) Learning to holistically detect bridges from large-size vhr remote sensing imagery. IEEE transactions on pattern analysis and machine intelligence
Liang D, Zhang J-W, Tang Y-P, Huang S-J (2023) Mus-cdb: mixed uncertainty sampling with class distribution balancing for active annotation in aerial object detection. IEEE Trans Geosci Remote Sens 61:1–13
Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: common objects in context. In: Computer vision–ECCV 2014: 13th European conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, p 740–755. Springer
Lin T-Y, Dollár P, Girshick R, He K, Hariharan B, Belongie S (2017a) Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, p 2117–2125
Lin T-Y, Goyal P, Girshick R, He K, Dollár P (2017b) Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision, p 2980–2988
Lin H, Zhou J, Gan Y, Vong C-M, Liu Q (2020) Novel up-scale feature aggregation for object detection in aerial images. Neurocomputing 411:364–374. https://doi.org/10.1016/j.neucom.2020.06.011
Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y, Berg AC (2016) Ssd: single shot multibox detector. In: Computer vision–ECCV 2016: 14th European conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, p 21–37. Springer
Liu B-Y, Chen H-X, Huang Z, Liu X, Yang Y-Z (2021a) Zoominnet: a novel small object detector in drone images with cross-scale knowledge distillation. Remote Sens 13(6):1198. https://doi.org/10.3390/rs13061198
Liu Z, Gao G, Sun L, Fang Z (2021b) Hrdnet: high-resolution detection network for small objects. In: 2021 IEEE international conference on multimedia and expo (ICME), p 1–6. IEEE

Liu E, Zheng Y, Pan B, Xu X, Shi Z (2021c) Dcl-net: augmenting the capability of classification and localization for remote sensing object detection. IEEE Trans Geosci Remote Sens 59(9):7933–7944. https://doi.org/10.1109/TGRS.2020.3048384
Liu S, Zhang L, Lu H, He Y (2022a) Center-boundary dual attention for oriented object detection in remote sensing images. IEEE Trans Geosci Remote Sens 60:1–14. https://doi.org/10.1109/TGRS.2021.3069056
Liu F, Chen R, Zhang J, Xing K, Liu H, Qin J (2022b) R2yolox: a lightweight refined anchor-free rotated detector for object detection in aerial images. IEEE Trans Geosci Remote Sens 60:1–15. https://doi.org/10.1109/TGRS.2022.3215472
Liu Y, Li Q, Yuan Y, Du Q, Wang Q (2022c) Abnet: adaptive balanced network for multiscale object detection in remote sensing imagery. IEEE Trans Geosci Remote Sens 60:1–14. https://doi.org/10.1109/TGRS.2021.3133956
Liu J, Li S, Zhou C, Cao X, Gao Y, Wang B (2022d) Sraf-net: a scene-relevant anchor-free object detection network in remote sensing images. IEEE Trans Geosci Remote Sens 60:1–14. https://doi.org/10.1109/TGRS.2021.3124959
Liu X, Ma S, He L, Wang C, Chen Z (2022e) Hybrid network model: transconvnet for oriented object detection in remote sensing images. Remote Sens 14(9):2090. https://doi.org/10.3390/rs14092090
Liu K, Huang J, Li X (2022f) Eagle-eye-inspired attention for object detection in remote sensing. Remote Sens 14(7):1743. https://doi.org/10.3390/rs14071743
Liu S, Li F, Zhang H, Yang X, Qi X, Su H, Zhu J, Zhang L (2022g) DAB-DETR: dynamic anchor boxes are better queries for DETR. In: International conference on learning representations (ICLR)
Liu N, Li W, Sun X, Tao R, Chanussot J (2023) Remote sensing image fusion with task-inspired multiscale nonlocal-attention network. IEEE Geosci Remote Sens Lett 20:1–5. https://doi.org/10.1109/LGRS.2023.3254049
Liu J, Cui J, Ye M, Zhu X, Tang S (2024a) Shooting condition insensitive unmanned aerial vehicle object detection. Expert Syst Appl 246:123221
Liu N, Xu X, Gao Y, Zhao Y, Li H-C (2024b) Semi-supervised object detection with uncurated unlabeled data for remote sensing images. Int J Appl Earth Obs Geoinf 129:103814
Liu Y, Yan G, Ma F, Zhou Y, Zhang F (2024c) Sar ship detection based on explainable evidence learning under intra-class imbalance. IEEE transactions on geoscience and remote sensing
Liu H-I, Tseng Y-W, Chang K-C, Wang P-J, Shuai H-H, Cheng W-H (2024d) A denoising fpn with transformer r-cnn for tiny object detection. IEEE transactions on geoscience and remote sensing
Liu C, Gao G, Huang Z, Hu Z, Liu Q, Wang Y (2024e) Yolc: you only look clusters for tiny object detection in aerial images. IEEE transactions on intelligent transportation systems
Lu X, Li B, Yue Y, Li Q, Yan J (2019) Grid r-cnn. In: 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR), p 7355–7364. https://doi.org/10.1109/CVPR.2019.00754
Luo J, Yang X, Yu Y, Li Q, Yan J, Li Y (2024) Pointobb: learning oriented object detection via single point supervision. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, p 16730–16740
Lv H, Qian W, Chen T, Yang H, Zhou X (2022) Multiscale feature adaptive fusion for object detection in optical remote sensing images. IEEE Geosci Remote Sens Lett 19:1–5. https://doi.org/10.1109/LGRS.2022.3178787
Ma T, Mao M, Zheng H, Gao P, Wang X, Han S, Ding E, Zhang B, Doermann D (2021) Oriented object detection with transformer. arXiv preprint arXiv:2106.03146
Ma W, Li N, Zhu H, Jiao L, Tang X, Guo Y, Hou B (2022) Feature split–merge–enhancement network for remote sensing object detection. IEEE Trans Geosci Remote Sens 60:1–17. https://doi.org/10.1109/TGRS.2022.3140856
Ma Y, Chai L, Jin L (2023) Scale decoupled pyramid for object detection in aerial images. IEEE transactions on geoscience and remote sensing
Ma Y, Chai L, Jin L, Yan J (2024a) Hierarchical alignment network for domain adaptive object detection in aerial images. ISPRS J Photogramm Remote Sens 208:39–52
Ma W, Wang X, Zhu H, Yang X, Yi X, Jiao L (2024b) Significant feature elimination and sample assessment for remote sensing small objects' detection. IEEE Trans Geosci Remote Sens 62:1–15
Meethal A, Granger E, Pedersoli M (2023) Cascaded zoom-in detector for high resolution aerial images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, p 2046–2055
Mei S, Jiang R, Ma M, Song C (2023) Rotation-invariant feature learning via convolutional neural network with cyclic polar coordinates convolutional layer. IEEE Trans Geosci Remote Sens 61:1–13
Meng D, Chen X, Fan Z, Zeng G, Li H, Yuan Y, Sun L, Wang J (2021) Conditional DETR for fast training convergence. In: 2019 IEEE/CVF international conference on computer vision, ICCV 2019, Seoul, Korea (South), October 27–November 2, 2019, p 3651–3660

Ming Q, Miao L, Zhou Z, Dong Y (2022) Cfc-net: a critical feature capturing network for arbitrary-oriented object detection in remote-sensing images. IEEE Trans Geosci Remote Sens 60:1–14. https://doi.org/10.1109/TGRS.2021.3095186
Ming Q, Miao L, Zhou Z, Song J, Dong Y, Yang X (2023) Task interleaving and orientation estimation for high-precision oriented object detection in aerial images. ISPRS J Photogramm Remote Sens 196:241–255
Ming Q, Miao L, Zhou Z, Song J, Pižurica A (2024) Gradient calibration loss for fast and accurate oriented bounding box regression. IEEE transactions on geoscience and remote sensing
Mokayed H, Nayebiastaneh A, De K, Sozos S, Hagner O, Backe B (2023) Nordic vehicle dataset (nvd): performance of vehicle detectors using newly captured nvd from uav in different snowy weather conditions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, p 5314–5322
Moon J, Jeon M, Jeong S, Oh K-Y (2024) Romp-transformer: rotational bounding box with multi-level feature pyramid transformer for object detection. Pattern Recogn 147:110067
Ouyang L, Guo G, Fang L, Ghamisi P, Yue J (2023) Pcldet: prototypical contrastive learning for fine-grained object detection in remote sensing images. IEEE Trans Geosci Remote Sens 61:1–11
Pang J, Chen K, Shi J, Feng H, Ouyang W, Lin D (2019) Libra R-CNN: towards balanced learning for object detection. In: IEEE conference on computer vision and pattern recognition, CVPR 2019, Long Beach, CA, USA, June 16–20, 2019, p 821–830
Pang Y, Zhang Y, Kong Q, Wang Y, Chen B, Cao X (2023) Socdet: a lightweight and accurate oriented object detection network for satellite on-orbit computing. IEEE Trans Geosci Remote Sens 61:1–15
Peng G, Yang Z, Wang S, Zhou Y (2023) Amflw-yolo: a lightweight network for remote sensing image detection based on attention mechanism and multi-scale feature fusion. IEEE transactions on geoscience and remote sensing
Pu Y, Wang Y, Xia Z, Han Y, Wang Y, Gan W, Wang Z, Song S, Huang G (2023) Adaptive rotated convolution for rotated object detection. In: Proceedings of the IEEE/CVF international conference on computer vision, p 6589–6600
Qin R, Liu Q, Gao G, Huang D, Wang Y (2022) Mrdet: a multihead network for accurate rotated object detection in aerial images. IEEE Trans Geosci Remote Sens 60:1–12. https://doi.org/10.1109/TGRS.2021.3113473
Qian X, Zhang N, Wang W (2023a) Smooth giou loss for oriented object detection in remote sensing images. Remote Sens 15(5):1259. https://doi.org/10.3390/rs15051259
Qian X, Wu B, Cheng G, Yao X, Wang W, Han J (2023b) Building a bridge of bounding box regression between oriented and horizontal object detection in remote sensing images. IEEE Trans Geosci Remote Sens 61:1–9. https://doi.org/10.1109/TGRS.2023.3256373
Rao C, Wang J, Cheng G, Xie X, Han J (2023) Learning orientation-aware distances for oriented object detection. IEEE Trans Geosci Remote Sens 61:1–11
Redmon J, Farhadi A (2018) YOLOv3: an incremental improvement
Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, p 779–788
Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: towards real-time object detection with region proposal networks. Advances in neural information processing systems 28
Ren S, He K, Girshick R, Sun J (2017) Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 39(6):1137–1149
Ren X, Sun M, Zhang X, Liu L, Zhou H, Ren X (2022) An improved mask-rcnn algorithm for uav tir video stream target detection. Int J Appl Earth Obs Geoinf 106:102660
Rezatofighi H, Tsoi N, Gwak J, Sadeghian A, Reid I, Savarese S (2019) Generalized intersection over union: a metric and a loss for bounding box regression. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, p 658–666
Roh B, Shin J, Shin W, Kim S (2022) Sparse detr: efficient end-to-end object detection with learnable sparsity. In: ICLR
Roy AM, Bhaduri J (2022) Real-time growth stage detection model for high degree of occultation using densenet-fused yolov4. Comput Electron Agric 193:106694
Sairam RVC, Keswani M, Sinha U, Shah N, Balasubramanian VN (2023) Aruba: an architecture-agnostic balanced loss for aerial object detection. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision (WACV), p 3719–3728
Shen Y, Zhang F, Liu D, Pu W, Zhang Q (2022) Manhattan-distance iou loss for fast and accurate bounding box regression and object detection. Neurocomputing 500:99–114. https://doi.org/10.1016/j.neucom.2022.05.052
Shen J, Zhang C, Yuan Y, Wang Q (2023a) Enhancing prospective consistency for semi-supervised object detection in remote sensing images. IEEE transactions on geoscience and remote sensing

Shen Y, Zhang D, Song Z, Jiang X, Ye Q (2023b) Learning to reduce information bottleneck for object detection in aerial images. IEEE Geosci Remote Sens Lett 20:1–5. https://doi.org/10.1109/LGRS.2023.3264455
Shen C, Qian J, Wang C, Yan D, Zhong C (2024) Dynamic sensing and correlation loss detector for small object detection in remote sensing images. IEEE transactions on geoscience and remote sensing
Shi T, Gong J, Hu J, Zhi X, Zhang W, Zhang Y, Zhang P, Bao G (2022) Feature-enhanced centernet for small object detection in remote sensing images. Remote Sens 14(21):5488. https://doi.org/10.3390/rs14215488
Stateczny A, Uday Kiran G, Bindu G, Ravi Chythanya K, Ayyappa Swamy K (2022) Spiral search grasshopper features selection with vgg19-resnet50 for remote sensing object detection. Remote Sens 14(21):5398. https://doi.org/10.3390/rs14215398
Sun X, Liu Y, Yan Z, Wang P, Diao W, Fu K (2021) Sraf-net: shape robust anchor-free network for garbage dumps in remote sensing imagery. IEEE Trans Geosci Remote Sens 59(7):6154–6168. https://doi.org/10.1109/TGRS.2020.3023928
Sun X, Wang P, Yan Z, Xu F, Wang R, Diao W, Chen J, Li J, Feng Y, Xu T, Weinmann M, Hinz S, Wang C, Fu K (2022) Fair1m: a benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery. ISPRS J Photogramm Remote Sens 184:116–130. https://doi.org/10.1016/j.isprsjprs.2021.12.004
Sun P, Zhang R, Jiang Y, Kong T, Xu C, Zhan W, Tomizuka M, Yuan Z, Luo P (2023) Sparse R-CNN: an end-to-end framework for object detection. IEEE Trans Pattern Anal Mach Intell 45(12):15650–15664. https://doi.org/10.1109/TPAMI.2023.3292030
Tan Z, Jiang Z, Guo C, Zhang H (2023) Wsodet: a weakly supervised oriented detector for aerial object detection. IEEE Trans Geosci Remote Sens 61:1–12
Tang S, Zhang S, Fang Y (2024) Hic-yolov5: improved yolov5 for small object detection. In: 2024 IEEE international conference on robotics and automation (ICRA), p 6614–6619. IEEE
Teng Z, Duan Y, Liu Y, Zhang B, Fan J (2022) Global to local: clip-lstm-based object detection from remote sensing images. IEEE Trans Geosci Remote Sens 60:1–13
Tian Z, Shen C, Chen H, He T (2019) FCOS: fully convolutional one-stage object detection. In: 2019 IEEE/CVF international conference on computer vision, ICCV 2019, Seoul, Korea (South), October 27–November 2, 2019, p 9626–9635
Tian G, Liu J, Yang W (2021) A dual neural network for object detection in uav images. Neurocomputing 443:292–301. https://doi.org/10.1016/j.neucom.2021.03.016
Tian S, Cao L, Kang L, Xing X, Tian J, Du K, Sun K, Fan C, Fu Y, Zhang Y (2022) A novel hybrid attention-driven multistream hierarchical graph embedding network for remote sensing object detection. Remote Sens 14(19):4951. https://doi.org/10.3390/rs14194951
Wan J, Zhang B, Zhao Y, Du Y, Tong Z (2021) Vistrongerdet: Stronger visual information for object detection in visdrone images. In: Proceedings of the IEEE/CVF international conference on computer vision, p 2820–2829
Wang J, Guo W, Pan T, Yu H, Duan L, Yang W (2018) Bottle detection in the wild using low-altitude unmanned aerial vehicles. In: 2018 21st International conference on information fusion (FUSION), p 439–444. IEEE
Wang Y, Yang Y, Zhao X (2020) Object detection using clustering algorithm adaptive searching regions in aerial images. In: Computer vision–ECCV 2020 workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16, p 651–664. Springer
Wang J, Xu C, Yang W, Yu L (2021) A normalized Gaussian Wasserstein distance for tiny object detection. arXiv preprint arXiv:2110.13389
Wang J, Shao F, He X, Lu G (2022a) A novel method of small object detection in uav remote sensing images based on feature alignment of candidate regions. Drones 6(10):292. https://doi.org/10.3390/drones6100292
Wang H, Huang Z, Chen Z, Song Y, Li W (2022b) Multigrained angle representation for remote-sensing object detection. IEEE Trans Geosci Remote Sens 60:1–13. https://doi.org/10.1109/TGRS.2022.3212592
Wang J, Wang Y, Wu Y, Zhang K, Wang Q (2022c) Frpnet: a feature-reflowing pyramid network for object detection of remote sensing images. IEEE Geosci Remote Sens Lett 19:1–5. https://doi.org/10.1109/LGRS.2020.3040308
Wang L, Mu X, Ma C, Zhang J (2022d) Hausdorff iou and context maximum selection nms: improving object detection in remote sensing images with a novel metric and postprocessing module. IEEE Geosci Remote Sens Lett 19:1–5. https://doi.org/10.1109/LGRS.2021.3093577
Wang P, Niu Y, Wang J, Ma F, Zhang C (2022e) Arbitrarily oriented dense object detection based on center point network in remote sensing images. Remote Sens 14(7):1536. https://doi.org/10.3390/rs14071536
Wang J, Li F, Bi H (2022f) Gaussian focal loss: learning distribution polarized angle prediction for rotated object detection in aerial images. IEEE Trans Geosci Remote Sens 60:1–13. https://doi.org/10.1109/TGRS.2022.3175520

Wang S, Huang S, Liu S, Bi Y (2023a) Not just select samples, but exploration: genetic programming aided remote sensing target detection under deep learning. Appl Soft Comput 145:110570
Wang G, Zhang X, Peng Z, Jia X, Tang X, Jiao L (2023b) Mol: towards accurate weakly supervised remote sensing object detection via multi-view noisy learning. ISPRS J Photogramm Remote Sens 196:457–470
Wang Y, Zhang Z, Xu W, Chen L, Wang G, Yan L, Zhong S, Zou X (2023c) Learning oriented object detection via naive geometric computing. IEEE transactions on neural networks and learning systems
Wang D, Zhang Q, Xu Y, Zhang J, Du B, Tao D, Zhang L (2023d) Advancing plain vision transformer toward remote sensing foundation model. IEEE Trans Geosci Remote Sens 61:1–15. https://doi.org/10.1109/TGRS.2022.3222818
Wang L, Zhan Y, Liu W, Yu B, Tao D (2023e) Bounding box vectorization for oriented object detection with tanimoto coefficient regression. IEEE transactions on multimedia
Wang C-Y, Liao H-YM (2024) YOLOv9: learning what you want to learn using programmable gradient information. arXiv preprint arXiv:2402.13616
Wang C, Guo G, Liu C, Shao D, Gao S (2024a) Effective rotate: learning rotation-robust prototype for aerial object detection. IEEE transactions on geoscience and remote sensing
Wang Y, Zhang T, Zhao L, Hu L, Wang Z, Niu Z, Cheng P, Chen K, Zeng X, Wang Z et al (2024b) Ringmo-lite: a remote sensing lightweight network with cnn-transformer hybrid framework. IEEE transactions on geoscience and remote sensing
Wang A, Chen H, Liu L, Chen K, Lin Z, Han J, Ding G (2024c) Yolov10: real-time end-to-end object detection. arXiv preprint arXiv:2405.14458
Wei C, Ni W, Qin Y, Wu J, Zhang H, Liu Q, Cheng K, Bian H (2023) Ridop: a rotation-invariant detector with simple oriented proposals in remote sensing images. Remote Sens 15(3):594. https://doi.org/10.3390/rs15030594
Wu J, Xu S (2021) From point to region: accurate and efficient hierarchical small object detection in low-resolution remote sensing images. Remote Sens 13(13):2620. https://doi.org/10.3390/rs13132620
Wu J, Pan Z, Lei B, Hu Y (2022a) Fsanet: feature-and-spatial-aligned network for tiny object detection in remote sensing images. IEEE Trans Geosci Remote Sens 60:1–17. https://doi.org/10.1109/TGRS.2022.3205052
Wu Y, Zhang K, Wang J, Wang Y, Wang Q, Li Q (2022b) Cdd-net: a context-driven detection network for multiclass object detection. IEEE Geosci Remote Sens Lett 19:1–5. https://doi.org/10.1109/LGRS.2020.3042465
Wu W, Wong H-S, Wu S (2024a) Pseudo-siamese teacher for semi-supervised oriented object detection. IEEE transactions on geoscience and remote sensing
Wu W, Wong H-S, Wu S, Zhang T (2024b) Relational matching for weakly semi-supervised oriented object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, p 27800–27810
Xia G-S, Bai X, Ding J, Zhu Z, Belongie S, Luo J, Datcu M, Pelillo M, Zhang L (2018) Dota: A large-scale dataset for object detection in aerial images. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)
Xiao Z, Xu B, Zhang Y, Wang K, Wan Q, Tan X (2022a) Aspect ratio-based bidirectional label encoding for square-like rotation detection. IEEE geoscience and remote sensing letters 1–1. https://doi.org/10.1109/LGRS.2023.3247027
Xiao J, Guo H, Yao Y, Zhang S, Zhou J, Jiang Z (2022b) Multi-scale object detection with the pixel attention mechanism in a complex background. Remote Sens 14(16):3969. https://doi.org/10.3390/rs14163969
Xiao J, Yao Y, Zhou J, Guo H, Yu Q, Wang Y-F (2023) Fdlr-net: a feature decoupling and localization refinement network for object detection in remote sensing images. Expert Syst Appl 225:120068. https://doi.org/10.1016/j.eswa.2023.120068
Xiao Z, Yang G, Yang X, Mu T, Yan J, Hu S (2024) Theoretically achieving continuous representation of oriented bounding boxes. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, p 16912–16922
Xie X, Cheng G, Feng X, Yao X, Qian X, Han J (2023a) Attention erasing and instance sampling for weakly supervised object detection. IEEE transactions on geoscience and remote sensing
Xie Y, Hou X, Guo Y, Wang X, Zheng J (2023b) Joint-guided distillation binary neural network via dynamic channel-wise diversity enhancement for object detection. IEEE Trans Circuits Syst Video Technol 34(1):448–460
Xie X, Cheng G, Li Q, Miao S, Li K, Han J (2024a) Fewer is more: efficient object detection in large aerial images. Science China Inf Sci 67(1):112106
Xie X, Cheng G, Rao C, Lang C, Han J (2024b) Oriented object detection via contextual dependence mining and penalty-incentive allocation. IEEE transactions on geoscience and remote sensing

Xu X, Feng Z, Cao C, Li M, Wu J, Wu Z, Shang Y, Ye S (2021) An improved swin transformer-based model for remote sensing object detection and instance segmentation. Remote Sens 13(23):4779. https://doi.org/10.3390/rs13234779
Xu J, Li Y-L, Wang S (2022a) Adazoom: towards scale-aware large scene object detection. IEEE Trans Multimedia 25:4598–4609
Xu C, Wang J, Yang W, Yu H, Yu L, Xia G-S (2022b) Detecting tiny objects in aerial images: a normalized wasserstein distance and a new benchmark. ISPRS J Photogramm Remote Sens 190:79–93. https://doi.org/10.1016/j.isprsjprs.2022.06.002
Xu G, Song T, Sun X, Gao C (2023) Transmin: transformer-guided multi-interaction network for remote sensing object detection. IEEE Geosci Remote Sens Lett 20:1–5. https://doi.org/10.1109/LGRS.2022.3230973
Xu J, Fan X, Jian H, Xu C, Bei W, Ge Q, Zhao T (2024) Yoloow: a spatial scale adaptive real-time object detection neural network for open water search and rescue from uav aerial imagery. IEEE transactions on geoscience and remote sensing
Yamani A, Alyami A, Luqman H, Ghanem B, Giancola S (2024) Active learning for single-stage object detection in uav images. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, p 1860–1869
Yan P, Zhao J, Hou R, Duan X, Cai S, Wang X (2024) Clustered remote sensing target distribution detection aided by density-based spatial analysis. Int J Appl Earth Obs Geoinf 132:104019
Yang F, Fan H, Chu P, Blasch E, Ling H (2019a) Clustered object detection in aerial images. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV)
Yang F, Fan H, Chu P, Blasch E, Ling H (2019b) Clustered object detection in aerial images. In: Proceedings of the IEEE/CVF international conference on computer vision, p 8311–8320
Yang Z, Liu S, Hu H, Wang L, Lin S (2019c) Reppoints: point set representation for object detection. In: 2019 IEEE/CVF international conference on computer vision, ICCV 2019, Seoul, Korea (South), October 27–November 2, 2019, p 9656–9665
Yang Z, Kong J, Zheng B, Li M, Zhang WE, Chen T (2022a) Object detection in remote sensing images with balanced rotational and horizontal bounding boxes. IEEE Geosci Remote Sens Lett 19:1–5. https://doi.org/10.1109/LGRS.2022.3211325
Yang Y, Sun X, Diao W, Yin D, Yang Z, Li X (2022b) Statistical sample selection and multivariate knowledge mining for lightweight detectors in remote sensing imagery. IEEE Trans Geosci Remote Sens 60:1–14
Yang Y, Diao W, Rong X, Li X, Sun X (2022c) Dynamic interactive learning for lightweight detectors in remote sensing imagery. IEEE Trans Geosci Remote Sens 60:1–14
Yao Y, Cheng G, Wang G, Li S, Zhou P, Xie X, Han J (2023) On improving bounding box representations for oriented object detection. IEEE Trans Geosci Remote Sens 61:1–11. https://doi.org/10.1109/TGRS.2022.3231340
Yin N, Liu C, Tian R, Qian X (2024) Sdpdet: learning scale-separated dynamic proposals for end-to-end drone-view detection. IEEE transactions on multimedia
Ying Z, Zhou J, Zhai Y, Quan H, Li W, Genovese A, Piuri V, Scotti F (2024) Large-scale high-altitude uav-based vehicle detection via pyramid dual pooling attention path aggregation network. IEEE transactions on intelligent transportation systems
Yu D, Ji S (2022) A new spatial-oriented object detection framework for remote sensing images. IEEE Trans Geosci Remote Sens 60:1–16. https://doi.org/10.1109/TGRS.2021.3127232
Yu X, Gong Y, Jiang N, Ye Q, Han Z (2020) Scale match for tiny person detection. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision (WACV)
Yu W, Yang T, Chen C (2021) Towards resolving the challenge of long-tail distribution in uav images for object detection. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision (WACV), p 3258–3267
Yu D, Guo H, Zhao C, Liu X, Xu Q, Lin Y, Ding L (2023a) An anchor-free and angle-free detector for oriented object detection using bounding box projection. IEEE transactions on geoscience and remote sensing
Yu Y, Yang X, Li J, Gao X (2023b) Task-specific heterogeneous network for object detection in aerial images. IEEE transactions on geoscience and remote sensing
Yu Y, Zhang K, Wang X, Wang N, Gao X (2023c) An adaptive region proposal network with progressive attention propagation for tiny person detection from uav images. IEEE transactions on circuits and systems for video technology
Yu N, Ren H, Deng T, Fan X (2023d) Stepwise locating bidirectional pyramid network for object detection in remote sensing imagery. IEEE Geosci Remote Sens Lett 20:1–5. https://doi.org/10.1109/LGRS.2022.3223470
Yu Y, Yang X, Li Q, Da F, Dai J, Qiao Y, Yan J (2024a) Point2rbox: Combine knowledge from synthetic visual patterns for end-to-end oriented object detection with single point supervision. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, p 16783–16793

Yu X, Chen P, Wang K, Han X, Li G, Han Z, Ye Q, Jiao J (2024b) Cpr++: object localization via single coarse point supervision. IEEE transactions on pattern analysis and machine intelligence
Yu H, Tian Y, Ye Q, Liu Y (2024c) Spatial transform decoupling for oriented object detection. Proc AAAI Conf Artif Intell 38:6782–6790
Yuan Z, Liu Z, Zhu C, Qi J, Zhao D (2021) Object detection in remote sensing images via multi-feature pyramid network with receptive field block. Remote Sens 13(5):862. https://doi.org/10.3390/rs13050862
Yuan Y, Li Z, Ma D (2022) Feature-aligned single-stage rotation object detection with continuous boundary. IEEE Trans Geosci Remote Sens 60:1–11. https://doi.org/10.1109/TGRS.2022.3203983
Yue C, Yan J, Zhang Y, Luo Z, Liu Y, Guo P (2023) Scfnet: semantic correction and focus network for remote sensing image object detection. Expert Syst Appl 224:119980. https://doi.org/10.1016/j.eswa.2023.119980
Zeng Y, Chen Y, Yang X, Li Q, Yan J (2024) Ars-detr: aspect ratio-sensitive detection transformer for aerial oriented object detection. IEEE Trans Geosci Remote Sens 62:1–15
Zhang K, Shen H (2022) Multi-stage feature enhancement pyramid network for detecting objects in optical remote sensing images. Remote Sens 14(3):579. https://doi.org/10.3390/rs14030579
Zhang S, Wen L, Bian X, Lei Z, Li SZ (2018) Single-shot refinement neural network for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, p 4203–4212
Zhang S, Yuan Q, Li J, Sun J, Zhang X (2020a) Scene-adaptive remote sensing image super-resolution using a multiscale attention network. IEEE Trans Geosci Remote Sens 58(7):4764–4779. https://doi.org/10.1109/TGRS.2020.2966805
Zhang R, Shao Z, Huang X, Wang J, Li D (2020b) Object detection in uav images via global density fused convolutional network. Remote Sens 12(19):3140. https://doi.org/10.3390/rs12193140
Zhang S, Chi C, Yao Y, Lei Z, Li SZ (2020c) Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In: 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR), p 9756–9765. https://doi.org/10.1109/CVPR42600.2020.00978
Zhang H, Chang H, Ma B, Wang N, Chen X (2020d) Dynamic R-CNN: towards high quality object detection via dynamic training. In: Computer vision - ECCV 2020: 16th European conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XV, p 260–275. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-030-58555-6_16
Zhang H, Sun M, Li Q, Liu L, Liu M, Ji Y (2021a) An empirical study of multi-scale object detection in high resolution uav images. Neurocomputing 421:173–182. https://doi.org/10.1016/j.neucom.2020.08.074
Zhang H, An L, Chu VW, Stow DA, Liu X, Ding Q (2021b) Learning adjustable reduced downsampling network for small object detection in urban environments. Remote Sens 13(18):3608. https://doi.org/10.3390/rs13183608
Zhang Z, Lu X, Cao G, Yang Y, Jiao L, Liu F (2021c) Vit-yolo: transformer-based yolo for object detection. In: Proceedings of the IEEE/CVF international conference on computer vision, p 2799–2808
Zhang H, Wang Y, Dayoub F, Sünderhauf N (2021d) Varifocalnet: an iou-aware dense object detector. In: 2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR), p 8510–8519. https://doi.org/10.1109/CVPR46437.2021.00841
Zhang F, Wang X, Zhou S, Wang Y (2022a) Dardet: a dense anchor-free rotated object detector in aerial images. IEEE Geosci Remote Sens Lett 19:1–5. https://doi.org/10.1109/LGRS.2021.3122924
Zhang T, Zhang X, Zhu P, Chen P, Tang X, Li C, Jiao L (2022b) Foreground refinement network for rotated object detection in remote sensing images. IEEE Trans Geosci Remote Sens 60:1–13. https://doi.org/10.1109/TGRS.2021.3109145
Zhang K, Wu Y, Wang J, Wang Y, Wang Q (2022c) Semantic context-aware network for multiscale object detection in remote sensing images. IEEE Geosci Remote Sens Lett 19:1–5. https://doi.org/10.1109/LGRS.2021.3067313
Zhang T, Zhuang Y, Wang G, Dong S, Chen H, Li L (2022d) Multiscale semantic fusion-guided fractal convolutional object detection network for optical remote sensing imagery. IEEE Trans Geosci Remote Sens 60:1–20. https://doi.org/10.1109/TGRS.2021.3108476
Zhang C, Xiong B, Li X, Kuang G (2022e) Aspect-ratio-guided detection for oriented objects in remote sensing images. IEEE Geosci Remote Sens Lett 19:1–5. https://doi.org/10.1109/LGRS.2021.3125502
Zhang Y, Liu X, Wa S, Chen S, Ma Q (2022f) Gansformer: a detection network for aerial images with high performance combining convolutional network and transformer. Remote Sens 14(4):923. https://doi.org/10.3390/rs14040923
Zhang Y-F, Ren W, Zhang Z, Jia Z, Wang L, Tan T (2022g) Focal and efficient iou loss for accurate bounding box regression. Neurocomputing 506:146–157
Zhang C, Lam K-M, Wang Q (2023a) Cof-net: a progressive coarse-to-fine framework for object detection in remote-sensing imagery. IEEE Trans Geosci Remote Sens 61:1–17. https://doi.org/10.1109/TGRS.2022.3233881

Zhang Y, Guo W, Wu C, Li W, Tao R (2023b) Fanet: an arbitrary direction remote sensing object detection network based on feature fusion and angle classification. IEEE Trans Geosci Remote Sens 61:1–11
Zhang T, Sun X, Zhuang L, Dong X, Gao L, Zhang B, Zheng K (2023c) Ffn: fountain fusion net for arbitrary-oriented object detection. IEEE Trans Geosci Remote Sens 61:1–13
Zhang Y, Wu C, Guo W, Zhang T, Li W (2023d) Cfanet: efficient detection of uav image based on cross-layer feature aggregation. IEEE Trans Geosci Remote Sens 61:1–11
Zhang C, Xiong B, Li X, Kuang G (2023e) Tcd: task-collaborated detector for oriented objects in remote sensing images. IEEE Trans Geosci Remote Sens 61:1–14
Zhang J, Xia K, Huang Z, Wang S, Akindele RG (2023f) Etam: ensemble transformer with attention modules for detection of small objects. Expert Syst Appl 224:119997
Zhang J, Lei J, Xie W, Li Y, Yang G, Jia X (2023g) Guided hybrid quantization for object detection in remote sensing imagery via one-to-one self-teaching. IEEE transactions on geoscience and remote sensing
Zhang C, Su J, Ju Y, Lam K-M, Wang Q (2023h) Efficient inductive vision transformer for oriented object detection in remote sensing imagery. IEEE transactions on geoscience and remote sensing
Zhang Z, Mei S, Ma M, Han Z (2024a) Adaptive composite feature generation for object detection in remote sensing images. IEEE transactions on geoscience and remote sensing
Zhang Y, Wu C, Zhang T, Zheng Y (2024b) Full-scale feature aggregation and grouping feature reconstruction based uav image target detection. IEEE transactions on geoscience and remote sensing
Zhang T, Zhang X, Zhu X, Wang G, Han X, Tang X, Jiao L (2024c) Multistage enhancement network for tiny object detection in remote sensing images. IEEE Trans Geosci Remote Sens 62:1–12
Zhang C, Lam K-M, Liu T, Chan Y-L, Wang Q (2024d) Structured adversarial self-supervised learning for robust object detection in remote sensing images. IEEE transactions on geoscience and remote sensing
Zhang X, Feng Y, Zhang S, Wang N, Lu G, Mei S (2024e) Robust aerial person detection with lightweight distillation network for edge deployment. IEEE transactions on geoscience and remote sensing
Zhang Y, Zhang W, Li J, Qi X, Lu X, Wang L, Hou Y (2024f) Empowering lightweight detectors: orientation distillation via anti-ambiguous spatial transformation for remote sensing images. ISPRS J Photogramm Remote Sens 214:244–260
Zhang H, Liu K, Gan Z, Zhu G-N (2025) UAV-DETR: efficient end-to-end object detection for unmanned aerial vehicle imagery. arXiv:2501.01855
Zhao L, Liu T, Xie S, Huang H, Qi J (2022) Ortdet: an orientation robust detector via transformer for object detection in aerial images. Remote Sens 14(24):6329. https://doi.org/10.3390/rs14246329
Zhao Z, Du J, Li C, Fang X, Xiao Y, Tang J (2024a) Dense tiny object detection: a scene context guided approach and a unified benchmark. IEEE transactions on geoscience and remote sensing
Zhao Y, Lv W, Xu S, Wei J, Wang G, Dang Q, Liu Y, Chen J (2024b) Detrs beat yolos on real-time object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), p 16965–16974
Zheng Z, Zhong Y, Ma A, Han X, Zhao J, Liu Y, Zhang L (2020a) Hynet: hyper-scale object detection network framework for multiple spatial resolution remote sensing imagery. ISPRS J Photogramm Remote Sens 166:1–14. https://doi.org/10.1016/j.isprsjprs.2020.04.019
Zheng Z, Wang P, Liu W, Li J, Ye R, Ren D (2020b) Distance-iou loss: faster and better learning for bounding box regression. Proc AAAI Conf Artif Intell 34:12993–13000
Zheng X, Zhang W, Huan L, Gong J, Zhang H (2021a) Apronet: detecting objects with precise orientation from aerial images. ISPRS J Photogramm Remote Sens 181:99–112. https://doi.org/10.1016/j.isprsjprs.2021.08.023
Zheng Y, Sun P, Zhou Z, Xu W, Ren Q (2021b) Adt-det: adaptive dynamic refined single-stage transformer detector for arbitrary-oriented object detection in satellite optical imagery. Remote Sens 13(13):2623. https://doi.org/10.3390/rs13132623
Zheng S, Wu Z, Xu Y, Wei Z (2023) Instance-aware spatial-frequency feature fusion detector for oriented object detection in remote-sensing images. IEEE Trans Geosci Remote Sens 61:1–13
Zhou Q, Yu C (2022) Point rcnn: an angle-free framework for rotated object detection. Remote Sens 14(11):2605. https://doi.org/10.3390/rs14112605
Zhou Z, Zhu Y (2024) Kldet: detecting tiny objects in remote sensing images via kullback-leibler divergence. IEEE transactions on geoscience and remote sensing
Zhou S, Wang F, Huang Z, Wang J (2019a) Discriminative feature learning with consistent attention regularization for person re-identification. In: Proceedings of the IEEE/CVF international conference on computer vision, p 8040–8049
Zhou D, Fang J, Song X, Guan C, Yin J, Dai Y, Yang R (2019b) Iou loss for 2d/3d object detection. In: 2019 International conference on 3D vision (3DV), p 85–94. IEEE
Zhou H, Ma A, Niu Y, Ma Z (2022a) Small-object detection for uav-based images using a distance metric method. Drones 6(10):308

Zhou J, Feng K, Li W, Han J, Pan F (2022b) Ts4net: two-stage sample selective strategy for rotating object detection. Neurocomputing 501:753–764. https://doi.org/10.1016/j.neucom.2022.06.049
Zhu P, Wen L, Bian X, Ling H, Hu Q (2018) Vision meets drones: a challenge. arXiv preprint arXiv:1804.07437
Zhu X, Lyu S, Wang X, Zhao Q (2021a) Tph-yolov5: improved yolov5 based on transformer prediction head for object detection on drone-captured scenarios. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV) workshops, p 2778–2788
Zhu X, Su W, Lu L, Li B, Wang X, Dai J (2021b) Deformable DETR: deformable transformers for end-to-end object detection. In: 9th International conference on learning representations, ICLR 2021, Virtual Event, Austria, May 3–7, 2021. OpenReview.net,
Zhu S, Zhang J, Liang X, Guo Q (2022) Multiscale semantic guidance network for object detection in vhr remote sensing images. IEEE Geosci Remote Sens Lett 19:1–5. https://doi.org/10.1109/LGRS.2021.3089604
Zhu Y, Sun X, Diao W, Wei H, Fu K (2023) Dualda-net: dual-head rectification for cross-domain object detection of remote sensing. IEEE Trans Geosci Remote Sens 61:1–16
Zou Z, Chen K, Shi Z, Guo Y, Ye J (2023) Object detection in 20 years: a survey. Proc IEEE 111(3):257–276
Wang, L., Mu, X., Ma, C., Zhang, J.: Hausdorff iou and context maximum selection nms: Improving object detection in remote sensing images with a novel metric and postprocessing module. IEEE Geoscience and Remote Sensing Letters 19, 1–5 (2021)