


Adaptive Dynamic Label Assignment for Tiny Object Detection in Aerial Images

Lihui Ge , Student Member, IEEE, Guanqun Wang , Member, IEEE,

Tong Zhang , Graduate Student Member, IEEE, Yin Zhuang , Member, IEEE, He Chen , Member, IEEE,

Hao Dong , Member, IEEE, and Liang Chen , Member, IEEE

Abstract—Tiny object detection is one of the most difficult and critical tasks in remote sensing intelligent interpretation applications. Compared with standard-size object detection, detecting tiny objects is more challenging as they typically contain fewer pixels. Besides, the metrics based on intersection-over-union (IoU) are more sensitive to their positioning bias. However, current mainstream object detectors usually assign samples to the ground truth (GT) according to a fixed IoU threshold, which would lead to a certain number of tiny objects fail to be assigned with high IoU conditional anchors as positive sample candidates under a static threshold. Consequently, insufficient positive samples would affect model training to further constrain the detection performance for tiny objects. In this article, a sample selection strategy called adaptive dynamic label assignment is proposed to optimize the training effectiveness and improve tiny object detection performance. First, sample allocation thresholds are individually assigned for each GT based on their shape, size, and positions on the feature map. Second, the sample sets are dynamically adjusted during training by using a newly designed indicator called dynamic IoU. Finally, with the guidance of this adaptive dynamic label assignment strategy, each GT can acquire sufficient positive samples for practical training. Extensive experiments on the AI-TOD and Levir-Ship datasets show that, compared with the baseline model, the tiny object detectors trained by our proposed adaptive dynamic label assignment strategy can significantly improve the tiny object detection performance without increasing storage space and inference time. Our method exhibits high portability and outperforms the state-of-the-art methods.

Index Terms—Adaptive threshold (AT), label assignment, remote sensing, tiny object detection.

I. INTRODUCTION

TINY object detection is a classical problem in remote sensing intelligent interpretation with broad application

Manuscript received 14 August 2023; revised 7 November 2023 and 29 January 2024; accepted 17 March 2024. Date of publication 20 March 2024; date of current version 23 May 2024. This work was supported in part by the National Science Foundation for Young Scientists of China under Grant 62101046, in part by Space based on Orbit Real-Time Processing Technology under Grant 2018-JCJQ-ZQ-046, in part by the National Natural Science Foundation of China under Grant 62136001, and in part by Multisource Satellite Data Hardware Acceleration Computing Method with Low Energy Consumption under Grant 2021YFA0715204. (Corresponding author: Guanqun Wang.)

Lihui Ge, Tong Zhang, Yin Zhuang, He Chen, and Liang Chen are with the National Key Laboratory of Science and Technology on Space-Born Intelligent Information Processing, Beijing Institute of Technology, Beijing 100081, China.

Guanqun Wang is with the National Key Laboratory of Science and Technology on Space-Born Intelligent Information Processing, Beijing Institute of Technology, Beijing 100081, China, and also with the School of Computer Science, Peking University, Beijing 100871, China (e-mail: wgq@pku.edu.cn).

Hao Dong is with the Center on Frontiers of Computing Studies (CFCS), Peking University, Beijing 100871, China.

Digital Object Identifier 10.1109/JSTARS.2024.3379515

scenarios, including marine surveillance and intelligent transportation [1], [2]. Followed MS COCO [3] definition, the objects with fewer than 16^2 pixels are tiny objects, whereas objects within 16^2 and 32^2 pixels are defined as small objects. Remote sensing images usually have a long shooting distance with an overhead view, which generates massive tiny objects in them. These objects typically exhibit characteristics of blurred appearances, dense distribution, and complex background environments, which are visually demonstrated in Fig. 1. Despite the impressive performance of mainstream detectors [4], [5], [6], [7], [8], [9] in object detection tasks within natural scenes, their performance is suboptimal when it comes to tiny object detection tasks in remote sensing scenes. Therefore, it is worth conducting in-depth research on tiny object detection in remote sensing images.

To enhance the tiny object detection performance, research works from multiple aspects have been carried out and made progress. First, research works on feature extraction and fusion techniques have been extensively conducted. Enhanced feature extraction modules [9], [10], [11] are designed to extract richer feature information from ambiguous object appearances. The integration of contextual information [12], [13] can effectively reduce the interference from complex backgrounds. Improvements in feature extraction and fusion can enhance the tiny object detection capability. However, due to the low tolerance of tiny objects to bounding box perturbations, a certain number of tiny objects fail to be assigned with high intersection-over-union (IoU) conditional anchors as positive sample candidates under a static threshold. The suboptimal label assignment results in the features of tiny objects, which are not assigned with positive samples ignored, and improvements in feature extraction and fusion cannot change this dilemma either. Therefore, only the improvement of the label assignment strategy can fundamentally enhance the detection performance of tiny objects. Second, methods [14], [15], [16] based on generative adversarial networks (GAN) [17] utilize image super-resolution techniques to enrich the feature representation of tiny objects within the image, thereby achieving better tiny object detection. However, GAN-based approaches often lack stability, as the generated fake object information would impact the relationship between context and objects. Besides, the generated super-resolution images increase the size of the input image, resulting in longer inference times for the detector. Third, research on postprocessing procedures has also attracted some attention. By replacing

nonmaximum suppression (NMS) [18] with Soft-NMS [19], the detection challenges arising from dense distribution of objects can be effectively addressed. However, the effectiveness of this method is based on the premise that the detector has a fairly high recall rate of tiny objects, which is precisely where tiny object detectors fall short. The aforementioned methods have indeed demonstrated improvements in the detection performance of tiny objects to some extent. However, they overlook the fundamental issue in training the detection model, that is how to assign labels more accurately, enabling the detector to distinguish between objects and backgrounds precisely. Label assignment strategies for standard-size objects, such as adaptive training sample selection (ATSS) [20] and AutoAssign [21], which entail high computational complexity and depend on hyperparameters, they struggle to effectively adapt to the challenges of tiny object detection label assignment. Related works on tiny object detection label assignment, for example, Wang et al. [22] conducted a study on the quality of anchors, modeled the bounding box as a 2-D Gaussian distribution, and proposed a new Wasserstein distance-based evaluation metric for tiny object detection label assignment. Xu et al. [23] proposed Gaussian receptive field-based label assignment (RFLA), which utilizes the receptive field distance (RFD) to measure the similarity between the Gaussian receptive field and ground truth (GT), and assigns labels accordingly. These related works primarily innovate the evaluation process of label assignment by introducing new metrics to enhance the assessment of the matching degree between anchors and GT. Moreover, the influence of the uniqueness of each GT (e.g., size, shape, and distribution) on label assignment is neglected.

In our study, we delve into the assignment process of label assignment, conducting a systematic analysis on how tiny object GT characteristics, such as size, shape, and distribution, influence label assignment. This facilitates a more adaptable determination of positive sample quantities for each GT, improving the detection performance of existing detectors for tiny objects without introducing additional model parameters or increasing inference time. In this article, an adaptive dynamic label assignment method is proposed, which can improve model detection performance by guiding the sample selection of each GT according to its shape, size, and distribution characteristics. First, a simple yet effective threshold-calculating formula is designed for assigning positive and negative samples. Specifically, each GT is assigned with individual thresholds, and the calculation takes into account of factors, such as the size, shape, and distribution characteristics of each GT. It ensures that each GT can be matched with sufficient positive samples during training. Second, an indicator is designed to evaluate the potential of anchor localization objects, which comprehensively considers both the quality of the preset anchors and the progress in model optimization. It dynamically adjusts the sets of positive and negative samples, which prioritize high-quality anchors as positive samples while discarding negatively optimized anchors as negative samples. Experimental results on AI-TOD and Levir-Ship datasets demonstrate that our proposed label assignment strategy improves the adaptability of the existing detectors to tiny object detection without increasing any model parameters or inference

time. Our method fundamentally addresses the issue of learning confusion in the training phase of object detectors and exhibits excellent compatibility, outperforming existing methods.

The contributions of this article are summarized as follows.

- 1) A strategy named adaptive dynamic label assignment is proposed, which can be used in the training stage to enhance the detection performance of the baseline model for tiny objects. This strategy holds significant importance for the advancement of tiny object detection techniques.
- 2) The low tolerance of tiny objects to bounding box perturbations is comprehensively analyzed and demonstrated that even slight regression deviations would lead to drastic changes in the IoU, thus affecting label assignment and object localization.
- 3) An adaptive threshold (AT) calculation method is proposed, which aims to set positive and negative sample thresholds for each GT separately to achieve high-quality sample selection.
- 4) A label assignment metric named dynamic IoU is proposed from the aspect of matching degree [24], which can flexibly adjust the sets of positive and negative samples during training to improve the quality of tiny object detection.

II. RELATED WORKS

A. Object Detection

Object detection technology is mainly used to classify and locate objects in images. With the rapid development of deep learning and convolutional neural networks (CNNs) [25], [26], [27], [28], [29], research on object detection has made remarkable achievements. Existing object detectors can be categorized into anchor based and anchor-free based on whether the detector requires predefined anchors or not. Moreover, according to whether the region of interest needs to be extracted from input images, anchor-based object detection methods can be further divided into one-stage detection models [4], [5], [30], [31] and two-stage detection models [7], [8], [32]. Typically, two-stage detectors demonstrate superior detection performance, whereas one-stage detectors exhibit faster detection speed. With the proposal of RetinaNet [8] and the utilization of focal loss, the issue of suboptimal detection performance in detectors caused by sample imbalance has been alleviated. Consequently, the disparity in detection accuracy between one-stage and two-stage detectors has significantly decreased. This progress has enabled one-stage detectors to achieve efficient and high-precision detection. Besides, inspired by pixel-level semantic segmentation, the DenseBox [33] embarked on pixelwise classification and regression experiments on the output feature map, representing an early exploration of anchor-free object detection algorithms. Subsequently, key-point prediction models represented by CornerNet [34] and CenterNet [35], as well as semantic maps-based detection models represented by FCOS [36] and FoveaNet [37], have emerged successively. Anchor-free object detection models discard predefined anchors and significantly improve object detection speed. In recent years, with the proposal of attention mechanism [38] and the application of transformer in natural

language processing, researchers have designed object detection models [39], [40], [41] based on vision transformer to realize the prediction of object location and category information through a set of input queries. Despite the numerous outstanding research efforts and significant technological advancements in the field of object detection, the performance in detecting tiny objects in remote sensing images still remains unsatisfactory. Therefore, this article aims to address and explore the challenges of tiny object detection in aerial images.

B. Tiny Object Detection

Tiny objects often lack detailed information about their appearance compared with objects of standard size. Mainstream feature extractors typically reduce spatial redundancy and learn high-level features by downsampling the feature map, which inevitably causes the loss of feature information of tiny objects. In addition, tiny objects have a low tolerance for bounding box perturbation. Therefore, although object detection algorithms have made significant progress, the detection performance for tiny objects is still unsatisfactory. In order to enhance the detection performance of tiny objects, extensive efforts have been undertaken in the following aspects.

Feature Extraction and Fusion: Some researchers put forward feature extraction and fusion methods based on multiscale features. Lin et al. [42] improved object detection accuracy, especially for tiny objects, by incorporating a feature pyramid network (FPN) in the object detection model. Pang et al. [11] proposed a unified and self-reinforced CNN under the end-to-end training framework called \mathcal{R}^2 -CNN, which uses the intermediate global attention block to enlarge the receptive field to inhibit false positives. Gong et al. [12] proposed the concept of fusion factor for controlling the information passed from deep features to shallow features to make FPN more adaptable to tiny object detection. Lim et al. [13] used additional features from different layers as contexts by connecting multiscale features to improve the accuracy of detecting tiny objects. Liu et al. [10] designed the high-resolution detection network, which uses multidepth backbones to receive multiresolution inputs, maintaining the advantages of high-resolution images without introducing new problems. Li et al. [9] experimentally verified the effect of the receptive field on the detection of objects at different scales and proposed TridentNet, which uses a dilated convolution instead of a standard convolution to adjust the receptive field to improve the detection performance of the model for objects at different scales. Qiao et al. [43] introduced recursive feature pyramid and switchable atrous convolution and proposed DetectoRS. DetectoRS demonstrates heightened robustness in detecting occluded objects, resulting in a significant enhancement in object detection performance. Chalavadi et al. [44] proposed network for multiscale object detection in aerial images using hierarchical dilated convolutions (mSODANet), which utilizes the bidirectional feature aggregation module to incorporate dense multiscale contextual features. mSODANet achieved effective multiscale object detection in aerial images. Wu et al. [45]

proposed feature-and-spatial aligned network (FSANet), which utilized the alignment mechanism and progressive optimization strategy to obtain more discriminative features and accurate localization results. Li et al. [46] proposed a mask augmented attention feature pyramid network (MA2-FPN) to detect tiny objects in remote sensing images. In this network, attention enhancement module (AEM) aggregates tiny target context and spatial feature information by large kernel separable convolutional attention mechanism, and mask supervision module supervises AEM through a segmentation attention loss to aggregate attention information more accurately while suppressing the influence of irrelevant background. Zhang et al. [47] introduced the attention module (SPAM) to filter out the background noise in the shallow feature extraction to better extract small object features.

GAN-based Methods: Some researchers, on the other hand, have implemented image super-resolution based on GAN [17] to improve tiny object detection performance by enriching the feature representation of tiny objects. The perceptual GAN [14] internally enhance the representation of tiny objects to super-resolution representation to achieve a feature representation similar to that of large objects to improve the discriminative power of the detection model. Inspired by the success of edge-enhanced GAN [48] and enhanced super-resolution GAN [49], Rabbi et al. [15] adopted a novel edge-enhanced super-resolution GAN to improve the quality of remote sensing images. An attention-based feature interaction method, multiresolution attention extractor [16], is proposed as a general-purpose feature extractor with significant improvements in tiny object detection capacity.

Tiny Object Learning Strategy: In addition, some researchers expect to improve the tiny object detection performance by designing better detection mechanisms and training strategies. Chen et al. [50] designed an additional network branch called degraded reconstruction enhancer, which uses the object-aware blurred version of the learning regression input image in the training phase, and the reconstruction branch enables the backbone network to focus more on the object region rather than the background. Wang et al. [22] conducted a study on the quality of anchors, modeled the bounding box as a 2-D Gaussian distribution, and proposed a new Wasserstein distance-based evaluation metric for tiny object detection instead of the commonly used IoU criterion, which improved the sensitivity of tiny objects to location. Li et al. [51] designed a new tiny object loss constraint term to attempt at overcoming the challenge of tiny object detection in few-shot aerial image object detection. QueryDet [52] uses a novel query mechanism to speed up the inference of an FPN-based object detector. Koyun et al. [53] and Bosquet et al. [54] investigated the tiny object detection model and divide the tiny object detection into two stages: the first stage filters the original image to generate the focal region containing the object, and the second stage detects the tiny objects within the focal region. Wang et al. [55] proposed a unified framework called feature-merged single-shot detection (FMSSD) network, which aggregates the context information both in multiple scales and the same scale feature maps. In

addition, a novel area-weighted loss function is introduced to guide the framework to pay more attention to small objects.

However, the studies mentioned above have failed to address the underlying cause of the suboptimal performance of detectors in detecting tiny objects. The less-than-ideal label assignment results hinder the model's ability to learn the complete features of samples, ultimately limiting the achievement of optimal performance in tiny object detection.

C. Label Assignment

Label assignment refers to the process by which the object detector classifies anchors into positive and negative samples during the training phase. Label assignment affects model detection performance by determining the learning target for each anchor on the feature map. For the classical anchor-based detection method RetinaNet [8], it assigns positive and negative samples by calculating the IoU between the predefined anchors and GTs where anchors with IoU above the static threshold are assigned as positive samples and vice versa as negative samples. The anchor-free method FCOS [36] determines the positive and negative samples based on the spatial distance between the points on the feature map and the center of GT. CenterNet [35] introduces the Gaussian kernel function, which is applied to map the GT to the feature map. The positive and negative samples are determined according to whether the pixels fall into GT's Gaussian distribution. Those contained within the Gaussian kernel are assigned as positive samples and vice versa as negative samples. Besides, ATSS [20] is a label assignment strategy that can automatically select positive and negative samples based on the statistical characteristics of the GT. AutoAssign [21] implements label assignment through a CNN-based adaptive learning method. Ge et al. [56] formulated the label assignment process as an optimization problem in optical transport. Li et al. [57] explored a novel weighting paradigm known as dual weighting, which assigns separate weights to positive and negative samples. Ming et al. [24] proposed a dynamic anchor learning method that utilizes the newly defined matching degree to comprehensively evaluate the localization potential of the anchors and carries out a more efficient label assignment process. Zhang et al. [58] determined the thresholds for positive and negative samples at different training iterations based on a combination of anchor IoU and bounding box IoU calculation. Qian et al. [59] proposed a pseudosoft label assignment strategy to assign a more precise soft label for each instance, where the soft label is determined by the spatial distance between each instance and its nearest pseudoground-truth instance. For tiny object detection label assignment, Xu et al. [23] proposed Gaussian RFLA, which utilizes the RFD to measure the similarity between the Gaussian receptive field and GT, and assigns labels accordingly. On this basis, Fu et al. [60] proposed Gaussian probabilistic distribution-based fuzzy similarity metric (GPM) and the adaptive dynamic anchor mining strategy (ADAS). The combination of ADAS and GPM in the anchor-based object detector addresses the issue of inaccurate similarity measurement between small bounding boxes and predefined anchors, achieves accurate label assignment. Improvements in label assignment can fundamentally address the issue of detection performance during model training,

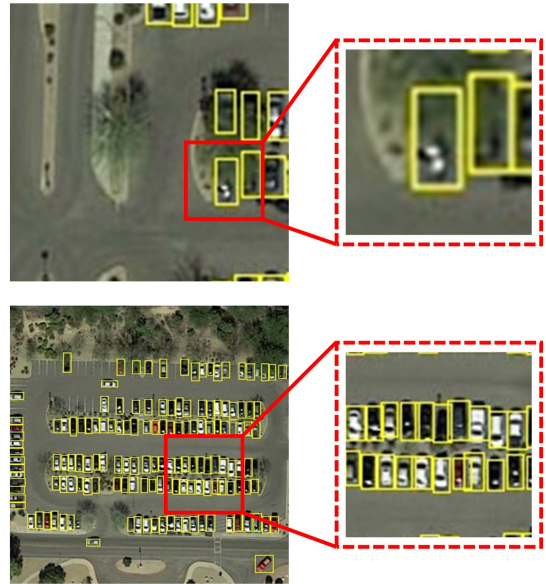


Fig. 1. Tiny objects in remote sensing images typically exhibit the following characteristics: features are blurry, distributed densely, and the background is complex.

making it worthy of thorough exploration. Thus, we rethink the label assignment strategy and delve into the sample selection problem in the context of tiny object detection, conducting an in-depth investigation in this article.

III. PROPOSED METHOD

In this section, we first conducted a meticulous analysis of the challenges in tiny object detection compared with standard-size object detection. Subsequently, our proposed AT calculation method and dynamic IoU metric are described in detail as follows.

A. Problem Analysis

The object detection model assigns learning objectives to each anchor based on the label assignment strategy in the training phase, and the quality of positive and negative samples directly affects the final detection performance of the model. First and foremost, compared with detecting objects of standard sizes, tiny object detection exhibits lower tolerance to bounding box perturbations. Fig. 2 demonstrates the sensitivity of different object sizes to positional deviation through changes in IoU. In Fig. 2(a), when the GT size is same as the anchor size, the IoU variation becomes more noticeable as the GT size decreases. In Fig. 2(b), GT with different anchor sizes, the IoU variation becomes more dramatic as the size gap between GT and anchors increases. The second point is, the shape of the GT affects its distribution of IoU with anchors on the feature map. When the GT sizes are the same, standard square GT tends to match more high IoU conditional anchors as candidate positive samples compared with elongated GT. Fig. 3 presents the distribution of IoU between the anchors and GT on the feature map for both square-shaped GT ($A_r = 1$) and elongated GT ($A_r = 0.25$). With a fixed threshold of 0.5, the square-shaped GT in Fig. 3(a)

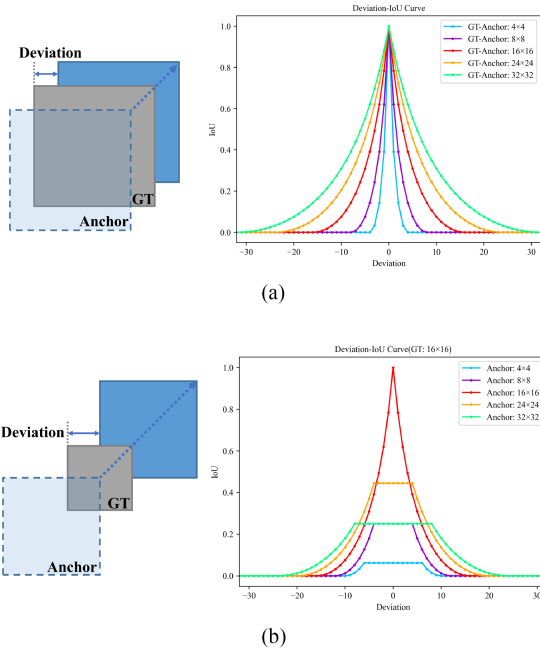


Fig. 2. Deviation-IoU curve of GT. (a) Same size GT and anchor. When the GT and anchor deviate from optimal spatial matching by one unit, the IoU of the tiny object GT changes more dramatically. (b) Different sizes of GT and anchor. The curve with a significant difference in size between GT and anchor changes more dramatically. Furthermore, the standard size anchors are often much larger than the tiny object GT, which makes even if the anchors are spatially matched to the GT, the IoU of both fail to reach the positive sample threshold.

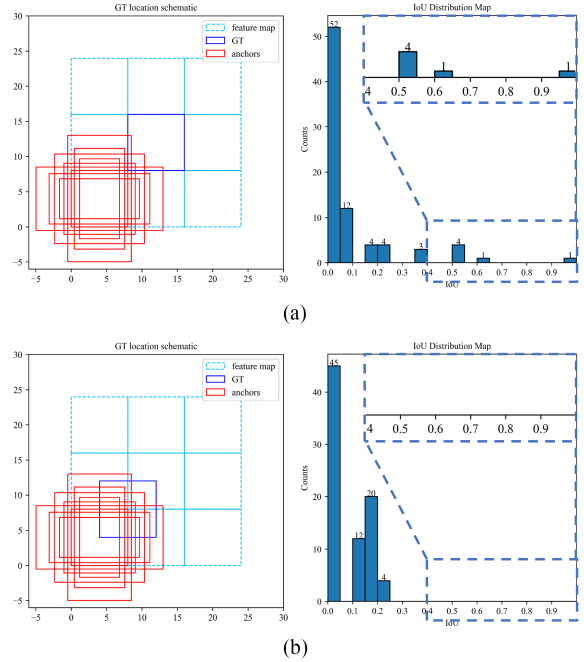


Fig. 4. IoU distribution of GT at different positions relative to the anchors on the feature map. (a) GT at the center of the feature pixel. (b) GT deviating from the center position of the feature pixel. When the GT deviates from the center of the feature pixel, the anchors distributed in the high IoU conditional region decrease, leading to a reduction in candidate positive samples.

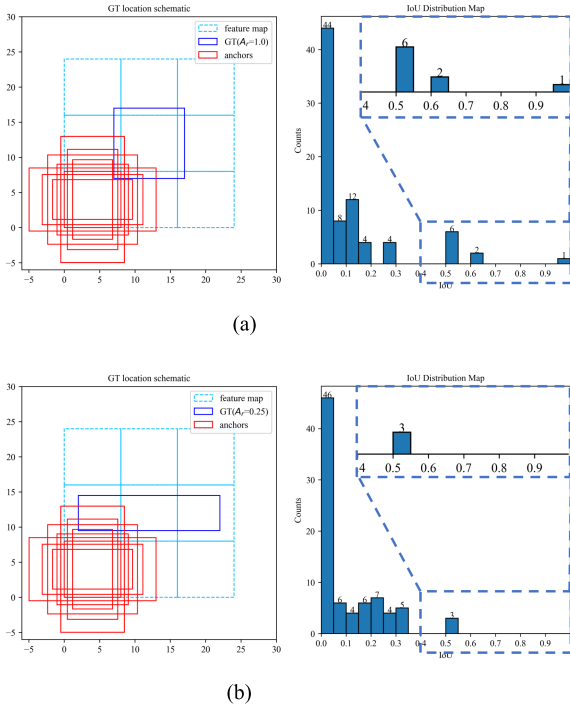


Fig. 3. IoU distribution of GT with different aspect ratios relative to the anchors on the feature map. As the A_T approaches 1, the GT becomes closer to a standard square shape. (a) $A_T = 1$, using 0.5 as the threshold, nine candidate positive samples can be found on the feature map. (b) $A_T = 0.25$, using 0.5 as the threshold, only three candidate positive samples can be found on the feature map. As the A_T deviates from 1, the anchors distributed in the high IoU conditional region decrease, leading to a reduction in candidate positive samples.

obtained nine positive samples, while the elongated GT in Fig. 3(b) got none. Furthermore, the position of the object on the feature map affects the distribution of its IoU with anchors. GTs located at the center of feature pixels have more anchors distributed in the high IoU region, whereas when the object deviates from the center of feature pixels, the number of high IoU anchors decreases. Fig. 4 shows the statistical results of the distribution of IoU between objects and anchors for different relative positions with respect to feature pixels. While the GT sizes are the same in Fig. 4(a) and (b), the GTs in Fig. 4(b) that are not located at the center of the feature pixels have no candidate positive samples, whereas the GTs in Fig. 4(a) at the center of the feature pixels have five candidate positive samples. Obviously, the number of high IoU anchors decreases as the object deviates from the center of the feature pixel.

The presence of the abovementioned characteristics in tiny objects often makes it challenging for corresponding GTs to achieve high IoU matching with the preset anchors on the feature map. In the training phase of the detector, the traditional label assignment method using a fixed threshold fails to allocate sufficient positive samples to each tiny object GT. As a result, the detector cannot achieve optimal performance in tiny object detection due to suboptimal learning targets. Although this problem can be effectively alleviated by using high-resolution feature layers or laying 1 more anchors on the feature map, but this also leads to an increase in model size and a decrease in inference efficiency. We aim to achieve optimal detection of tiny objects by improving the sample selection strategy to assign reasonable

learning targets for each anchor, without increasing the model size and inference time.

B. Adaptive Threshold

Setting appropriate positive and negative sample thresholds is crucial to achieving accurate label assignment and improving the detection performance of object detectors for tiny objects, without relying on more high-resolution feature maps or increasing the number of preset anchors. Through the study of sample set partitioning threshold, a method called AT is proposed. This method addresses the challenge of designing thresholds for positive and negative samples in the training stage of tiny object detection models. Specifically, considering factors, such as the GT size, shape, and distribution, an appropriate threshold for the positive and negative sample division is calculated for each GT, enabling the assignment of an appropriate set of positive and negative samples. The proposed AT can be formed as follows:

$$T_{rp} = \text{IoU}_{\max} \times \frac{3 - \text{IoU}_{\max}^2}{4} \times \text{mp} \quad (1)$$

$$T_{rn} = T_{rp} \times \frac{4 - (\text{IoU}_{\max} - 1)^2}{5} \quad (2)$$

where IoU_{\max} represents the maximum IoU of the GT with all anchors, and T_{rp} and T_{rn} are the reference thresholds for positive and negative samples, respectively, which are calculated by using our proposed method. mp is a modulation parameter used to modulate the effect of the aspect ratio (ratio of the short side to the long side) of the GT on the positive sample threshold, and it can be calculated from the following equation based on aspect ratio:

$$\text{mp} = \frac{4 + A_r}{5} \quad (3)$$

$$A_r = \frac{\text{GT}_{\text{short-side}}}{\text{GT}_{\text{long-side}}}. \quad (4)$$

For GT with the same area, GT with shape closer to square is more likely to match suitable anchors. Conversely, as the aspect ratio deviates from 1, the GT would only match fewer anchors when the same threshold is applied. To address this, the modulation parameter A_r was introduced, which represents the aspect ratio of the GT and is calculated by (4), where $\text{GT}_{\text{long-side}}$ represents the long side of the GT, and $\text{GT}_{\text{short-side}}$ represents the short side of the GT. The value of A_r falls within the range of $(0, 1]$. As A_r approaches 1, indicating a more square-shaped GT, the influence of the aspect ratio is not considered when calculating the positive sample threshold. Consequently, when the shape of the GT changes, A_r can adjust the threshold downward enables the GT to match a sufficient number of anchors.

As shown in Fig. 5, two GTs with the same size but different shapes were selected, and their IoU distribution with anchors on the feature map were calculated. Fig. 5(a) and (b) shows the positive sample assignment thresholds without considering the influence of GT aspect ratio, the square GT ($A_r = 1$) in Fig. 5(a) obtained five anchors as positive samples, while the nonsquare GT ($A_r = 0.25$) in Fig. 5(b) only got two positive samples. In order to make GT with different shapes can be assigned

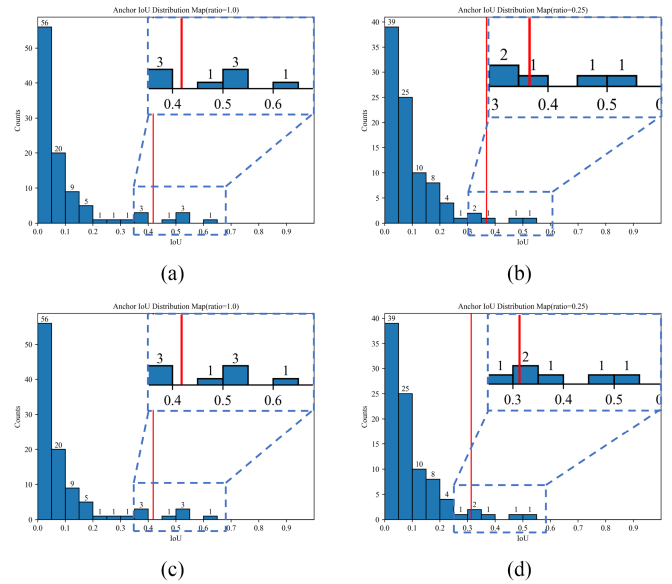


Fig. 5. Relationship between the object adaptation threshold and the IoU distribution of the anchor and GT before and after using A_r (the red line represents the object adaptation threshold). The number of positive samples that can be allocated to the two GTs of the same size but different aspect ratios according to the object adaptation threshold without considering the effect of A_r on the threshold is shown as: (a) GT ($A_r = 1$) obtained five positive samples; (b) GT ($A_r = 0.25$) obtained two positive samples. The positive samples obtained by the above two GTs after adjusting the threshold with A_r are shown as: (c) GT ($A_r = 1$) still obtained five positive samples; (d) GT ($A_r = 0.25$) obtained five positive samples.

to enough positive samples, when calculating the threshold for each GT, the object with a smaller aspect ratio should have a relatively smaller positive sample threshold. With the addition of the shape modulation parameter A_r , both square and nonsquare GT can be assigned to five positive samples, as shown in Fig. 5(c) and (d)

$$T_p = \max(T_{rp}, T_{p\min}) \quad (5a)$$

$$T_n = \max(T_{rn}, T_{n\min}). \quad (5b)$$

Mislabeled GT or other reasons may cause the $\text{IoU}_{\max} \ll 1$ and thus affect the training stability. To suppress this interference, (5) is used to constrain the minimum values of positive and negative sample thresholds. In the experiment, we set $T_{p\min} = 0.1$ and $T_{n\min} = 0.05$. Through this constraint, we eliminate the interference of mislabeled GT on training, but may cause some GT ($\text{IoU}_{\max} \ll 1$) with extreme shape sizes to fail to assign positive samples. Some methods to forcibly specify positive samples can solve this problem. However, we cannot ensure that anchors of low quality in the initial stage can return to high-quality prediction after training. To solve this problem, a regression IoU-guided dynamic label assignment method was proposed.

C. Training-Based Dynamic Label Assignment

When studying arbitrary-oriented object detection, Ming et al. [24] observed that the localization performances of the samples assigned to the GT were inconsistent with the learning goal.

After training, partial anchors selected as positive samples are returned to low-quality positioning boxes, whereas the anchor of partially negative samples return to high-quality positioning boxes. This phenomenon can be called as sample isolation. They introduced the concept of matching degree (md), which comprehensively utilizes the spatial matching of anchors and the posterior information of training regression. The matching degree is defined as follows:

$$md = \alpha A_{IoU} + (1 - \alpha) R_{IoU} - u^\gamma \quad (6)$$

$$u = |A_{IoU} - R_{IoU}| \quad (7)$$

where A_{IoU} represents the IoU of anchor and GT, R_{IoU} represents the IoU of regression box and GT, and α and γ are hyperparameters used for weighting. u is a penalty factor, which is obtained by calculating the difference of IoU between anchor and GT before and after regression, and u characterizes the change value of anchors before and after regression.

It is believed that this idea can also be used to solve the problem that GT cannot be assigned to positive samples because of the low maximum IoU of the anchor to GT. At the same time, as tiny objects are more sensitive to position deviation, the problem of sample isolation is more serious. The adaptability of (6) was researched to guide label assignment in tiny object detection models. Theoretically, when training improves the positioning quality of the anchor ($R_{IoU} > A_{IoU}$), the md should be greater than A_{IoU} , but there is a problem that md is less than A_{IoU} when the value of A_{IoU} is low ($A_{IoU} \leq \alpha$), which is considered illogical, as shown in Fig. 6(a). Moreover, due to the high sensitivity of tiny object detection to position deviation, low IoU samples are very common, which leads to the fact that (6) does not have the ability to be used directly in the assignment of tiny object detection labels. Based on this, the training-based dynamic label assignment (DLA) was proposed, which employs dynamic IoU to dynamically adjust the sample sets during training. D_{IoU} is defined as

$$D_{IoU} = \alpha A_{IoU} + (1 - \alpha) R_{IoU} - \beta u^\gamma. \quad (8)$$

The definitions of A_{IoU} , R_{IoU} , α , γ , and u are the same as those in (6). The weight factor β is added to adjust the penalty term, so that the indicator is applicable to the low IoU samples in tiny object detection. β can be determined in two ways: one is determined dynamically according to (9) based on the value of α , and the other is to use a fixed value according to experience. The range of D_{IoU} is from 0 to 1. The higher the value of D_{IoU} , the more likely the anchor is to regress to accurate detection. Fig. 6(b) shows that our D_{IoU} has better adaptability to positive samples of low IoU

$$\beta = 1 - \alpha. \quad (9)$$

In particular, our adaptive dynamic label assignment strategy follows the following procedure. First, before training commences, we calculate the IoU between GT and predefined anchors and derive T_p and T_n based on (1) and (2). Subsequently, during label assignment, the D_{IoU} of each anchor is calculated, and anchors are designated as positive samples when D_{IoU} is greater than T_p , while anchors with D_{IoU} less than T_n are

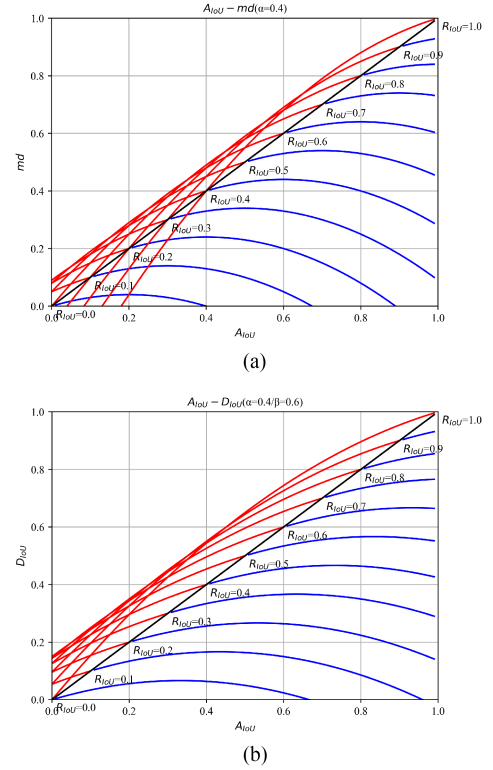


Fig. 6. Differences between md and D_{IoU} were compared under the same conditions. (a) When using md to evaluate the quality of the anchor, there is a problem that when the anchor IoU is lower than α , although the network makes the anchor tend to detect accurately ($R_{IoU} > A_{IoU}$), the red line in the figure), md is smaller than the anchor IoU, which does not meet the theoretical requirements. (b) Using D_{IoU} ($\beta = 1 - \alpha$) to evaluate the quality of the anchor, when the network makes the anchor tend to detect accurately, D_{IoU} constant is greater than anchor IoU, which meets our theoretical requirements.

categorized as negative samples, with the remaining anchors being disregarded. The entire process is detailed in Algorithm 1. For the stability of training, dynamic IoU is not used as the basis for label assignment in the initial stage of training, but keep increasing the influence of regression IoU on sample selection with training iteration. The α for each stage is adjusted following the following criteria:

$$\alpha(p, \alpha_0) = \begin{cases} 1, & 0 \leq p < 0.1 \\ \frac{\alpha_0 - 1}{0.5 - 0.1}(p - 0.1) + 1, & 0.1 \leq p < 0.5 \\ \alpha_0, & p \geq 0.5 \end{cases} \quad (10)$$

where $p = \frac{\text{epoch}}{\text{epochs}}$, epochs is the total number of training epochs, epoch is the current training epoch, and α_0 is a predefined hyperparameter. α in (8) and (9) represents the actual weight of α_0 at the current training epoch, which is determined specifically by (10).

IV. EXPERIMENTS

A. Datasets

AI-TOD [61] benchmark is a remote sensing dataset for tiny object detection, which is built by Wuhan University based on public large aerial image datasets. It contains 280 036 aerial

Algorithm 1: Adaptive Dynamic Label Assignment Strategy.

Input: \mathcal{G} is a set of ground-truth boxes on the image \mathcal{A} is a set of all anchor boxes $\mathcal{M}(\cdot)$ represents the regression process of the model to the anchor $IoU(a, b)$ represents the calculation process of IoU between a and b $epochs$ is the total number of training iterations α_0 and γ are the predefined hyper-parameters**Output:** \mathcal{P} is a set of positive samples \mathcal{N} is a set of negative samples \mathcal{I} is a set of ignore samples

- 1: $\mathcal{P}, \mathcal{N}, \mathcal{I} \leftarrow \emptyset$
 - 2: Compute T_p and T_n for each GT using (1), (2) and (5)
 - 3: **for** $epoch = 1, 2, 3, \dots, epochs$ **do**
 - 4: $p = \frac{epoch}{epochs}$
 - 5: $\alpha = \alpha(p, \alpha_0)$, where $\alpha(p, \alpha_0)$ is expressed by (10)
 - 6: $\beta = 1 - \alpha$
 - 7: **for** each anchor box $a \in \mathcal{A}$ **do**
 - 8: $A_{IoU} = IoU(a, \mathcal{G})$
 - 9: $R_{IoU} = IoU(\mathcal{M}(a), \mathcal{G})$
 - 10: $D_{IoU} = \alpha A_{IoU} + (1 - \alpha)R_{IoU} - \beta(A_{IoU} - R_{IoU})^\gamma$
 - 11: Compare D_{IoU} with T_p and T_n
 - 12: Put a into \mathcal{P}, \mathcal{N} or \mathcal{I}
 - 13: **end for**
 - 14: **end for**
 - 15: **return** $\mathcal{P}, \mathcal{N}, \mathcal{I}$
-

images with the size of 800×800 pixels, containing a total of 700 621 object instances. The dataset contains eight object classes: *airplane* (AI), *bridge* (BR), *storage-tank* (ST), *ship* (SH), *swimming-pool* (SP), *vehicle* (VE), *person* (PE), and *wind-mill* (WM). Compared with the existing object detection datasets in aerial images, the average side length of the object in AI-TOD is only 12.8 pixels, which is smaller than the general remote sensing dataset DOTA [62] (55.2 pixels). The training set of 11 214 images and the validation set of 2804 with a total of 14 018 images are used for training, and the test set of 14 018 images are used to evaluate the model performance.

LEVIR-SHIP [50] concentrate on tiny ship objects with a lower spatial resolution. The dataset consists of images captured from multispectral cameras of GaoFen-1 and GaoFen-6 satellites. Initially, 85 scenes with pixel resolution between $10\,000 \times 10\,000$ and $50\,000 \times 20\,000$ were cropped to create 3896 remote sensing images with a resolution of 512×512 . These images were then split into training, validation, and test sets in a 3:1:1 ratio. The dataset contains labels for 1973 tiny ship objects. In Levir-Ship, ship pixels are typically below 20×20 and centralizes at around 10×10 , which is relatively small compared with the vast background.

TABLE I

EVALUATION RESULTS OF DIFFERENT ANCHOR SIZES ON AI-TOD DATASET

Baseline	Anchor size	AP ₅₀	AP ₇₅	AP
RetinaNet	Standard-size subsize	11.2 31.3	2.6 6.0	4.7 11.3

B. Evaluation Metric

Average Precision (AP): The AP metric is used to evaluate the performance of the proposed method. Specifically, AP₅₀ means the IoU threshold of defining true positive (TP) is 0.5, AP₇₅ means the IoU threshold of defining TP is 0.75, and AP means the average value from AP₅₀ to AP₉₅, with an IoU interval of 0.05. AP_{vt} represents the AP for objects with an absolute size within $[0, 8^2]$ pixels, AP_t represents the AP for objects with an absolute size within $[8^2, 16^2]$ pixels, AP_s represents the AP for objects with an absolute size within $[16^2, 32^2]$ pixels, and AP_m represents the AP for objects with an absolute size within $[32^2, +\infty]$ pixels.

Parameters: Parameters metric represents the total number of parameters in the detector that need to be trained. More parameters means more storage space. Megabits are used to measure parameters.

Floating Point Operations (FLOPs): FLOPs metric refers to a floating-point operand, understood as the amount of computation. FLOPs can be used to measure the complexity of an algorithm/model. GFLOPs are used to measure FLOPs.

C. Experiment Settings

The official RetinaNet code was used for initial experiments and other codes were built upon MMDetection [63]. The ImageNet [64] pretrained model is used as the backbone. To make the model more applicable for tiny object detection, the P6 and P7 characteristic layers for detecting large objects have been eliminated, and the P3–P5 characteristic layers for detecting tiny and medium objects have been retained. This adjustment can save network storage space and improve detector operation speed without compromising the detector's performance for tiny object detection. The stochastic gradient descent [65] optimizer was utilized, and the initial learning rate was set at 0.1, with a momentum of 0.9. The total number of epochs was set to 12. Experiments were conducted on an NVIDIA GeForce RTX 3090 GPU with a batch size set to 4.

D. Implement Details

1) *Anchor Size*: Anchors with the original size [see Fig. 7(a)] and anchors with side length reduced by two times [see Fig. 7(b)] are used to verify the influence of anchor size on the performance of tiny object detection in the experiment. The original anchors have areas of 32^2 to 128^2 on feature maps P3–P5, respectively. At each feature map, we use anchors at three aspect ratios $\{1 : 2, 1 : 1, 2 : 1\}$ and add anchors of sizes $\{2^0, 2^{\frac{1}{3}}, 2^{\frac{2}{3}}\}$ of the original set of three aspect ratio anchors. While our adjusted anchors have areas of 16^2 to 64^2 on feature maps P3–P5 with the same aspect ratios and anchor sizes as the original anchors, respectively. The experimental results are given in Table I. Compared with the

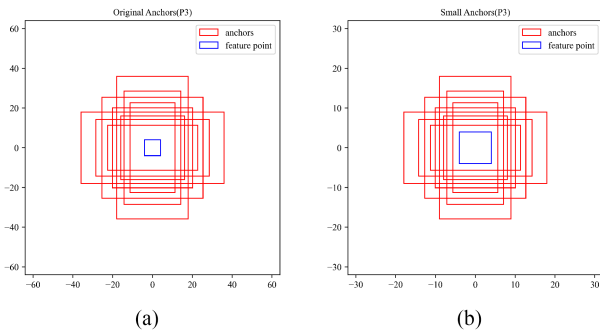


Fig. 7. Comparisons of feature point with anchors of different sizes on the P3 feature layer are shown. (a) Original anchors have areas of 32^2 on feature layer P3, with anchors at three aspect ratios $\{1:2, 1:1, 2:1\}$ and add anchors of sizes $\{2^0, 2^{\frac{1}{3}}, 2^{\frac{2}{3}}\}$ of the original set of three aspect ratio anchors. (b) Our adjusted anchors have areas of 16^2 on feature layer P3 with the same aspect ratios and anchor sizes as the original anchors.

TABLE II
EVALUATION RESULTS OF DIFFERENT BACKBONES ON AI-TOD DATASET

Backbone	Baseline	AP ₅₀	AP ₇₅	AP
ResNet-18	RetinaNet	29.3	5.6	10.9
	RetinaNet*	42.7	10.4	17.0
ResNet-50	RetinaNet	31.3	6.0	11.3
	RetinaNet*	<u>50.2</u>	<u>11.7</u>	<u>19.4</u>
ResNet-101	RetinaNet	34.0	6.9	13.0
	RetinaNet*	50.5	11.8	20.4

* The adaptive dynamic label assignment strategy proposed in this article is adopted.

The best indicators are highlighted in bold, and the suboptimal indicators are underlined.

set of standard-size anchors, reducing the size of the anchors to make the set of anchors more suitable for the GT of tiny object dataset can improve the AP₅₀ metric of the baseline model by 20.04%. Therefore, our subsequent ablation experiments all utilize subsize anchors.

2) *Backbone*: Our method is designed to be a plug-and-play solution, capable of enhancing the model’s detection performance for tiny objects using any backbone. Experimental results on various backbones are presented in Table II, with model’s trained using our method denoted by an asterisk (*). First, we confirm the compatibility of our method with diverse backbones. Second, the performance of the object detection network trained using our method improves with the enhancement of the backbone’s feature extraction capabilities. However, the performance improvement is not very pronounced when transitioning from ResNet50 [66] to ResNet101 as the backbone. Consequently, taking into consideration the storage space and inference time requirements for remote sensing object recognition tasks, we opt to use ResNet50 as the backbone in subsequent experiments.

E. Comparative Experiments

Comparative experiments have been carried out. Different baseline models have been used to train on the AI-TOD and Levir-Ship datasets and detection results have been obtained. Subsequently, under the same experimental conditions, these

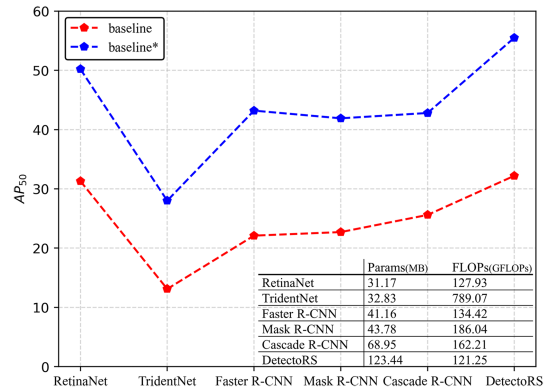


Fig. 8. Using our label assignment strategy (blue line) can improve the detection performance of the baseline model (red line) for tiny objects without increasing model parameters and inference time.

baseline models were retrained using the label assignment strategy proposed in this study, and experimental results were counted.

Main results on AI-TOD: Multiple different types of detectors were selected for conducting comparative experiments. The results are given in Table III, and models trained by using our proposed method are marked with an asterisk (*) for reference. The best indicators are highlighted in bold, and the suboptimal indicators are underlined. The experimental results on AI-TOD demonstrate that our method exhibits excellent compatibility. When applied to RetinaNet, Faster R-CNN, Cascade R-CNN, Mask R-CNN, TridentNet, and DetectoRS, it significantly enhanced the AP₅₀ metrics by 18.9%, 10.5%, 10.2%, 11.7%, 12.2%, and 23.3%, respectively. Fig. 8 visually demonstrates the variations of each metric. Second, other label assignment methods, such as ATSS [20] and AutoAssign [21], are also used for comparison with our approach. The experimental results demonstrate that our approach is both simpler and more effective. The models trained using our proposed label assignment strategy achieve higher AP₅₀ metric, with improvements of 33.0% and 19.3% compared with ATSS and AutoAssign, respectively. This superiority is attributed to the fact that ATSS and AutoAssign, as closely related comparison methods, overlook the uniqueness of tiny object samples, leading to suboptimal performance when compared with our proposed method. The effectiveness of our method has been empirically validated, and it outperforms existing label assignment strategies in improving the detection performance of tiny objects. In addition, employing DetectoRS as the baseline model, the AP₅₀ metric exhibits optimal performance, reaching 55.5%. In comparison with other state-of-the-art (SOTA) tiny object detection algorithms, our best AP₅₀ metric surpasses FSANet [45] and RFLA [23] by 2.7%, accompanied by improvements in other evaluation metrics. The experimental results indicate that our method has achieved SOTA performance. Fig. 9 showcases a selection of our detecting results on the AI-TOD benchmark.

Main results on Levir-Ship: To demonstrate the generalization of our method, besides the AI-TOD, the Levir-Ship dataset is also utilized. The experimental results on Levir-Ship show that

TABLE III
EVALUATION RESULTS OF DIFFERENT MODELS ON AI-TOD DATASET

Method	Backbone	AI	BR	ST	SH	SP	VE	PE	WM	AP ₅₀	AP ₇₅	AP _{vt}	AP _t	AP _s	AP _m
ATSS [20]	ResNet-50	0.5	9.8	16.8	26.4	0.6	12.0	4.0	1.2	22.5	8.9	2.1	10.1	11.4	15.0
AutoAssign [21]	ResNet-50	14.7	10.7	25.5	26.4	4.1	16.5	6.0	3.0	36.2	13.4	3.5	13.5	17.0	24.9
Dynamic R-CNN [65]	ResNet-50	10.4	15.5	30.7	38.7	7.0	19.5	9.4	3.9	39.4	16.9	6.3	18.2	21.2	25.6
SSD-512 [7]	VGG-16 [27]	10.9	2.1	10.7	14.3	1.8	9.2	2.2	0.9	21.0	6.5	1.0	5.3	11.9	21.2
YOLOv3 [66]	DarkNet-53	19.1	9.7	21.7	20.9	4.7	14.5	5.2	2.7	35.3	12.3	3.2	12.0	18.5	24.2
YOLOX [67]	DarkNet-53	20.2	7.6	28.8	24.0	5.0	22.3	6.8	1.4	37.2	14.5	3.9	14.1	20.6	24.2
FCOS [36]	ResNet-50	17.2	1.6	21.2	19.8	0.8	13.3	4.9	0.0	24.1	9.8	1.2	7.8	16.1	27.2
RepPoints [68]	ResNet-50	0.0	0.0	12.7	16.1	0.0	10.6	2.4	0.0	15.2	5.2	1.3	5.4	7.5	13.7
FoveaBox [69]	ResNet-50	14.7	0.0	19.2	17.7	0.0	12.4	4.2	0.0	20.6	8.5	0.9	6.1	14.1	27.2
CenterNet [35]	ResNet-18	12.8	2.0	15.3	15.9	1.7	11.7	4.6	1.5	24.3	8.2	1.3	6.9	12.9	22.3
M-CenterNet [59]	DLA-34	18.6	10.6	27.6	22.3	7.5	18.6	9.2	2.0	40.7	14.5	6.1	15.0	19.4	20.4
FSANet [45]	Swin-T	30.9	15.1	35.0	40.3	19.8	<u>24.9</u>	8.9	5.6	<u>52.8</u>	22.6	7.4	21.6	<u>29.1</u>	38.5
DetectoRS w/RFLA [23]	ResNet-50	<u>27.6</u>	16.8	37.6	45.9	15.5	25.0	<u>11.3</u>	5.7	<u>52.8</u>	<u>23.2</u>	<u>9.0</u>	<u>23.2</u>	28.7	36.4
Faster R-CNN [4]	ResNet-50	19.6	1.4	17.3	17.4	7.9	10.4	3.6	0.0	22.1	9.7	0.0	5.2	21.7	32.5
Faster R-CNN*	ResNet-50	14.8	16.1	28.8	35.2	9.2	19.7	8.8	4.9	43.2	17.2	7.8	17.5	21.7	27.4
Mask R-CNN [5]	ResNet-50	20.0	1.9	17.4	18.6	8.0	10.9	3.8	0.0	22.7	10.1	0.0	5.8	22.1	32.6
Mask R-CNN*	ResNet-50	15.8	14.5	28.9	38.3	5.3	19.1	8.8	4.6	41.9	16.9	7.0	16.9	21.1	28.8
Cascade R-CNN [6]	ResNet-50	21.0	6.7	20.3	21.0	7.7	12.8	4.8	0.0	25.6	11.8	0.1	7.9	23.0	34.8
Cascade R-CNN*	ResNet-50	15.1	17.5	29.2	40.2	7.8	20.5	8.8	4.5	42.8	18.0	7.8	18.2	22.3	29.7
TridentNet [9]	ResNet-50	7.6	0.0	10.8	9.6	2.5	7.0	1.9	0.0	13.1	4.9	0.0	2.8	10.8	19.2
TridentNet*	ResNet-50	8.2	1.5	17.6	29.8	3.0	15.7	5.2	1.6	28.0	10.3	3.6	9.5	14.6	21.0
RetinaNet [8]	ResNet-50	6.1	10.6	22.5	24.5	6.8	15.7	1.8	2.8	31.3	13.2	0.2	8.6	11.9	11.5
RetinaNet*	ResNet-50	11.9	<u>17.9</u>	27.7	<u>49.9</u>	14.7	21.7	5.3	<u>6.4</u>	50.2	19.4	8.1	19.0	23.3	22.4
DetectoRS [43]	ResNet-50	25.3	8.2	22.9	25.4	13.2	15.2	5.9	0.1	32.2	14.5	0.1	10.6	27.7	<u>37.7</u>
DetectoRS*	ResNet-50	26.0	20.1	<u>36.5</u>	57.1	<u>18.2</u>	24.0	11.3	7.2	55.5	25.0	13.5	25.1	29.3	35.7

* The adaptive dynamic label assignment strategy proposed in this article is adopted. The best indicators are highlighted in bold, and the suboptimal indicators are underlined.

our method exhibits good generalization on other datasets as well. When using Mask R-CNN as the baseline model, the AP₅₀ metric is improved by 3.2%, reaching 83.8% and achieving the SOTA performance. The results are given in Table IV.

Experimental results show that guiding the model to accurately differentiate between objects and background during training can effectively improve the detector's performance in tiny object detection. Compared with the methods that concentrate on enhancing the detection network structure and post-processing techniques, our proposed label assignment strategy fundamentally tackles the problem of learning confusion arising from improper label assignment. Furthermore, our approach demonstrates outstanding compatibility, enabling seamless integration with other optimization techniques.

F. Ablation Study

1) *Evaluation of Different Components*: The effectiveness of AT and D_{IoU} is verified separately. Compared with the label assignment strategy using fixed thresholds, using AT can improve the AP₅₀ metric by 14.5%, while the use of D_{IoU} ($\alpha = 0.6$, $\gamma = 1.5$) can increase the AP₅₀ metric by 5.5%. On this basis, by combining the use of AT and D_{IoU} , we achieved better tiny object detection performance than using either method alone, with an AP₅₀ score exceeding the baseline model by 19.1%. The effectiveness of our methods have been demonstrated. Table V gives the results of our experiment.

2) *Evaluation of md and D_{IoU}* : The adaptability of md and D_{IoU} in tiny object detection label assignment was independently

TABLE IV
EVALUATION RESULTS OF DIFFERENT MODELS ON LEVIR-SHIP DATASET

Method	Backbone	AP ₅₀	AP ₇₅	AP
Dynamic R-CNN	ResNet-50	77.9	10.1	28.5
YOLOv3	Darknet-53	72.7	7.3	25.1
ATSS	ResNet-50	63.8	4.1	19.9
AutoAssign	ResNet-50	80.6	11.2	28.4
FCOS	ResNet-50	69.7	3.6	21.6
FoveaBox	ResNet-50	73.9	8.6	26.2
Faster R-CNN	ResNet-50	66.7	4.4	20.5
Faster R-CNN*	ResNet-50	79.5	10.2	28.4
Mask R-CNN	ResNet-50	80.6	<u>12.5</u>	<u>31.1</u>
Mask R-CNN*	ResNet-50	83.8	14.7	32.7
Cascade R-CNN	ResNet-50	66.3	3.6	20.2
Cascade R-CNN*	ResNet-50	80.1	9.3	29.4
SSD-512	VGG-16	77.1	9.6	27.6
SSD-512*	VGG-16	<u>80.7</u>	9.3	28.5
TridentNet	ResNet-50	75.1	12.0	27.5
TridentNet*	ResNet-50	80.5	10.9	29.9
RetinaNet	ResNet-50	63.5	4.0	19.4
RetinaNet*	ResNet-50	77.5	7.3	25.7

* The adaptive dynamic label assignment strategy proposed in this article is adopted. The best indicators are highlighted in bold, and the suboptimal indicators are underlined.

verified. Leveraging AT as the basis, md and D_{IoU} were employed as dynamic label assignment metrics, and the experimental data are presented in Table VI. Obviously, when md is used,

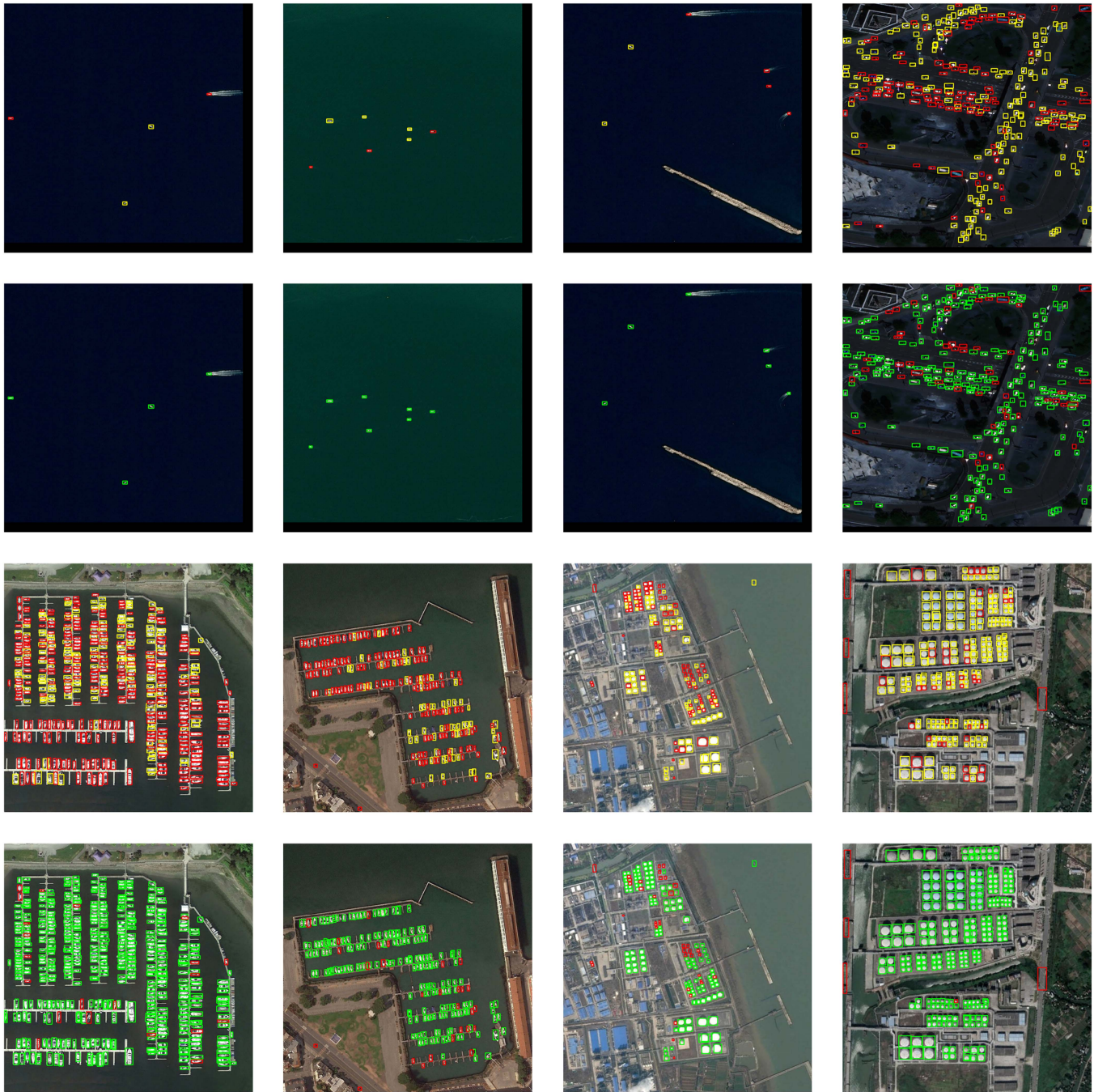


Fig. 9. Visualization of detection results by using different label assignment methods on the AI-TOD dataset. The yellow boxes (odd row) represent the TP detections by the model trained using a fixed threshold method. The green boxes (even row) represent the TP predictions by the model trained using our AT method. The red boxes represent the false negative predictions.

TABLE V
EFFECTIVENESS OF THE PROPOSED METHODS

Baseline	AT	D_{10U}	AP_{50}	AP_{75}	AP
RetinaNet	✓		31.3	6.0	11.3
		✓	45.8	11.6	19.0
	✓		36.8	7.0	13.9
		✓	50.2	11.7	19.4

The best indicators are highlighted in bold.

the model’s detection performance drops, exhibiting reverse optimization during training. Conversely, when D_{10U} is applied, the model’s detection performance for tiny objects is further improved. This experiment confirms that md is not suitable for label assignment in tiny object detection and underscores the enhanced robustness of our adjusted D_{10U} .

3) *Parameters in AT*: Guided by theoretical insights, we explored potential parameter combinations for calculating AT through a series of rigorous experiments. The experimental

TABLE VI
COMPARISON OF DIFFERENT DYNAMIC LABEL ASSIGNMENT INDICATORS ON AI-TOD

Method	md [24]	D_{IoU}	AP_{50}	AP_{75}	AP
RetinaNet+AT	✓	✓	45.8	11.6	19.0
			50.2	11.7	19.4

The best indicators are highlighted in bold.

TABLE VII
EFFECTS OF DIFFERENT AT CALCULATION PARAMETERS ON DETECTOR PERFORMANCE

P_{T_n}	P_{T_p}	AP_{50}	P_{T_n}	P_{T_p}	AP_{50}	P_{T_n}	P_{T_p}	AP_{50}
2/3	5/6	48.2	3/4	5/6	47.8	4/5	5/6	48.0
	4/5	48.4		4/5	49.1		4/5	49.3
	3/4	48.7		3/4	49.8		3/4	50.2
	2/3	48.6		2/3	49.3		2/3	49.4

The best indicators are highlighted in bold, and the suboptimal indicators are underlined.

results for different parameter combinations are presented in Table VII, where P_{T_p} represents the parameter in (1), and P_{T_n} represents the parameter in (2). The experiments clearly indicate that our method achieves optimal performance in tiny object detection when $P_{T_p} = 3/4$ and $P_{T_n} = 4/5$. Based on these stringent experimental results and for the sake of clarity and practicality, we have opted to employ fixed values for the parameters in (1) and (2).

4) *Hyperparameters in D_{IoU}* : The interplay of various hyperparameters and their influence on detector performance are investigated, so as to find the appropriate hyperparameters configuration. First, the way to determine β was investigated. In the training phase of the detector, α decays gradually from 1 to a set value, and the results are shown in Fig. 10. If a fixed β is used and set to fit the stabilized α , in the phase where α decays, β would cause the same problem as md due to the excessive impact of the penalty term on D_{IoU} caused by overweighting. Therefore, in order to adapt β well to α at each stage of training, the adjustment strategy that β changes dynamically according to α was adopted. Then, the relationship between hyperparameters and detector performance was experimented, where α balances the effect of feature alignment on label assignment, which increases as α decreases; γ reflects the adjustment effect of u on D_{IoU} , where u refers to the change of IoU before and after regression, and the larger the γ is, the smaller the effect of u on D_{IoU} . The specific experimental results are given in Table VIII. As α decays, the impact of feature alignment on D_{IoU} becomes more pronounced, resulting in an increase in AP_{50} , which indicates that using the indicator of adding R_{IoU} to assess anchor quality as the basis for label assignment is beneficial to improve detector performance, while as α decreases further, the effect of feature alignment further increases, AP_{50} decreases instead, suggesting that excessively high feature alignment weights damages the stability of training.

TABLE VIII
EFFECTS OF DIFFERENT HYPERPARAMETERS CONFIGURATIONS ON DETECTOR PERFORMANCE

γ	α	AP_{50}	γ	α	AP_{50}	γ	α	AP_{50}
1.5	0.2	48.9	2	0.2	47.6	3	0.2	47.0
	0.3	49.0		0.3	47.7		0.3	47.3
	0.4	49.1		0.4	48.5		0.4	47.9
	0.5	<u>49.6</u>		0.5	48.8		0.5	48.1
	0.6	50.2		0.6	49.4		0.6	48.2
	0.7	49.3		0.7	48.9		0.7	48.6
	0.8	49.0		0.8	48.7		0.8	48.5

The best indicators are highlighted in bold, and the suboptimal indicators are underlined.

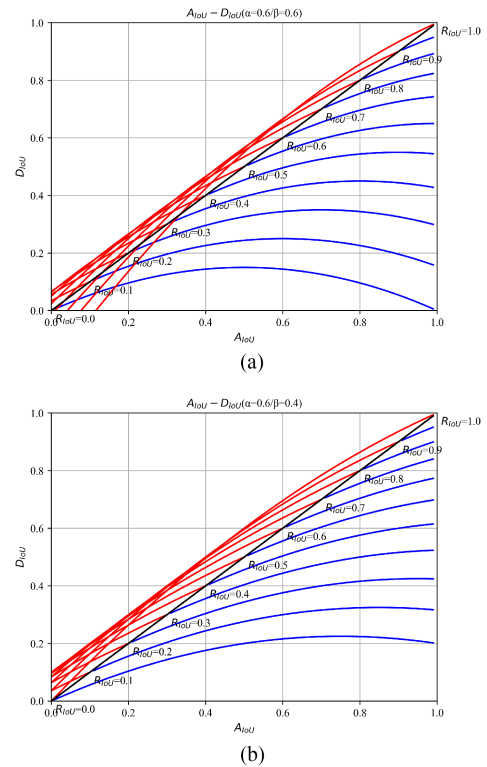


Fig. 10. This figure shows the relationship between β and D_{IoU} in the decay phase of α . We set $\alpha = 0.4$ in the stabilization phase, when $\beta = 0.6$. (a) β is constant, when α decays, β is too large, and the penalty term has an excessive impact on D_{IoU} , this makes the calculation result of D_{IoU} illogical. (b) β is dynamic, α and β have a good fit at different stages, and D_{IoU} meets theoretical needs.

V. CONCLUSION

In this article, the phenomenon that the IoU metric for tiny objects is highly susceptible to position deviations is thoroughly analyzed. This poses a challenge for the standard label assignment strategy in assigning appropriate positive samples to tiny object GT, thereby impacting the performance of the detector. To address this issue, a novel method called AT was proposed to determine the positive and negative sample thresholds for GT. This method enables the assignment of label assignment thresholds for each GT individually. In addition, the concept of dynamic IoU was introduced to resolve the issue of GT not receiving positive samples when $IoU_{max} \leq T_{pmin}$, thereby

mitigating sample isolation. Experimental results demonstrate that our proposed method significantly enhances the detection performance of tiny object without the need for increased feature layers or additional predefined anchors, and achieving SOTA performance on both the AI-TOD and Levir-Ship datasets. Furthermore, our method can be easily integrated as a plug-in into various models.

Moreover, the concept of adaptive threshold is not limited to the IoU-based models. As research progresses and more advanced evaluation metrics regarding anchor quality are investigated and proposed, these studies delve deeper into the potential of anchors, thereby positively influencing label assignment guidance. However, regardless of the chosen standard, the selection of thresholds remains an inevitable topic. Our approach can be applied to any anchor quality evaluation metric, making the process of setting label assignment thresholds more flexible and straightforward. This training-based adaptive concept holds promise for further promotion in future research.

REFERENCES

- [1] G. Chen et al., "A survey of the four pillars for small object detection: Multiscale representation, contextual information, super-resolution, and region proposal," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 52, no. 2, pp. 936–953, Feb. 2022.
- [2] X. Yu et al., "The 1st tiny object detection challenge: Methods and results," in *Proc. Comput. Vis. Workshops*, 2020, pp. 315–323.
- [3] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Comput. Vis.–ECCV 2014: 13th Eur. Conf. Part V 13*, Zurich, Switzerland, Sep. 6–12, 2014, pp. 740–755. [Online]. Available: <https://www.microsoft.com/en-us/research/publication/microsoft-coco-common-objects-in-context/>
- [4] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2016.
- [5] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2961–2969.
- [6] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6154–6162.
- [7] W. Liu et al., "SSD: Single shot multibox detector," in *Proc. Comput. Vis. 14th Euro. Conf.*, 2016, pp. 21–37.
- [8] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.
- [9] Y. Li, Y. Chen, N. Wang, and Z. Zhang, "Scale-aware trident networks for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6054–6063.
- [10] Z. Liu, G. Gao, L. Sun, and Z. Fang, "HRDNet: High-resolution detection network for small objects," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2021, pp. 1–6.
- [11] J. Pang, C. Li, J. Shi, Z. Xu, and H. Feng, " \mathcal{R}^2 -CNN: Fast tiny object detection in large-scale remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5512–5524, Aug. 2019.
- [12] Y. Gong, X. Yu, Y. Ding, X. Peng, J. Zhao, and Z. Han, "Effective fusion factor in FPN for tiny object detection," in *Proc. IEEE/CVF winter Conf. Appl. Comput. Vis.*, 2021, pp. 1160–1168.
- [13] J.-S. Lim et al., "Small object detection using context and attention," in *Proc. Int. Conf. Artif. Intell. Inf. Commun.*, 2021, pp. 181–186.
- [14] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, and S. Yan, "Perceptual generative adversarial networks for small object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1222–1230.
- [15] J. Rabbi, N. Ray, M. Schubert, S. Chowdhury, and D. Chao, "Small-object detection in remote sensing images with end-to-end edge-enhanced GAN and object detector network," *Remote Sens.*, vol. 12, no. 9, 2020, Art. no. 1432.
- [16] F. Zhang, L. Jiao, L. Li, F. Liu, and X. Liu, "Multiresolution attention extractor for small object detection," 2020, *arXiv:2006.05941*.
- [17] I. Goodfellow et al., "Generative adversarial networks," in *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.
- [18] J. R. Uijlings, K. E. A. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, pp. 154–171, 2013.
- [19] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-NMS - improving object detection with one line of code," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5562–5570.
- [20] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9759–9768.
- [21] B. Zhu et al., "Autoassign: Differentiable label assignment for dense object detection," 2020, *arXiv:2007.03496*.
- [22] J. Wang, C. Xu, W. Yang, and L. Yu, "A normalized Gaussian Wasserstein distance for tiny object detection," 2021, *arXiv:2110.13389*.
- [23] C. Xu, J. Wang, W. Yang, H. Yu, L. Yu, and G.-S. Xia, "RFLA: Gaussian receptive field based label assignment for tiny object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 526–543.
- [24] Q. Ming, Z. Zhou, L. Miao, H. Zhang, and L. Li, "Dynamic anchor learning for arbitrary-oriented object detection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, 2021, pp. 2355–2363.
- [25] Y. Lecun and L. Bottou, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
- [26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017.
- [27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014. [Online]. Available: <https://api.semanticscholar.org/CorpusID:14124313>
- [28] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.
- [29] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.
- [30] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 580–587.
- [31] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.
- [32] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.
- [33] L. Huang, Y. Yang, Y. Deng, and Y. Yu, "Densebox: Unifying landmark localization with end to end object detection," *Comput. Sci.*, 2015.
- [34] H. Law and J. Deng, "CornerNet: Detecting objects as paired keypoints," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 734–750.
- [35] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," 2019, *arXiv:1904.07850*.
- [36] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9627–9636.
- [37] X. Li et al., "FoveaNet: Perspective-aware urban scene parsing," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 784–792.
- [38] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–3010.
- [39] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Comput. Vis. 16th Euro. Conf.*, 2020, pp. 213–229.
- [40] F. Sung et al., "Learning to compare: Relation network for few-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1199–1208.
- [41] Z. Liu et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.
- [42] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2117–2125.
- [43] S. Qiao, L. C. Chen, and A. Yuille, "Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 10213–10224.
- [44] V. Chalavadi et al., "mSODANet: A network for multi-scale object detection in aerial images using hierarchical dilated convolutions," *Pattern Recognit.*, vol. 126, 2022, Art. no. 108548.

- [45] J. Wu, Z. Pan, B. Lei, and Y. Hu, "FSANet: Feature-and-spatial-aligned network for tiny object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–17, 2022.
- [46] S. Li, Q. Tong, X. Liu, Z. Cui, and X. Liu, "MA2-FPN for tiny object detection from remote sensing images," in *Proc. 15th Int. Congr. Image Signal Process., BioMedical Eng. Inform.*, 2022, pp. 1–8.
- [47] L. Zhang, M. Wang, Y. Jiang, D. Li, and Y. Zhou, "SSRDnet: Small object detection based on feature pyramid network," *IEEE Access*, vol. 11, pp. 96743–96752, 2023.
- [48] K. Jiang, Z. Wang, P. Yi, G. Wang, T. Lu, and J. Jiang, "Edge-enhanced GAN for remote sensing image superresolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5799–5812, Aug. 2019.
- [49] X. Wang et al., "ESRGAN: Enhanced super-resolution generative adversarial networks," in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2018.
- [50] J. Chen, K. Chen, H. Chen, Z. Zou, and Z. Shi, "A degraded reconstruction enhancement-based method for tiny ship detection in remote sensing images with a new large-scale dataset," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.
- [51] L. Li, X. Yao, X. Wang, D. Hong, G. Cheng, and J. Han, "Robust few-shot aerial image object detection via unbiased proposals filtration," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art no. 5617011.
- [52] C. Yang, Z. Huang, and N. Wang, "QueryDet: Cascaded sparse query for accelerating high-resolution small object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 13668–13677.
- [53] O. C. Koyun, R. K. Keser, I. B. Akkaya, and B. U. Töreyn, "focus-and-detect: A small object detection framework for aerial images," *Signal Process.: Image Commun.*, vol. 104, 2022, Art. no. 116675.
- [54] B. Bosquet, M. Mucientes, and V. M. Brea, "STDNet: A convnet for small target detection," in *Proc. BMVC*, 2018, p. 253.
- [55] P. Wang, X. Sun, W. Diao, and K. Fu, "Fmssd: Feature-merged single-shot detection for multiscale objects in large-scale remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3377–3390, May 2020.
- [56] Z. Ge, S. Liu, Z. Li, O. Yoshie, and J. Sun, "OTA: Optimal transport assignment for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 303–312.
- [57] S. Li, C. He, R. Li, and L. Zhang, "A dual weighting label assignment scheme for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 9387–9396.
- [58] T. Zhang, B. Luo, A. Sharda, and G. Wang, "Dynamic label assignment for object detection by combining predicted IoUs and anchor IoUs," *J. Imag.*, vol. 8, no. 7, p. 193, 2022.
- [59] X. Qian, Y. Huo, G. Cheng, C. Gao, X. Yao, and W. Wang, "Mining high-quality pseudo instance soft labels for weakly supervised object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art no. 5607615.
- [60] R. Fu et al., "Gaussian similarity-based adaptive dynamic label assignment for tiny object detection," *Neurocomputing*, vol. 543, 2023, Art. no. 126285.
- [61] J. Wang, W. Yang, H. Guo, R. Zhang, and G.-S. Xia, "Tiny object detection in aerial images," in *Proc. 25th Int. Conf. Pattern Recognit.*, 2021, pp. 3791–3798.
- [62] G. S. Xia et al., "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3974–3983.
- [63] K. Chen et al., "MMDetection: Open MMLab detection toolbox and benchmark, 2019, *arXiv:1906.07155*.
- [64] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.
- [65] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Statist.*, vol. 22, no. 3, pp. 400–407, 1951.
- [66] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
- [67] H. Zhang, H. Chang, B. Ma, N. Wang, and X. Chen, "Dynamic R-CNN: Towards high quality object detection via dynamic training," in *Proc. Comput. Vis. 16th Euro. Conf.*, 2020, pp. 260–275.
- [68] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767v1*.
- [69] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "Yolox: Exceeding YOLO series in 2021," 2021, *arXiv:2107.08430*.
- [70] Z. Yang, S. Liu, H. Hu, L. Wang, and S. Lin, "RepPoints: Point set representation for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9657–9666.
- [71] T. Kong, F. Sun, H. Liu, Y. Jiang, L. Li, and J. Shi, "FoveaBox: Beyond anchor-based object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 7389–7398, 2020.