remote sensing

MDPI

Article

Vision-Language Guided Semantic Diffusion Sampling for Small Object Detection in Remote Sensing Imagery

Jian Ma [1,*], Mingming Bian [1], Fan Fan [2], Hui Kuang [1], Lei Liu [1], Zhibing Wang [1], Ting Li [1] and Running Zhang [1]

[1] Institute of Remote Sensing Satellite, China Academy of Space Technology, Beijing 100048, China; bianmingming2008@163.com (M.B.); liting20011@sina.com (T.L.)
[2] China Academy of Space Technology, Beijing 100048, China
* Correspondence: majian0508@163.com; Tel.: +86-1860-105-0109

Abstract

Synthetic aperture radar (SAR), with its all-weather and all-day active imaging capability, has become indispensable for geoscientific analysis and socio-economic applications. Despite advances in deep learning–based object detection, the rapid and accurate detection of small objects in SAR imagery remains a major challenge due to their extremely limited pixel representation, blurred boundaries in dense distributions, and the imbalance of positive–negative samples during training. Recently, vision–language models such as Contrastive Language-Image Pre-Training (CLIP) have attracted widespread research interest for their powerful cross-modal semantic modeling capabilities. Nevertheless, their potential to guide precise localization and detection of small objects in SAR imagery has not yet been fully exploited. To overcome these limitations, we propose the CLIP-Driven Adaptive Tiny Object Detection Diffusion Network (CDATOD-Diff). This framework introduces a CLIP image–text encoding-guided dynamic sampling strategy that leverages cross-modal semantic priors to alleviate the scarcity of effective positive samples. Furthermore, a generative diffusion-based module reformulates the sampling process through iterative denoising, enhancing contextual awareness. To address regression instability, we design a Balanced Corner–IoU (BC-IoU) loss, which decouples corner localization from scale variation and reduces sensitivity to minor positional errors, thereby stabilizing bounding box predictions. Extensive experiments conducted on multiple SAR and optical remote sensing datasets demonstrate that CDATOD-Diff achieves state-of-the-art performance, delivering significant improvements in detection robustness and localization accuracy under challenging small-object scenarios with complex backgrounds and dense distributions.

Keywords: synthetic aperture radar (SAR); object detection; sampling optimization; diffusion model; vision–language model; bounding box regression

check for updates

Academic Editor: Dusan Gleich

Received: 6 July 2025
Revised: 28 August 2025
Accepted: 30 August 2025
Published: 17 September 2025

Citation: Ma, J.; Bian, M.; Fan, F.; Kuang, H.; Liu, L.; Wang, Z.; Li, T.; Zhang, R. Vision-Language Guided Semantic Diffusion Sampling for Small Object Detection in Remote Sensing Imagery. *Remote Sens.* **2025**, *17*, 3203. https://doi.org/10.3390/rs17183203

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

With the advantage of all-weather, all-day active imaging, synthetic aperture radar (SAR) has facilitated critical applications in geoscientific element perception and socio-economic research applications in geoscience element perception and socio-economic research [1]. The task of detecting objects in SAR remote sensing images is one of the essential techniques for analyzing and interpreting SAR remote sensing data. For SAR targets, the fundamental objective of detection frameworks lies in simultaneously achieving spatial localization and semantic categorization of visual entities within digital imagery. Current predominant detection techniques primarily focus on large-scale targets, where

deep learning methodologies can effectively obtain more intuitive feature representations for relatively large objects. In response to the distinctive characteristics of remote sensing image targets, existing studies have conducted algorithmic improvements from multiple perspectives [2,3]. Regarding the issue of target size distribution variation, detection networks necessitate multiscale feature processing. Conventional object detection frameworks employ feature pyramid networks (FPNs) as the neck network of detectors, allocating targets to distinct feature hierarchies for regression and detection. Multiscale detection in remote sensing primarily focuses on the fusion of multi-level information. By integrating hierarchical information from different pyramid levels, these approaches effectively combine low-level texture features with high-level semantic representations, thereby enabling effective identification of targets across various scales. Furthermore, attention mechanism-based multiscale detection has demonstrated promising performance in remote sensing applications, where spatial-channel attention modules enhance feature discriminability while scale-aware architectures optimize multi-receptive-field processing.

However, rapid and accurate detection of small targets remains a significant challenge due to difficulties in acquiring discriminative features and insufficient representational capacity. In contrast to natural images, minuscule targets in SAR imagery exhibit substantially smaller dimensions, typically spanning merely tens of pixels. The limited availability of texture, shape, and chromatic information leads to elevated missed detection rates in conventional detectors. In specific scenarios such as airport aprons, harbor docks, and parking facilities, small targets are frequently arranged in densely packed configurations. When processed through feature extraction pipelines, the corresponding feature maps of clustered small targets often manifest blurred boundary delineation, resulting in compromised localization precision during network prediction. Concurrently, the proximity of adjacent targets may induce bounding box overlap artifacts, potentially causing duplicate detections during inference. In contemporary deep learning frameworks, prevailing approaches to enhancing small target detection predominantly employ three strategic dimensions: data augmentation techniques, contextual information fusion mechanisms, and specialized training paradigms. Data augmentation methodologies typically leverage Generative Adversarial Networks (GANs) to synthesize enhanced small target details, where generative models produce photorealistic instances mimicking original data distributions, while discriminative models authenticate the synthetic samples' fidelity. The inherent limitations of small target samples—characterized by sparse pixel information and inadequate semantic representations in feature extraction—are exacerbated by progressive resolution reduction in feature maps during convolutional processing, inevitably inducing cascading information degradation. Consequently, strategic acquisition of supplementary small-target-related features emerges as an essential countermeasure against informational paucity. Optimized training configurations specifically tailored for small target detection—through refined anchor configuration, adaptive positive−negative sample assignment, and specialized loss function formulation—demonstrate significant precision improvements when systematically implemented within detector architectures.

Recent advances in vision–language models (VLMs), particularly Contrastive Language-Image Pre-training (CLIP) [4], have revolutionized object detection frameworks through semantic-aware feature learning. The inherent cross-modal alignment capability of CLIP enables detectors to leverage textual semantics for enhancing visual representation, which proves particularly effective in addressing open-vocabulary detection challenges. In remote sensing, pioneering works have adapted CLIP's zero-shot transfer capability to aircraft identification and land use classification, demonstrating improved generalization on low-data regimes [4–7]. Current research extends these advantages to detection tasks by the following: (1) utilizing semantic embeddings to refine region proposals in weakly super-

vised scenarios; (2) enabling query-based detection through natural language prompts; (3) mitigating domain shifts via joint image–text feature space regularization. However, the direct application of CLIP in small object detection remains underexplored, especially concerning the precision localization requirements and scale variance in SAR imagery.

Notwithstanding these advancements, persistent challenges remain in small target detection methodologies, fundamentally stemming from rigid modeling of sampling mechanisms. The minuscule dimensions of small targets create inherent matching incompatibilities within static sampling paradigms—either failing to establish effective correspondences with existing sampling points or achieving only marginal matches (as shown in Figure 1). Furthermore, conventional anchor sampling processes inadequately incorporate contextual priors, inducing severe positive–negative sample imbalance during training phases that substantially compromises network optimization efficacy. This dual deficiency in adaptive sampling coordination and prior knowledge integration fundamentally constrains the theoretical upper bound of current detection frameworks. This study aims to advance the adaptation of general-purpose detectors for small object detection from an architectural perspective. Specifically, we propose a vision–language model-guided dynamic positional prior to optimize the sampling process, named CDATOD-Diff (CLIP-Driven Adaptive Tiny Object Detection Diffusion Network), coupled with a redesigned regression loss function that enhances bounding box regression robustness. These methodological innovations collectively address two critical challenges in small object detection: severe sample imbalance and stringent localization accuracy requirements.

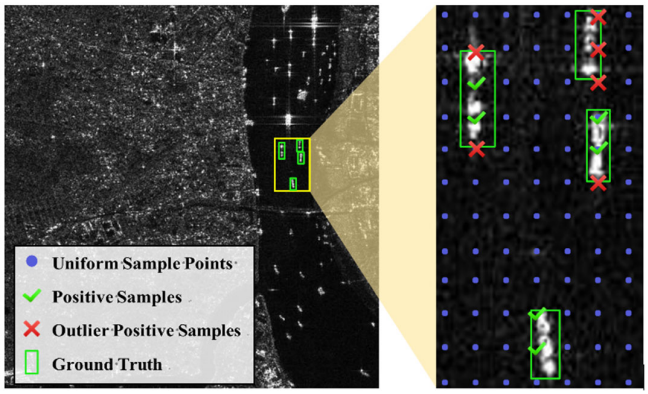

Figure 1. Schematic representation of the small target sampling problem.

The main contributions of this paper are as follows:

1. CDATOD-Diff, an integrated detection framework, is specifically designed for small target detection under scale-variation conditions, significantly improving detection robustness and localization accuracy where conventional architectures falter.
2. Innovative CLIP-driven dynamic anchor sampling strategies are introduced, solving the inadequate positive-sample issue under the guidance of CLIP in the sampling phase.
3. A robust adaptive regression loss function named BC-IoU is proposed for the sample regression phase, ameliorating the regression challenges encountered by small targets in specific scenarios and diminishing the repercussions of scale fluctuations.

2. Related Works

2.1. Object Detection

Compared with conventional algorithms, deep learning-based detection approaches overcome the limitations of handcrafted feature extraction in traditional detectors, demonstrating superior feature representation capabilities and enhanced performance. Contem-

porary neural network-based object detectors can be categorized into anchor-based and anchor-free paradigms based on sampling strategies. The anchor-based framework further divides into two-stage and single-stage detectors. Two-stage detectors generally achieve higher detection accuracy at the expense of computational speed, whereas single-stage detectors prioritize real-time performance with marginally reduced precision. In two-stage architectures, the seminal R-CNN framework [8] decomposes object detection into four sequential steps: region proposal generation via selective search, CNN-based feature extraction, SVM classification, and bounding box regression. The evolution continued with Faster R-CNN [9], which innovatively integrated a Region Proposal Network (RPN) to generate high-quality anchors through convolutional operations, employing softmax layers for foreground/background discrimination. Subsequent advancements include Mask R-CNN [10], which introduces a parallel segmentation branch with ROIAlign layers to mitigate quantization errors, and Cascade R-CNN [11], which progressively refines detection accuracy through multi-stage regressors with increasing IoU thresholds.

The fundamental distinction between single-stage and two-stage detectors lies in the presence of region proposal generation. Two-stage architectures (e.g., Faster R-CNN series) first generate sparse region proposals through dedicated subnetworks, subsequently refining these candidates via RoI-based operations. In contrast, single-stage detectors perform dense prediction directly on input images without intermediate proposal sampling. The YOLO framework [12] pioneered this paradigm by formulating detection as a unified regression task, simultaneously predicting bounding box coordinates and class probabilities through a single neural network pass. Subsequent YOLO variants [12–16] enhanced scale robustness through multiscale prediction layers and advanced backbone architectures.

While both single-stage and two-stage frameworks employ anchor boxes as predefined positional priors, their reliance on massive anchor populations exacerbates class imbalance during training, and the hyperparameter sensitivity (e.g., aspect ratios, scales) significantly impacts model robustness and inference efficiency. To address these limitations, CornerNet [17] pioneered an anchor-free detection paradigm by eliminating predefined anchors, instead predicting paired heatmaps for top-left and bottom-right corner points of bounding boxes, thereby reducing hyperparameter dependency. Subsequently, CenterNet [18] enhanced this approach by regressing object dimensions and orientations solely from center point predictions, establishing a one-to-one correspondence between targets and keypoints. In contrast, ExtremeNet [19] reformulated detection as extreme keypoint estimation, requiring simultaneous prediction of five geometric extremum points (topmost, bottommost, leftmost, rightmost, and center). RepPoints [20] employs adaptive point sets to delineate object geometries. FoveaBox [21] predicts category-specific semantic maps with anchor-like positional references. FCOS [22] introduced a simplified framework with centerness-based feature alignment, achieving state-of-the-art performance. Recently, diffusion models have recently been applied to detection tasks for candidate generation. For instance, DiffDet4SAR [23] models aircraft target detection as a bounding-box denoising process and employs a scattering feature enhancement module to suppress clutter. Different from this, our method integrates a Gaussian generative sampling process with an adaptive BC-IoU loss, enabling explicit anchor distribution modeling and improved robustness for small targets in complex scenes.

Compared to general object detection tasks, small objects are more pervasive and challenging in remote sensing imagery due to their extremely limited scale, dense distribution, and complex backgrounds. Our proposed CDATOD-Diff is specifically designed for small object detection in remote sensing scenarios. Our approach combines a generative sample point strategy—enabling semantically guided and scale-adaptive anchor generation—with a stability-enhanced regression mechanism. This dual design effectively alleviates the diffi-

culties posed by micro-scale targets and dense spatial arrangements, thereby significantly improving detection robustness and localization accuracy under scale variation conditions.

2.2. Tiny Object Detection

Augmenting contextual information acquisition has proven effective in mitigating the inherent limitations of insufficient feature representation for small objects. IONet [24] pioneered spatial recurrent neural networks to synergize intra-region and peripheral contextual cues within RoIs. PyramidBox [25] innovatively observed that facial targets exhibit co-occurrence dependencies with body parts, thereby developing an environment-sensitive prediction module that integrates the following: (a) low-level texture features, (b) context-aware data augmentation strategies, and (c) hierarchical context aggregation, achieving a 23.6% *AP* improvement on sub-20px targets in aerial imagery. CABNet [26] introduced a pyramid dilated convolution block that preserves spatial resolution while capturing multi-level semantics, coupled with reduced downsampling rates in backbone networks to retain fine-grained patterns. GCI-Net [27] leverages a Gaussian curvature-based branch and a complementary patch-group attention module to enhance detection through multi-scale feature aggregation and context modeling. The SAGCP [28] framework leverages the intrinsic sparsity of infrared small targets to efficiently remove redundant channels without complex auxiliary structures or metrics, significantly reducing model size and computation while maintaining high detection performance. Unlike these approaches, our framework focuses on diffusion-driven anchor generation and vision–language alignment, providing a complementary perspective for robust small target detection.

Current research has introduced targeted refinements to anchor-based mechanisms: FaceBoxes [29] proposed increasing anchor sampling density by reducing inter-anchor stride distances, thereby enhancing small object detection accuracy. Xu et al. [30] developed prior extraction blocks for rotating small objects, incorporating dynamic adjustments in label assignment and ground-truth representation to address mismatch errors. Wang et al. [31] established neural network-based anchor modeling approaches, both creating non-uniform and non-intensive positional priors. Sun et al. [32,33] abandoned predefined anchors, directly generating sparse region proposals through learnable candidate sets for classification and regression.

Label assignment design critically determines training sample quality in detection frameworks. While widely adopted, such static thresholding proves suboptimal for small objects due to their heightened IoU sensitivity, where semantically valid samples often reside in the 0.3–0.7 IoU range. Recent advancements address this limitation through four methodological innovations: S3FD [34] implements two-phase matching with a lower initial IoU threshold for candidate selection and Top-N ranking for unmatched ground truths. Zhang et al. [35] dynamically adjust assignment thresholds based on target statistical characteristics. Zhu et al. [36] replaces binary categorization with dual-weight assignments, balancing positive/negative contributions per sample. Xu et al. [37] proposes Gaussian receptive field-based matching with primary selection via distance-score-ranked top-K candidates and secondary refinement through decayed field radius.

Unlike these works that focus primarily on improving sampling density or fixed assignment criteria, our method introduces a CLIP-driven dynamic anchor sampling strategy. By leveraging semantic priors from the vision–language model, we overcome the insufficient positive sample issue, particularly for small objects with ambiguous or indistinct features.

2.3. Remote Sensing Image Processing Based on the Vision–Language Model

With the rapid advancement of deep learning, large-scale vision–language models (VLMs), such as CLIP, have emerged as a powerful paradigm integrating natural language processing and computer vision via contrastive learning. These models typically adopt a dual-tower architecture, comprising a visual encoder (e.g., ResNet [38], Vision Transformer [39]), and a text encoder (based on the Transformer architecture [40]), which are jointly optimized to align visual and textual representations in a shared embedding space. Current research efforts on VLMs are primarily centered on improving self-supervised training strategies, enhancing pre-training efficiency, and enabling few-shot transfer to downstream tasks. In the field of remote sensing, the application of VLMs is still in its early stages but shows considerable promise. Qiu et al. [41] leveraged a pre-trained CLIP model for feature extraction, achieving competitive scene classification performance with minimal supervision. Joufack Basso [42] explored cross-modal retrieval by integrating CLIP with a text-based remote sensing image retrieval system. Bazi et al. [43] introduced a visual question answering framework that combines image and textual features extracted by CLIP.

Sun et al. [44] proposed a mask-based visual foundation model that attained state-of-the-art results across eight remote sensing datasets on four downstream tasks. More recently, Chen et al. [45] incorporated the Segment Anything Model (SAM) [46] into a remote sensing foundation model to enhance instance segmentation performance.

Unlike prior works that utilize CLIP primarily for feature extraction or retrieval purposes, our approach incorporates CLIP-derived semantic embeddings directly into the generative anchor sampling process. Specifically, these embeddings serve as conditional priors in a diffusion-based sampling framework, enabling semantically aligned anchor generation and improving positive-sample quality for small object instances.

3. Methodology

This section presents the proposed CDAToD-Diff framework for small object detection in remote sensing, with the overall architecture illustrated in Figure 2. In Section 3.1, we first revisited the feature extraction process of the Contrastive Language-Image Pre-training (CLIP) method for both visual and textual data. Section 3.2 introduces the CLIP-guided diffusion-denoising sampling mechanism coupled with the positive–negative sample allocation strategy. Finally, Section 3.3 describes the proposed regression process for enhanced detection refinement.

3.1. CLIP Feature Extraction for Object Detection

The CLIP-based processing workflow for remote sensing imagery and its corresponding textual descriptions is depicted in Figure 3. The image and text data are processed through CLIP's visual encoder to generate fundamental visual representations. In the CLIP framework, image and text modalities are encoded by two separate transformer-based encoders, which are jointly trained to maximize the cosine similarity of matched image–text pairs while minimizing the similarity of unmatched pairs. This contrastive training paradigm allows CLIP to align heterogeneous modalities into a shared embedding space, enabling semantic comparisons between images and arbitrary textual prompts. Such a mechanism is particularly advantageous for object detection in remote sensing, where categories may be highly diverse and context-dependent. To fully leverage the CLIP-derived features, we implemented multiscale feature extraction across various network hierarchies rather than solely relying on the final fully connected layer outputs. These multi-hierarchical image features from the CLIP encoder were subsequently employed as conditional encodings for the diffusion-denoising sampling process. For input image

$I \in \mathbf{R}^{H\times W\times C}$, let E_{image} denote the vision encoder network and F_{image} denote the vision encoding feature. The hierarchical image encoding procedure can be formally expressed as:

$$I_{image} = E_{image}(I). \quad (1)$$

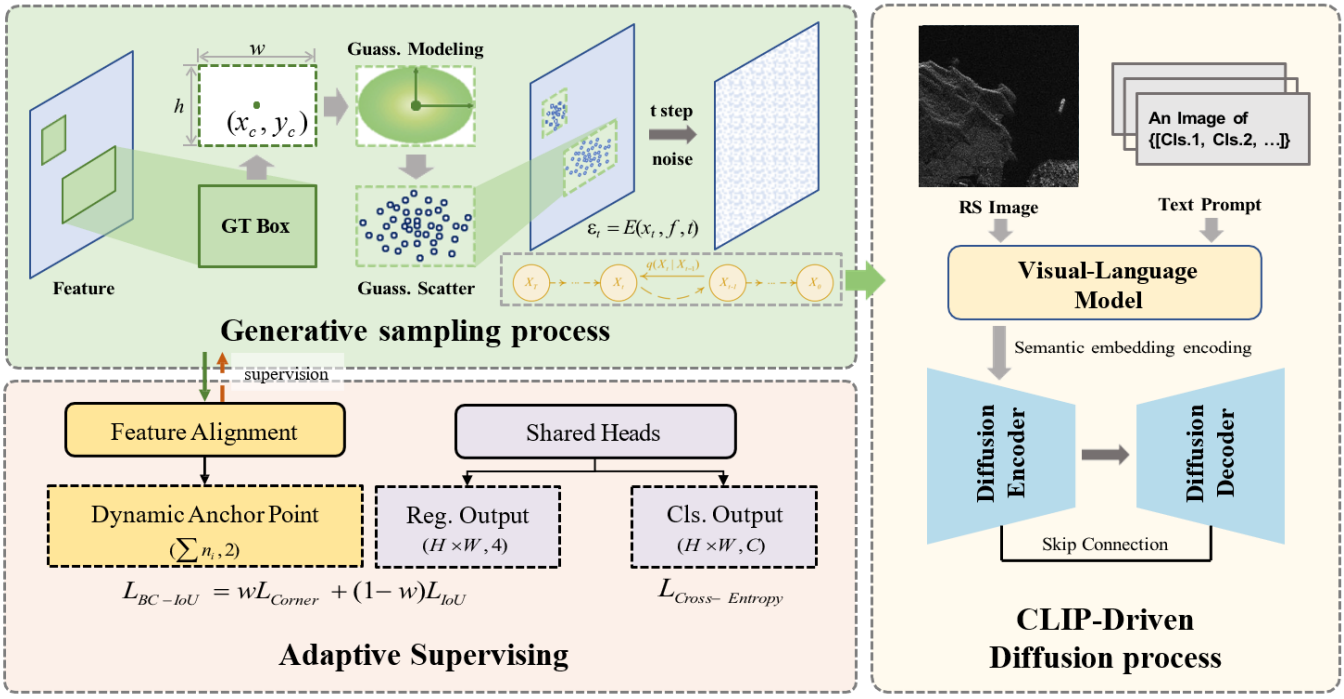

Figure 2. The illustration of the overall framework. The model consists of three main components: the first is generative sampling process part, where Gaussian modeling is applied to the ground-truth box to generate sample points, followed by a diffusion-based generative model to enrich training samples; the second is adaptive supervising part, where features are aligned and shared heads produce regression and classification outputs with the BC-IOU loss function for robust optimization; and the last one is CLIP-driven diffusion process part, where images and text prompts are embedded by a vision–language model and semantic features are refined through a diffusion encoder–decoder with skip connections, enhancing cross-modal representation learning.

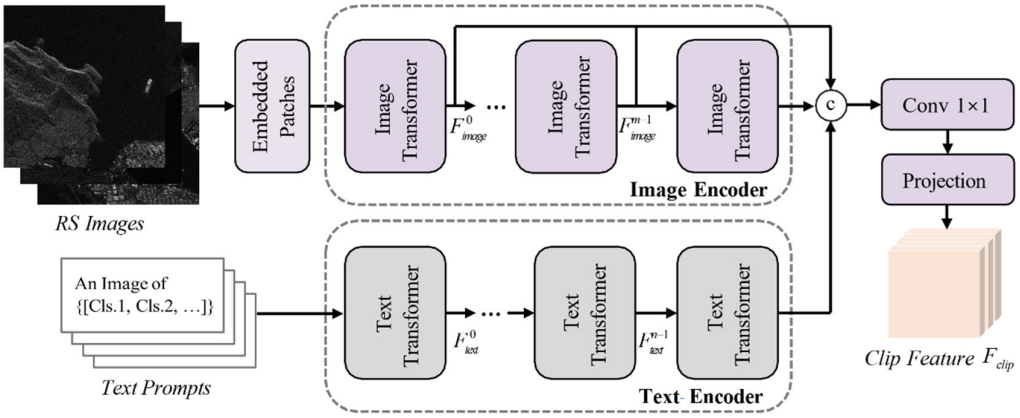

Figure 3. The illustration of the CLIP-driven feature extraction process. Images are first divided into embedded patches and then processed by a stack of image transformer blocks to obtain multi-layer visual features. Text prompts describing class categories are encoded by text transformer blocks to generate semantic representations. The encoded visual and textual features are then concatenated and projected through a convolution and a projection layer to produce the unified CLIP feature.

For the object detection task, a predefined textual input was employed to parameterize object categories. Specifically, the textual template T was configured as "an image of

[CLASS]" following CLIP's standard prompt engineering protocol, where the [CLASS] placeholder was dynamically populated with actual object category labels corresponding to the n_{cls} target classes in the detection task. This categorical parameterization enables semantic alignment between visual features and textual descriptors throughout the detection pipeline. Let E_{text} denote the text encoder network and F_{text} denote the text encoding feature. The text encoding procedure can be formally expressed as:

$$F_{text} = E_{text}(T). \tag{2}$$

3.2. CLIP-Driven Diffusion Anchor Point Sampling

3.2.1. Dynamic Anchor Point Sampling Procedure

In small object detection tasks, the network initially employs sampled points as positional priors, subsequently generating final prediction targets through optimization and regression processes. Addressing the characteristic requirement of anchor-free detection frameworks where object–anchor matching determines positive/negative training samples, our method introduces an adaptive sampling scheme using dynamic point selection mechanisms. This strategy effectively enhances the production of high-quality positive samples while suppressing outlier sample generation for small objects, thereby mitigating the prevalent challenge of imbalanced positive-sample distribution in small-scale target detection.

Let G denote the positive sample and M_d denote the dynamic matching process. The sample partitioning process can be expressed as:

$$G = M_d(F_d(P), GT), \tag{3}$$

where F_d denotes the sample generator with P as the input and GT denotes the ground truth.

To guarantee adequate positive-sample allocation for ground-truth bounding boxes during training, we implemented a hierarchical label assignment strategy building upon [37]. Following their paradigm of representing bounding boxes as two-dimensional Gaussian distributions, our approach extends this theoretical foundation through probabilistic modeling of spatial relationships between anchors and targets. Given receptive field $n_e = \mathcal{N}(\mu_e, \Sigma_e)$ and GT box $n_g = \mathcal{N}(\mu_g, \Sigma_g)$, the distance score WDS can be calculated based on the Wasserstein distance:

$$WDS = \frac{1}{1 + \left\| [x_n, y_n, e^{r_n}, e^{r_n}]^T - \left[x_g, y_g, \frac{w_g}{2}, \frac{h_g}{2} \right]^T \right\|_2^2} \tag{4}$$

The distance metric guides sample prioritization and label allocation, generating preliminary sampling output R_1 and mask m. Through progressive refinement, we strategically reduced the effective receptive field radius and reapply the ranking protocol, systematically augmenting each ground-truth instance with an additional positive sample to yield enhanced assignment R_2. The final sampling configuration emerges through strategic fusion of dual-phase outputs, formally expressed as:

$$R = R_1 m + R_2(1 - m) \tag{5}$$

Let s denote the sampling stride of the feature map with corresponding mapped coordinates (x, y). The updated sample points can be expressed as:

$$(x', y') = (x + \frac{s \cdot d_x}{2}, y + \frac{s \cdot d_y}{2}) \tag{6}$$

As demonstrated in Figure 2, our architecture incorporates an adaptive offset prediction module that dynamically estimates spatial displacements and updates sampling positions through learnable deformation fields. The module first processes the input feature tensor through a 3×3 convolutional layer to generate preliminary offset parameters for deformable convolution. Subsequently, these transformed features undergo spatial

adaptation via deformable convolution, where the kernel weights automatically adjust their sampling positions through parameterized coordinate perturbations $\Delta p = (d_x, d_y)$, ultimately producing the displacement vector that encodes optimized spatial offsets for target-aware feature sampling.

3.2.2. Diffusion-Based Anchor Point Sampling Procedure

Diffusion frameworks have emerged as prominent generative architectures in computer vision, distinguished by their stable optimization dynamics and exceptional distribution learning capacity. These models implement a stochastic denoising paradigm through Markov chain iterations, progressively injecting structured noise patterns during the forward phase (diffusion process), followed by learned iterative purification during the reverse phase (denoising process). For small object detection challenges, we adapted the framework to optimize anchor point generation through diffusion-based refinement, effectively enhancing the model's capacity to reconstruct critical sampling positions from noise-corrupted distributions. We modeled the annotated bounding boxes as Gaussian distributions and use a Gaussian sampler to perform random sampling, generating n sample points, the number of which is determined by the area of the ground-truth box. For timestep t, let α_t denote the predefined timestep noise schedule parameter ($\alpha_t \in (0, 1)$), and $\overline{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. The forward process for original sample x_0 can be formulated as:

$$x_t = \sqrt{\overline{\alpha}_t} x_0 + \epsilon \sqrt{1 - \overline{\alpha}_t}, \epsilon \in \mathcal{N}(0, I), \tag{7}$$

To optimize computational efficiency in processing noise-perturbed feature samples, our framework implements Denoising Diffusion Implicit Models (DDIMs) with hierarchical temporal refinement. This approach enables accelerated trajectory estimation through sequential approximation of latent distributions across temporal steps. For timestep p and predefined control parameter η ($\eta \in [0, 1]$):

$$\sigma^2 = \eta \cdot \frac{1 - \overline{\alpha_p}}{1 - \overline{\alpha_t}} (1 - \frac{\overline{\alpha_t}}{\overline{\alpha_p}}) \tag{8}$$

$$x_p = \sqrt{\overline{\alpha_p}} \left(\frac{x_t - \sqrt{1 - \overline{\alpha_t}} \epsilon_t}{\sqrt{\overline{\alpha_t}}} \right) + \sqrt{1 - \overline{\alpha_p} - \sigma^2} \epsilon_t + \sigma \epsilon \tag{9}$$

During the denoising stage, dynamically adjusted sampling points exhibit inconsistent spatial correspondence with the initial feature map. To harmonize these adaptive sampling locations with the feature map's perceptual regions, a spatial calibration module introduces learned positional adjustments. This mechanism first establishes regularly spaced sampling references and then calculates noise-reduction adjustments by measuring the positional variance between these predefined points and their denoised counterparts. These calculated deviations act as adjustment parameters that modify the feature map's spatial configuration through deformable convolution operations. The resulting geometrically refined feature representation is subsequently fed into the detection head to execute final coordinate prediction and category determination.

3.2.3. CLIP-Driven Conditional Encoding

To achieve deep semantic guidance in the conditional generation process of diffusion models, we implemented CLIP-encoded feature fusion for conditional semantic conditioning. The CLIP-derived prior encoding is integrated into the semantic encoding phase of the diffusion process to enhance sampling precision. Specifically, the textual prompt T is encoded through CLIP's text encoder, transforming descriptive text into 768-dimensional semantic vectors F_{text} that capture global textual semantics. For visual processing, the ViT-B/16 CLIP encoder decomposes input images into 16×16 patch sequences, extracting

hierarchical features $F_{image} = \left\{F_{clip}^1, F_{clip}^2, \ldots, F_{clip}^L\right\}$ through self-attention mechanisms, where $L = 12$ denotes the total transformer layers. Each layer's features maintained spatial dimensions of $H \times W \times D$ (e.g., $14 \times 14 \times 768$).

During the diffusion sampling iterations, we developed a dual-path CLIP guidance architecture that progressively incorporated both visual and textual encodings. The hierarchical CLIP features were sequentially injected into corresponding diffusion steps through residual connections. To address the modality gap between text and image feature distributions, we adopted an interactive modality-specific mapping module that generates text-visual prompts for cross-modal alignment. To implement cross-modal alignment, the text embedding F_{text} is first spatially replicated to match the resolution of the visual features, resulting in a text align feature $f_{ta} \in \mathbf{R}^{H \times W \times 768}$. Meanwhile, the visual features F_{image} are processed by a projection module composed of continuous multilayer perceptrons (MLPs), which generate image semantic representations $f_{is} \in \mathbf{R}^{H \times W \times d}$. These modality-specific representations are then fused through a conditional interaction mechanism defined as:

$$\mathbf{F}_{clip}^{fusion} = \sigma(W_{text} \odot f_{ta}) \oplus (W_{image} \odot f_{is}) \tag{10}$$

where σ denotes sigmoid activation, $\odot$ indicates element-wise multiplication, $\oplus$ indicates element-wise addition, and W_{text}/W_{image} represent learnable projection matrices. The U-Net encoder in diffusion models governs multiscale generation through its hierarchical structure, enabling fine-grained control over the synthesis process. To enhance the model's sensitivity to small objects and complex structural patterns—particularly relevant for detection tasks—we introduced a stage-adaptive injection mechanism that integrates CLIP-derived image features into the conditional encoding stream of the diffusion model. The fusion pipeline proceeds as follows.

Given the CLIP fusion representations $\left\{\mathbf{F}_{\text{clip}}^i\right\}_{i=1}^{L}$, where $\mathbf{F}_{\text{clip}}^i \in \mathbb{R}^{B \times C_{\text{clip}} \times H_i \times W_i}$, the spatial resolution must be aligned with that of the corresponding diffusion conditional encoding $\mathbf{F}_{\text{diff}}^l \in \mathbb{R}^{B \times C_l \times H_l \times W_l}$ to facilitate effective cross-modal fusion. The feature fusion pipeline comprises three principal stages of conditional adaptation. First, bilinear spatial interpolation is applied to align the spatial resolution of CLIP-encoded visual features $\mathbf{F}_{\text{clip}}$ with the diffusion conditional encoding features $\mathbf{F}_{\text{diff}}$:

$$\mathbf{F}_{\text{clip}}^{i \to l} \leftarrow \text{BilinearInterpolate}(\mathbf{F}_{\text{clip}}^i, (H_l, W_l)) \tag{11}$$

To ensure inter-modal compatibility in the fusion stage, we applied a bottleneck transformation to reduce the channel dimensionality while preserving spatial details. The transformation is defined as:

$$\mathbf{F}_{\text{clip}}^{i \to l} \leftarrow \text{BilinearInterpolate}(\mathbf{F}_{\text{clip}}^i, (H_l, W_l)) \tag{12}$$

where $\mathrm{C}_{k \times k}$ denotes convolution with kernel size k, progressively reducing channel dimensions from C_{in} to $C_{bottleneck}$ (default $C_{bottleneck} = C_{in}/4$), enabling efficient representation while retaining critical spatial information.

To achieve effective cross-modal feature fusion, we performed channel-wise concatenation between the CLIP-transformed image features F^{trans}.

The cross-modal feature fusion is implemented through channel-wise concatenation of conditional encoding features F_{diff} and CLIP-derived image features F^{trans}. This operation yields CLIP-guided conditional features that preserve complementary semantic information from both modalities:

$$\mathbf{F}_{CD} = \mathcal{C}_{1 \times 1}\left(\text{ReLU}\left(\left[\mathbf{F}_{clip}^{trans} \| \mathbf{F}_{diff}\right]\right)\right) \tag{13}$$

where $[\cdot\|\cdot]$ denotes concatenation along the channel dimension, followed by a 1×1 convolutional layer ($C_{1\times1}$) for feature recalibration.

3.3. Adaptive Loss

In the object detection process, factors that should be considered include the overlap area, distance, aspect ratio, and angle, among others. Accurately localizing small targets in remote sensing images poses a significant challenge. In the proposed adaptive loss function, we focused on two location-related key factors: overlap area and distance, which are collectively termed the BC-IoU loss.

In object detection, the IoU loss function is employed to measure the overlap area, reflecting the intersection between the predicted box and the ground-truth box. For a predicted box P and a ground-truth box G, the IoU loss can be formulated as:

$$L_{IOU} = 1 - \frac{P \cap G}{P \cup G} \tag{14}$$

To address the task of small object detection, the corner loss term L_{Corner} was designed to replace the central point distance loss, thereby circumventing the issue where the loss function becomes zero and cannot be optimized under conditions of central position overlap. L_{Corner} can be expressed as follows:

$$L_{Corner} = 1 - e^{-D/S}, \tag{15}$$

where S is the constant parameter that is related to the scale of the distance and D represents the distance between points.

For the central point distance, D_{cent} can be expressed as:

$$D_{cent} = \sqrt{(x_{1c} - x_{2c})^2 + (y_{1c} - y_{2c})^2}, \tag{16}$$

where (x_{1c}, y_{1c}) and (x_{2c}, y_{2c}) denote the center coordinates of the predicted box and ground-truth box, respectively. When the predicted center coincides with the ground-truth center, the loss value becomes zero, regardless of the width and height of the predicted box. In this case, the regression of box size receives no further supervision, which reduces the effectiveness of the central point distance loss in guiding localization. To overcome this limitation, we replaced the central point distance loss with a corner-based formulation, which considers the distances between the corners of the predicted and ground-truth boxes:

$$D_{corn} = \sum_{i=1,2} \sqrt{(x_{1i} - x_{2i})^2 + (y_{1i} - y_{2i})^2}, \tag{17}$$

where (x_{1i}, y_{1i}) and (x_{2i}, y_{2i}) denote the coordinates of the corresponding corners of the predicted and ground-truth boxes, where $i = 1, 2$ represents the top-left and bottom-right corners, respectively. The corner loss facilitates the optimization of spatial alignment for small object detection, particularly in scenarios where the targets are extremely small or the central points are overlapping.

The effectiveness of regression terms varies with object scale. For very small targets, IoU loss degenerates due to vanishing overlap, making the corner-based distance a more reliable supervisory signal. As object size increases, the IoU loss becomes more informative by encoding shape and overlap. Thus, we designed the weighting strategy that adaptively shifts from the corner loss to the IoU loss across scales, ensuring robust regression from small to large objects:

$$w = e^{-A/\beta}, \tag{18}$$

where A indicates the instance area and β indicates the scale factor of the area. As shown in Figure 4a, to ensure robustness across scales, the proposed weighting strategy adaptively

balances the two terms. The weight of the corner loss decreases as the target size increases. Further, Figure 5 analyzes the variation in loss values under different target scales. IoU loss is highly sensitive to small positional shifts in small targets but varies smoothly for large targets (Figure 5b). In contrast, the corner loss remains stable across scales (Figure 5a), making it more suitable for small instances. After applying the proposed scale-sensitive reweighting (Figure 5c,d), small targets retain more contribution from the corner loss, while large targets retain more from the IoU loss, enabling adaptive optimization during training.

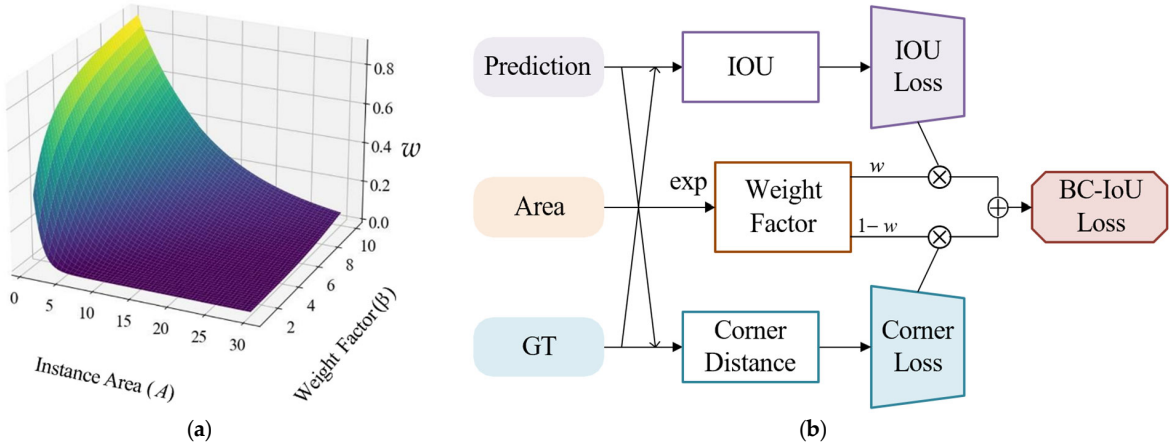

Figure 4. Proposed balance loss diagram. (**a**) The weighting factor w is adaptively determined by the instance area and width factor, emphasizing small targets. (**b**) The computation pipeline of the BC-IoU loss, where the weight factor is used to balance the IoU loss and corner distance loss, enabling more accurate localization of objects with varying scales and aspect ratios.

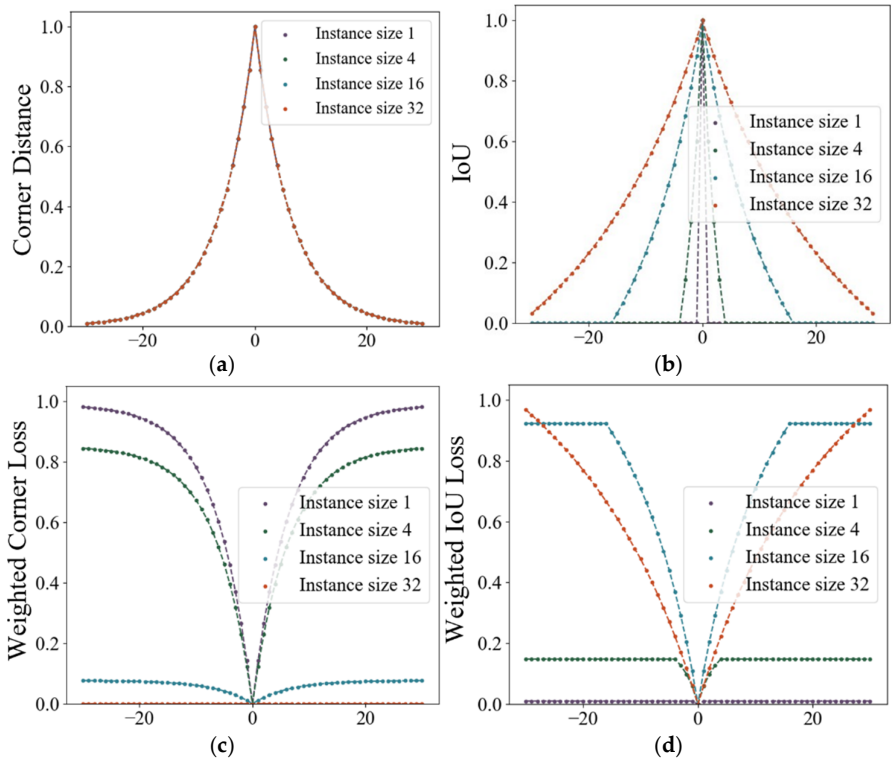

Figure 5. Loss change curves as the prediction box (green) shifts x pixels from the truth box (blue). (**a**) IoU change curve. (**b**) Curve of corner loss changing with the distance of the corner points. (**c**) IoU loss change curve. (**d**) Variation curve of corner loss with offset. The IoU loss is sensitive to small-target shifts while the corner loss is scale-stable, and the proposed scale-sensitive reweighting adaptively balances their contributions across target sizes.

Finally, Figure 4b presents the computation pipeline, where the weighted IoU and corner losses are summed to form the final regression objective. The complete BC-IoU loss is given as below, where both the IoU loss and the proposed corner loss are adaptively weighted according to the target scale:

$$L_{BC-IoU} = wL_{Corner} + (1 - w)L_{IoU}. \quad (19)$$

4. Results

4.1. Datasets

4.1.1. SAR Datasets

(1) MSAR-1.0 [47]: The MSAR-1.0 dataset includes 28,449 detection slices, using data from the Hise-1 and Gaofen-3 satellites. The polarization modes of the MSAR-1.0 dataset include HH, HV, VH, and VV. The dataset scenarios include airports, ports, nearshore, islands, offshore, urban areas, etc. The types of targets include aircraft, oil tanks, bridges, and ships, consisting of 1851 bridges, 39,858 ships, 12,319 oil tanks, and 6368 aircraft. We picked the images with pixel size of 256 × 256 and divided 80% images for the training and the rest for the test.

(2) HRSID [48]: The High-Resolution SAR Image Dataset (HRSID) is a dataset used for ship detection, semantic segmentation, and instance segmentation tasks in high-resolution SAR images. This dataset contains a total of 5604 high-resolution SAR images and 16951 ship instances. The HRSID dataset includes SAR images with different resolutions, polarizations, sea conditions, sea areas, and coastal ports. The resolutions of SAR images are as follows: 0.5 m, 1 m, and 3 m. We followed the division of dataset founders as 3642 images for the training and 1961 for the test.

4.1.2. Optical Datasets

(1) AI-TOD [49]: The AI-TOD benchmark contains 28,036 aerial images with 700,621 annotated instances, specifically designed for small target detection in remote sensing applications. This dataset features eight distinct object categories: airplanes, bridges, military tanks, maritime vessels, recreational swimming pools, transportation vehicles, individual persons, and renewable energy installations (wind turbines). Notably, the mean object dimension within AI-TOD measures 12.8 pixels in diameter, representing a substantial reduction in scale compared to conventional aerial detection datasets. This characteristic makes AI-TOD particularly valuable for evaluating models' capability in identifying sub-20-pixel targets that are prevalent in high-altitude observational scenarios.

(2) VEDAI [50]: The VEDAI benchmark constitutes a specialized resource for vehicular recognition in aerial surveillance systems, containing 1210 high-resolution (1024 × 1024 pixels) images with 3700 meticulously annotated instances. This dataset features fine-grained taxonomic classification across six vehicle types: recreational vehicles (campers), passenger cars (sedans), utility vehicles (pickups), agricultural machinery (tractors), commercial transports (trucks), and light commercial vehicles (vans). Unlike conventional remote sensing datasets focusing on macro-objects like industrial storage tanks or athletic facilities, VEDAI specifically addresses micro-scale targets with average spatial coverage below 0.05% of the total image area. The inherent challenges stem from vehicles' reduced spatial footprint (typically 32 × 32 pixels) and spectral similarity within complex urban backgrounds, establishing VEDAI as a critical evaluation platform for advanced pattern recognition in dense, cluttered environments.

(3) USOD [50]: Developed from the UNICORN2008 foundation, the USOD benchmark specializes in sub-pixel vehicle detection for aerial surveillance applications. Its source

imagery originates from electro-optical sensors with a 0.4 m spatial resolution, refined through a preprocessing pipeline involving spectral filtering, image segmentation, and expert verification to isolate vehicular targets. The curated collection contains 3000 annotated scenes encompassing 43,378 vehicular instances, divided into 70% training and 30% testing subsets through randomized stratification. Notably, 96.3% of targets fall within the micro-scale classification (16 × 16 pixels), cumulatively reaching 99.9% coverage in sub-32 × 32-pixel categories. This distribution establishes USOD as a challenging benchmark for evaluating detection algorithms in high-resolution surveillance scenarios. On this dataset, we performed quantitative evaluations and visual comparisons against the baseline methods.

4.2. Experimental Setup

The experimental implementation was developed using PyTorch 1.13.0 integrated with the MMDetection framework, executed on an Ubuntu system powered by an NVIDIA 3090 GPU (24 GB VRAM). Training employed stochastic gradient descent (SGD) optimization with an initial learning rate of 0.005, adjusted for hardware constraints through a reduced batch size of 2. The model underwent 12 training epochs, with the learning rate reduced to 0.0001 implemented at epochs 8 and 11. Built-in MMDetection preprocessing techniques were utilized, encompassing geometric transformations like random flipping, intensity normalization, and spatial padding operations to enhance data diversity.

The initial sampling architecture operates through two mechanisms: The offset sampling module dynamically adjusts sampling positions by applying learnable offsets to feature points on the two highest-resolution feature maps. Concurrently, the generative sampling module adopts a time-sequential approach with 1000 predefined time intervals, executing three progressive sampling iterations that yield three temporally ordered sampling pairs. For the robust regression component, the spatial scaling factor S is configured as 4 to modulate distance sensitivity, while the adaptive weighting parameter β is fixed at 12 to balance loss contributions. Remaining hyperparameters strictly follow the implementation specifications established in the FCOS baseline framework.

4.3. Experimental Metrics

To demonstrate the effectiveness of the proposed method, we employed average accuracy, specifically mean average precision (mAP), mAP at a 0.5 intersection over union (mAP@0.5), and mAP across a range of 0.5:0.95 (mAP@0.5:0.95) as evaluation metrics. The average precision (AP) is computed based on both precision p and recall rates r:

$$p = \frac{TP}{TP + FP}. \tag{20}$$

$$r = \frac{TP}{TP + FN}. \tag{21}$$

$$AP = \int_0^1 p(r)dr. \tag{22}$$

In the aforementioned equations, true positive (TP) is defined as the number of samples correctly identified as positive, false positive (FP) is defined as the number of samples incorrectly identified as positive from the negative class, false negative (FN) is defined as the number of samples incorrectly identified as negative from the positive class, and true negative (TN) is defined as the number of samples correctly identified as negative from the negative class. For multi-category detection tasks, the evaluation index mAP,

which is applicable to the complete dataset, can be obtained by averaging the AP values of all categories:

$$mAP = \frac{1}{n}\sum_{i=1}^{n} AP_i, \tag{23}$$

where n is the number of classes and AP_i is the AP at class i. According to the classification defined in the AI-TOD dataset, the Average Precision (AP) values for very small targets (less than 8 pixels in size), small targets (between 8 and 16 pixels), medium-small targets (between 16 and 32 pixels), and medium targets (larger than 32 pixels) are denoted as AP_{vt}, AP_t, AP_s, and AP_m, respectively. Targets larger than 32 pixels are uniformly categorized as medium targets for calculation purposes.

4.4. Comparison with State-of-the-Art Methods on SAR Datasets

This section analyzes the comparative performance of the proposed algorithm against other state-of-the-art algorithms on the MSAR-1.0 and HRSID datasets, with results tabulated for clarity.

4.4.1. Results on MSAR-1.0

As delineated in Table 1, our proposed method was compared against several state-of-the-art algorithms, encompassing both two-stage and one-stage approaches, as well as anchor-free methods. Among all contenders, our method demonstrated superior performance in terms of IOU-based AP, AP_{50}, and AP_{75} metrics, with particularly outstanding detection capabilities for small objects AP_s.

Table 1. Comparison experiments with the state of the art in MSAR-1.0.

Method	Backbone	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
Faster R-CNN [9]	ResNet50	43.0	65.6	49.2	36.5	64.6	56.9
TridentNet [51]	ResNet50	41.0	62.9	46.3	28.7	62.7	59.6
Sparse R-CNN [32]	ResNet50	42.2	65.0	46.4	33.4	62.2	57.6
SSD513 [52]	ResNet50	41.9	69.6	46.0	36.6	60.1	45.8
RetinaNet [53]	ResNet50	38.1	62.2	40.7	27.6	64.3	55.6
ATSS [35]	ResNet50	51.1	72.5	56.9	44.2	73.9	61.2
RepPoints [20]	ResNet50	44.4	69.0	50.3	34.7	67.2	60.6
AutoAssign [36]	ResNet50	52.5	76.4	59.8	47.5	71.1	64.3
Foveabox [21]	ResNet50	43.2	66.2	48.2	34.9	69.2	60.4
FCOS [22]	ResNet50	24.3	48.5	20.9	15.3	41.6	43.0
SASSDN [54]	CSPDarkner53	55.6	77.9	51.1	48.7	78.2	65.1
3SD-Net [27,55]	ResNet50	63.4	80.6	62.2	51.6	77.9	68.4
Proposed	ResNet50	64.1	83.8	68.9	58.9	80.9	71.9

4.4.2. Results on HRSID

In the HRSID dataset, we compared our proposed method with common mainstream detection algorithms and reported the detection results for ship targets. As shown in Table 2, our proposed method demonstrated desirable outcomes in detecting smaller-sized ships, achieving the best AP score of 0.397. In addition, the proposed method achieved the best performance with an AP of 0.698 for medium ships and the best AP of 0.507 for large ships, indicating that the method also performs well on objects of average size. The overall highest mean Average Precision (mAP) confirms that our proposed diffusion-denoising approach can achieve superior sampling and more robust bounding box regression on the SAR ship detection task.

Table 2. Comparison experiments with the state of the art in HRSID.

Method	Backbone	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
Faster R-CNN [9]	ResNet50	50.8	80.5	56.2	33.0	67.6	43.2
TridentNet [51]	ResNet50	48.5	78.4	53.0	28.7	67.1	48.4
Sparse R-CNN [32]	ResNet50	40.8	65.8	44.8	22.8	59.3	34.1
SSD513 [52]	ResNet50	44.6	76.9	47.0	25.8	63.0	26.5
RetinaNet [53]	ResNet50	41.6	70.1	44.7	18.6	63.6	33.3
ATSS [35]	ResNet50	50.2	81.2	55.0	31.3	67.9	39.6
RepPoints [20]	ResNet50	50.4	83.8	53.8	32.9	66.9	41.3
AutoAssign [36]	ResNet50	52.0	85.6	55.8	35.0	67.3	44.2
Foveabox [21]	ResNet50	44.2	75.3	47.2	23.0	64.1	36.3
FCOS [22]	ResNet50	44.7	75.1	48.0	24.6	64.0	35.1
SASSDN [54]	CSPDarkner53	48.6	78.7	56.7	37.3	66.9	46.9
3SD-Net [27,55]	ResNet50	51.9	81.3	60.8	36.9	67.7	48.6
Proposed	ResNet50	55.4	85.5	61.5	39.7	69.8	50.7

4.5. Comparison with State-of-the-Art Methods on Visible Datasets

This section analyzes the comparative performance of the proposed algorithm against other state-of-the-art algorithms on the AI-ToD, VEDAI, and USOD datasets, with results tabulated for clarity.

4.5.1. Results on AI-ToD

The detection results of the proposed method in the AI-TOD dataset and the USOD dataset are shown in Figure 6. Moreover, as delineated in Table 3, our proposed method was compared against several state-of-the-art algorithms, encompassing both two-stage and one-stage approaches, as well as anchor-free methods. Among all contenders, our method demonstrated superior performance in terms of IOU-based AP, AP_{50}, and AP_{75} metrics, with particularly outstanding detection capabilities for tiny objects AP_t. It also achieved the second-best performance in detecting very tiny and small-sized objects, thereby highlighting the efficacy of our proposed diffusion sampling-optimization scheme.



(**a**) (**b**)

Figure 6. The detection results of the proposed method in the AI-TOD dataset (**a**) and the USOD dataset (**b**).

Table 3. Comparison experiments with the state of the art on AI-TOD.

Method	Backbone	AP	AP_{50}	AP_{75}	AP_{vt}	AP_t	AP_s	AP_m
Faster R-CNN [9]	ResNet50	11.1	26.3	7.6	0.0	7.2	23.3	33.6
TridentNet [51]	ResNet50	7.5	20.9	3.6	1.0	5.8	12.6	14.0
DotD [56]	ResNet50	16.1	39.2	10.6	8.3	17.6	18.1	22.1
Sparse R-CNN [32]	ResNet50	7.2	19	4.1	3.5	8.4	7.2	7.3
SSD513 [52]	ResNet50	7.0	21.7	2.8	1.0	4.7	11.5	13.5
RetinaNet [53]	ResNet50	8.7	22.3	4.8	2.4	8.9	12.2	16.0
ATSS [35]	ResNet50	12.8	30.6	8.5	1.9	11.6	19.5	29.2
RepPoints [20]	ResNet50	9.2	23.6	5.3	2.5	9.2	12.9	14.4
AutoAssign [36]	ResNet50	12.2	32.0	6.8	3.4	13.7	16.0	19.1
Foveabox [21]	ResNet50	8.7	21.1	5.4	1.1	6.7	13.4	26.4
FCOS [22]	ResNet50	10.7	26.9	6.5	2.3	11	15.1	20.7
RFLA [37]	ResNet50	16.3	39.1	11.3	7.3	18.5	19.8	21.8
DiffusionDet [57]	ResNet50	11	30	5.7	4	10.7	14.3	19.1
Proposed	ResNet50	19.4	47	13	8.2	20.8	22.9	24.5

4.5.2. Results on VEDAI

In the VEDAI dataset, we compared our proposed method with common mainstream detection algorithms and reported the detection results for each subclass. As shown in Table 4, our proposed method demonstrated desirable outcomes in detecting smaller-sized objects, achieving AP scores of 0.458 for boats, 0.353 for tractors, and 0.495 for cars. In addition, the proposed method achieved the best performance with an AP of 0.444 for camping cars and the second-best AP of 0.359 for vans, indicating that the method also performs well on objects of average size. The overall highest mean Average Precision (mAP) confirms that our proposed diffusion-denoising approach can achieve superior sampling and more robust bounding box regression on the vehicle detection task.

Table 4. Comparison with the state of the art on VEDAI.

Method	BO	CP	CA	OT	PI	TR	TK	VA	mAP
FasterRCNN [9]	0.112	0.274	0.413	0.117	0.356	0.173	0.147	0.206	0.225
Retinanet [53]	0.176	0.365	0.398	0.113	0.33	0.197	0.207	0.144	0.241
ATSS [35]	0.315	0.371	0.445	0.185	0.417	0.274	0.245	0.348	0.315
RFLA [37]	0.22	0.402	0.413	0.087	0.422	0.326	0.197	0.355	0.302
Align DETR [58]	0.22	0.402	0.413	0.087	0.422	0.326	0.197	0.355	0.302
Diffusiondet [57]	0.433	0.427	0.442	0.177	0.443	0.317	0.175	0.298	0.340
Proposed	0.458	0.444	0.495	0.162	0.457	0.353	0.193	0.359	0.365

4.6. Ablation Study

We performed ablation experiments on three public benchmark datasets to prove the effects of the proposed components. To comprehensively evaluate the contribution of each module within our detection framework, we selected ResNet-50 as the common baseline model. Each design was intended to systematically assess the impact and efficacy of the various modules within the detector. The results of the ablation experiments are displayed in Table 5. Comparison of the loss convergence process on the USOD dataset is shown in Figure 7. Our method demonstrates a significant advantage in convergence speed.

Table 5. Ablation study on anchor point sampling methods for object detection.

Dataset	Method	AP	AP_{50}	AP_{75}	AP_{vt}	AP_t	AP_s	AP_m
AI-ToD	FCOS	0.107	0.269	0.065	0.023	0.11	0.151	0.207
	FCOS + RFLA	0.163	0.391	0.113	0.073	0.185	0.198	0.218
	FCOS + Diffusion	0.179	0.448	0.102	0.061	0.167	0.246	0.317
	FCOS + Diff-CLIP	0.194	0.47	0.13	0.082	0.208	0.229	0.245
USOD	FCOS	0.106	0.268	0.063	0.02	0.103	0.14	0.208
	FCOS + RFLA	0.192	0.613	0.05	0.087	0.206	0.302	0.182
	FCOS + Diffusion	0.232	0.712	0.062	0.118	0.247	0.281	0.320
	FCOS + Diff-CLIP	0.246	0.732	0.077	0.117	0.264	0.296	0.245
VEDAI	FCOS	0.017	0.086	0.003	-	-	0.012	0.055
	FCOS + RFLA	0.324	0.585	0.310	-	-	0.311	0.342
	FCOS + Diffusion	0.313	0.600	0.284	-	-	0.328	0.332
	FCOS + Diff-CLIP	0.365	0.683	0.342	-	-	0.345	0.388

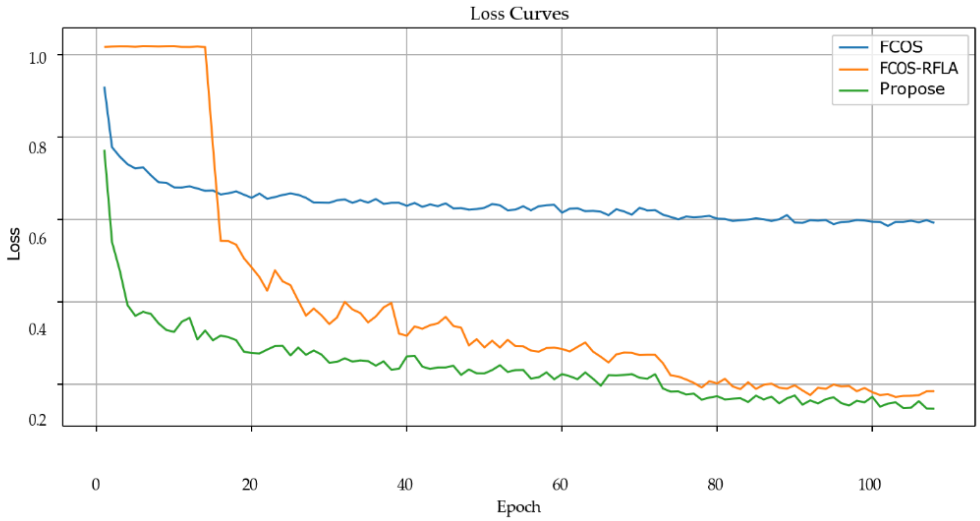

Figure 7. Comparison of the loss convergence process on the USOD dataset.

4.6.1. Baseline Setup

In this study, we adopted FCOS as our baseline algorithm, leveraging the ResNet50 architecture as its backbone network. To enhance the detection capabilities for small-scale targets, we selected the P2 feature layer from the Feature Pyramid Network (FPN), known for its higher resolution and detailed information, as the starting point for our detection layers. This choice is strategic for capturing fine-grained details essential for small object detection. Furthermore, we refined the sampling point matching through the application of a hierarchical sampling strategy, which is crucial for aligning the detection process with the distribution of small targets. For bounding box regression, we opted for the Generalized Intersection over Union (GIOU) loss function, which provides a more comprehensive measure of the overlap between predicted and ground-truth bounding boxes. A comparative analysis of the performance between the original and the optimized algorithms is presented in Table 5, demonstrating the optimized algorithm's improved performance and establishing it as a new benchmark for subsequent studies.

4.6.2. Ablation Study of the CLIP-Driven Diffusion Sampling Module

We systematically conducted comparative experiments to evaluate the performance of conventional sampling, diffusion-based generative sampling, and CLIP-driven diffusion-based generative sampling in object detection tasks. The diffusion-denoising model-based

generative sampling module further enhances detection performance, as evidenced by the overall mean Average Precision (mAP) metric: the generative sampling module achieved scores of 19.4, 24.6, and 36.5 on three benchmarks, respectively. During the detection of very tiny and tiny-sized targets in the AI-ToD dataset, the CLIP-driven diffusion-based generative sampling outperformed the diffusion-based generative sampling by 2.1 and 4.1, respectively, as illustrated in Table 5. On the remaining datasets, the proposed method also achieves the best performance compared to its ablated variants. On the USOD dataset, the baseline FCOS yields a 0.106 AP. With RFLA, the AP rises to 0.192, and diffusion further improves it to 0.233, confirming the benefits of enhanced feature learning and representation. Our proposed Diff-CLIP achieves the best result of a 0.246 AP, demonstrating superior robustness and effectiveness on this difficult dataset. On VEDAI, our method achieves substantial improvements across all metrics. These data significantly validate the effectiveness and superiority of the CLIP-driven diffusion-based generative sampling approach in the field of small object detection. This indicates that the generative sampling module, based on the denoising diffusion model, generates more robust sampling points, effectively mitigating the issue of sample imbalance.

4.6.3. Ablation Study of Each Loss Item

In this experimental section, we report validation results of the contributions of the proposed corner loss and IoU loss on the AI-ToD dataset. We tested the corner loss, IoU loss, and the proposed combined loss function. It is evident from Table 6 that using individual loss function terms leads to lower results. The proposed combined loss outperformed the individual corner loss and IoU loss by 1.9 and 1.3 in the AP metric, respectively, 3.7 and 2.8 in AP50, respectively, and 0.6 and 0.2 in AP75, respectively. This indicates that the BC-IoU loss demonstrates superior detection performance and can address the challenges of loss regression under specific relative positions.

Table 6. Ablation study on components of the BC-IoU loss function on the AI-ToD dataset.

IoU	Corner	AP	AP_{50}	AP_{75}	AP_{vt}	AP_t	AP_s	AP_m
✓		10.6	26.8	6.3	2	10.3	14	20.8
	✓	11.2	27.7	6.6	2.4	11.1	15.7	21.9
✓	✓	12.5	30.5	8	2.6	11.7	17	24.9

4.6.4. Effectiveness of Balanced Loss Function

To validate the effectiveness of the corner–IoU balanced loss, we compared the performance of the balanced loss with that of individual losses in Figure 8. The experimental results demonstrate that the proposed balanced loss algorithm achieves superior performance compared to individual losses, effectively mitigating the regression difficulties encountered in specific relative positioning scenarios. To further substantiate its validity, the proposed loss function was compared with other IoU-based loss functions, including GIoU, DIoU, CIoU, EIoU, and SIoU. These loss regression methods are constructed based on overlap-ping area, distance, width-height, and angle. In comparison, the balanced corner–IoU loss achieved optimal results across most metrics.

4.7. Discussion

As shown in Figure 6, an intuitive visualization of the proposed method is presented on three public benchmarks. Furthermore, we utilized Grad-CAM to visualize the attention of the target network. As illustrated in Figure 9, the proposed method not only achieves fewer false detections and missed detections but also assigns higher attention weights to genuine small targets, which is conducive to refining the detection outcomes.

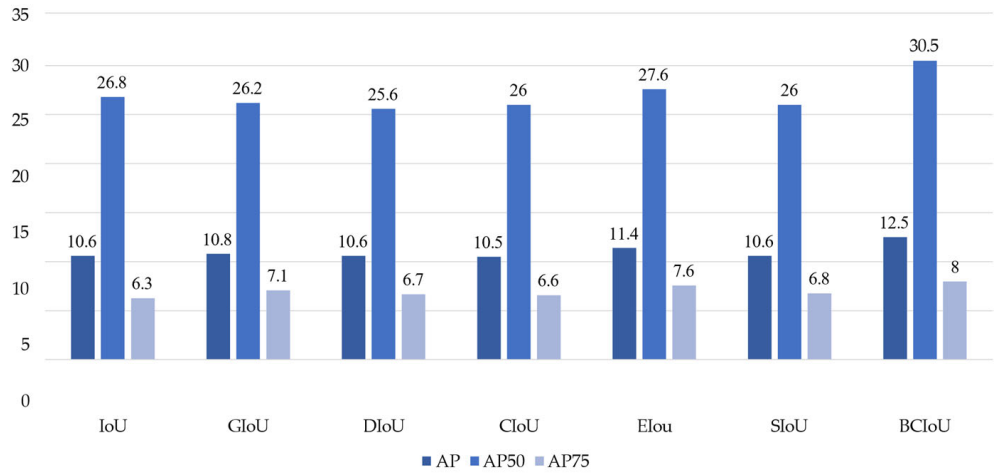

Figure 8. Comparison of detection results on the AI-ToD dataset using different loss functions.

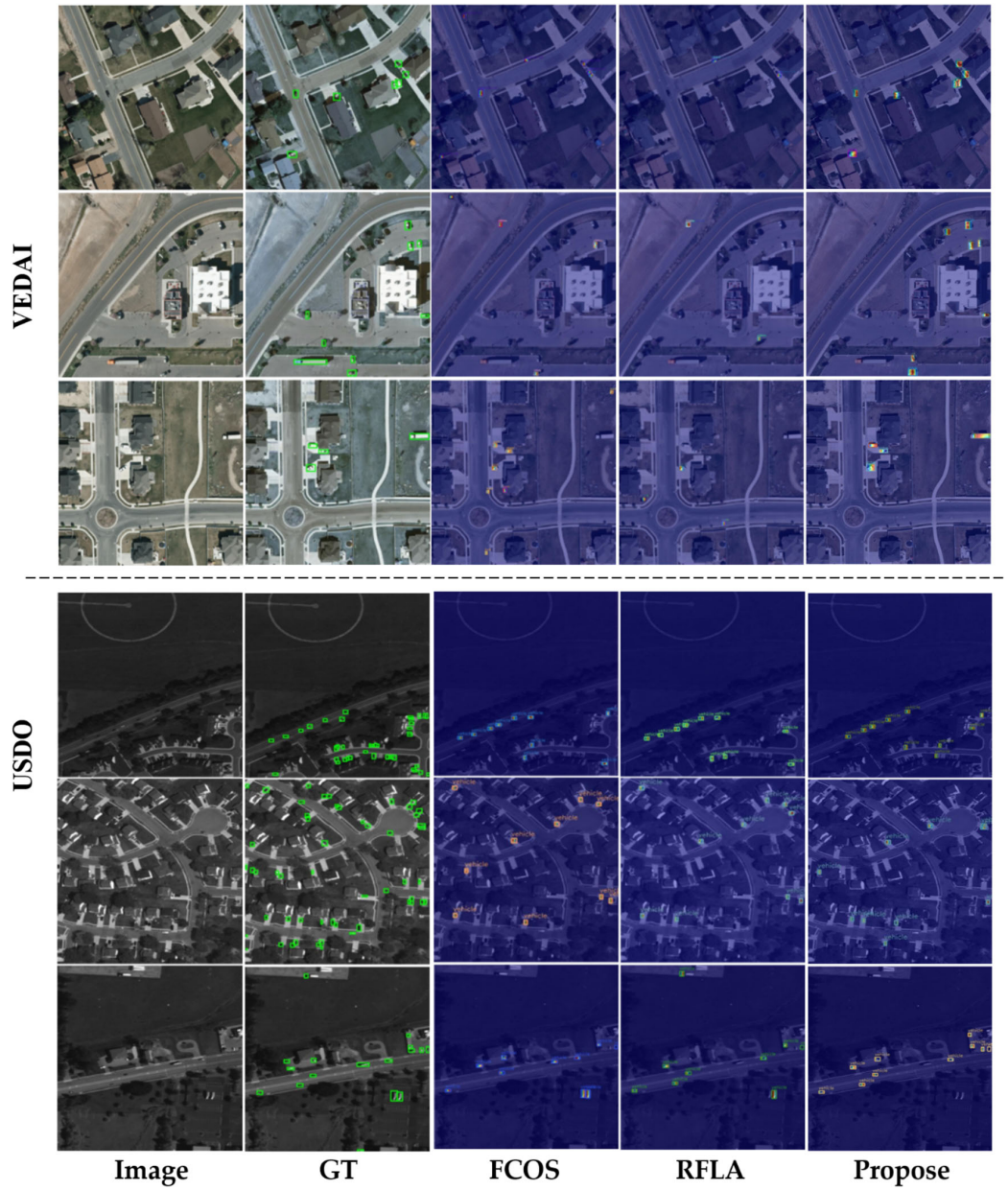

Figure 9. Comparison of class activation maps for detection results.

We compared the proposed method with the baseline method on these three benchmarks and displayed the results in Figures 10–13. We visualized and compared the detection results of remote sensing targets, examining the baseline network and the optimized diffusion sampling detection network. It can be observed that the baseline method has a relatively high false alarm rate in vehicle target detection, especially when there are interfering targets in the environment. In contrast, the proposed method effectively reduces false alarms. Specifically, in Figure 10 (VEDAI), our method demonstrates better robustness in detecting small vehicles within cluttered backgrounds, where the baselines either miss true objects (red boxes) or produce redundant predictions (blue boxes). In Figure 11 (AI-TOD), the proposed approach significantly improves the detection of densely distributed targets, yielding more complete coverage with fewer missed instances. In Figure 12 (USOD), our method exhibits superior capability in handling extremely small and low-contrast objects, where FCOS tends to produce fragmented or inaccurate bounding boxes. Overall, these visualizations highlight the strength of our approach in enhancing both detection accuracy and localization robustness across different datasets and challenging conditions. Figure 13 shows qualitative results on the HRSID dataset. Compared with FCOS and RFLA, the proposed method achieves more accurate detections of ships under complex backgrounds in SAR imagery. Specifically, FCOS tends to generate false positives around strong background scatterers (brown circles), and RFLA still misses several true ships (red boxes). In contrast, our method effectively suppresses background interference and yields more complete detections, with bounding boxes that better align with the true ship shapes. These results demonstrate that the proposed loss function and adaptive sampling strategy enhance robustness against clutter and improve detection performance in challenging SAR scenes.

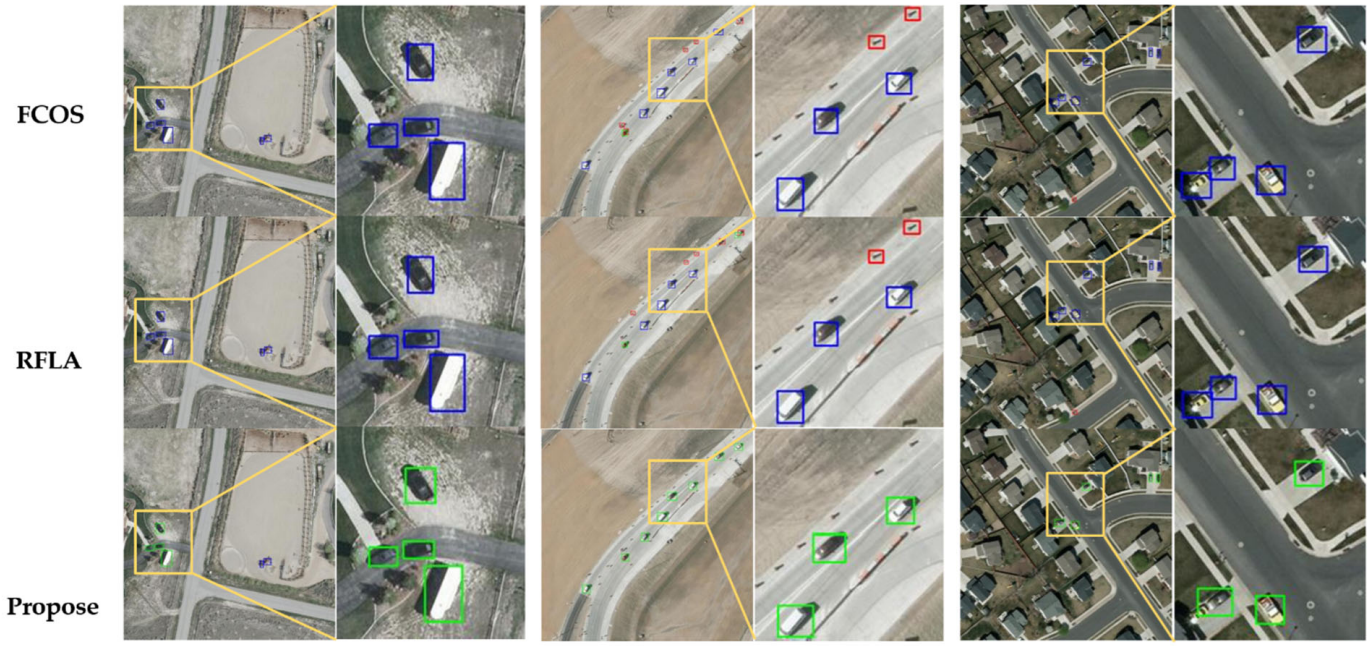

Figure 10. Comparison between detection results in the VEDAI dataset. The green, blue, and red boxes denote true positive (TP), false positive (FP), and false negative (FN) predictions, respectively.

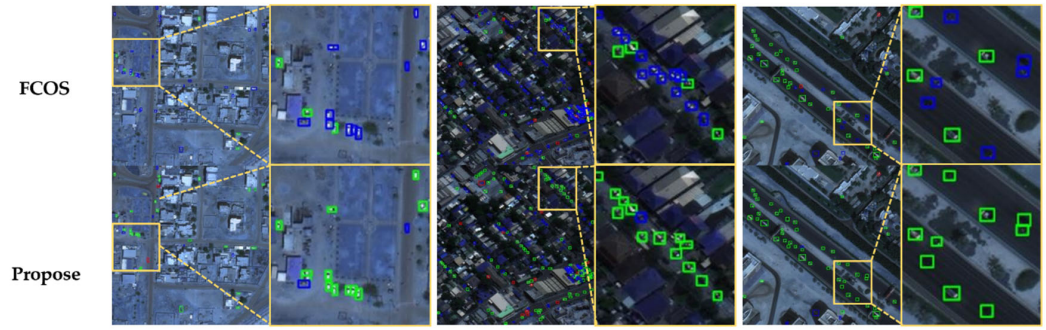

Figure 11. Comparison between detection results in the AI-ToD dataset. The green, blue, and red boxes denote true positive (TP), false positive (FP), and false negative (FN) predictions, respectively.

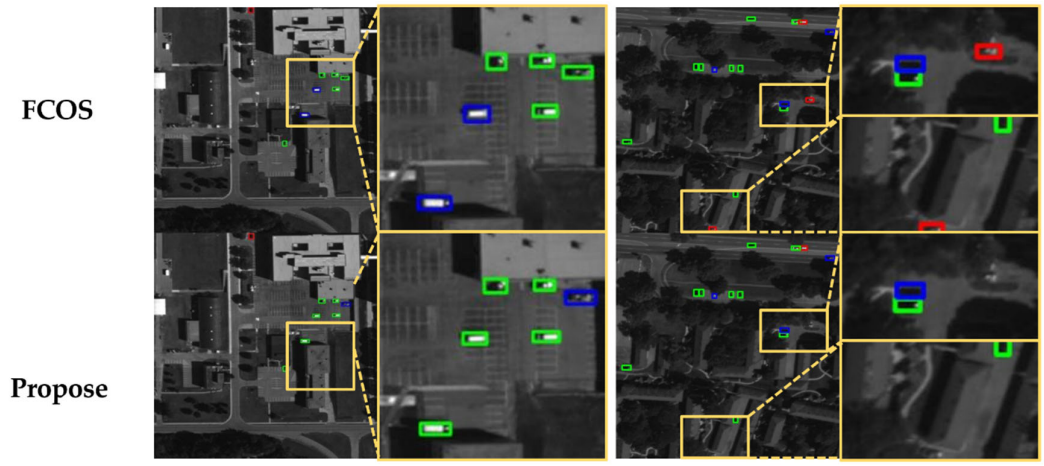

Figure 12. Comparison between detection results in the USOD dataset. The green, blue, and red boxes denote true positive (TP), false positive (FP), and false negative (FN) predictions, respectively.

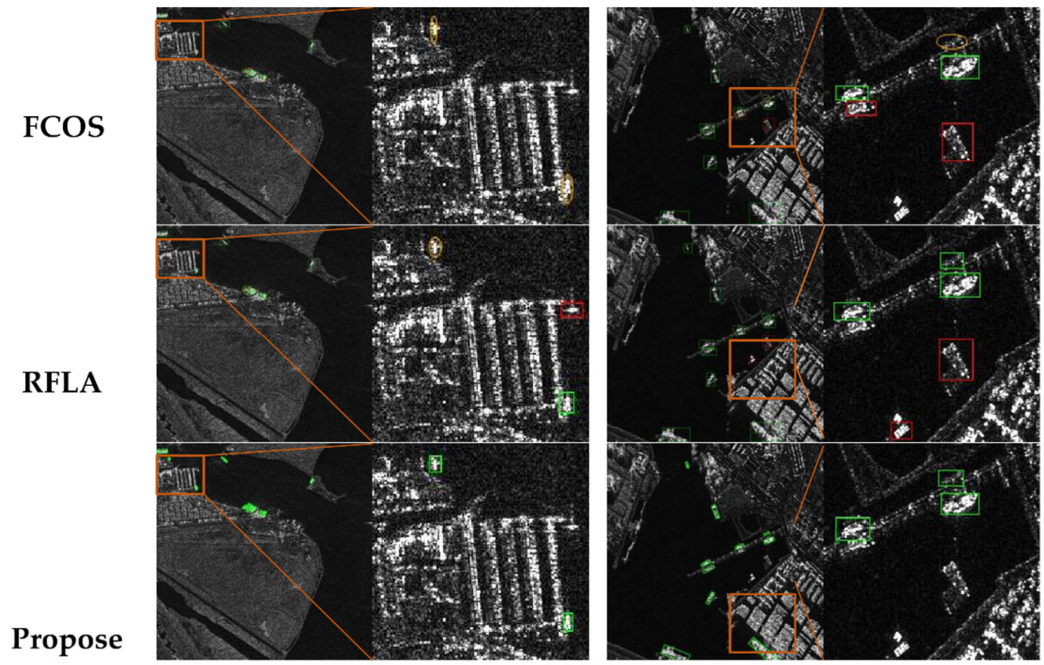

Figure 13. Comparison between detection results in the HRSID dataset. The green box, brown circle, and red box denote true positive (TP), false positive (FP), and false negative (FN) predictions, respectively.

5. Conclusions

This study systematically investigates algorithm design for small object detection in remote sensing imagery, addressing the challenges of sample imbalance during training and the difficulties in bounding box regression when boxes are in special relative positions. To tackle these issues, we propose a small object detection framework, CDATOD-Diff, which integrates sampling optimization and adaptive regression refinement. Building upon the anchor-free, single-stage FCOS algorithm as our baseline, we introduce a CLIP-driven generative sampling module that leverages vision–language models to bias sampling points toward genuine targets while increasing the quantity of small object samples, thereby enhancing detection performance. By designing a corner–IoU loss function with dynamic weight factors, we mitigate regression difficulties when predicted bounding boxes and ground-truth annotations exist in non-overlapping or inclusive states. This innovation enables the network to maintain effective regression under diverse conditions and further improves small object detection accuracy. Extensive experiments across multiple datasets demonstrate the effectiveness of our proposed method.

Author Contributions: Conceptualization, J.M. and M.B.; Methodology, J.M.; Software, J.M., F.F., L.L. and Z.W.; Formal analysis, J.M., H.K. and L.L.; Investigation, J.M.; Data curation, H.K.; Writing—original draft, J.M.; Writing—review & editing, T.L. and R.Z.; Visualization, Z.W. and T.L.; Supervision, R.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The datasets used in this study are all publicly available.

Acknowledgments: We would like to express our sincere gratitude to the anonymous reviewers for their insightful feedback and valuable recommendations.

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

References

1. Sun, Z.; Leng, X.; Zhang, M.; Ren, H.; Ji, K. SAR Image Object Detection and Information Extraction: Methods and Applications. *Remote Sens.* **2025**, *17*, 2098. [CrossRef]
2. Cao, S.; Deng, J.; Luo, J.; Li, Z.; Hu, J.; Peng, Z. Local convergence index-based infrared small target detection against complex scenes. *Remote Sens.* **2023**, *15*, 1464. [CrossRef]
3. Gui, S.; Song, S.; Qin, R.; Tang, Y. Remote sensing object detection in the deep learning era—A review. *Remote Sens.* **2024**, *16*, 327. [CrossRef]
4. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 18–24 July 2021; pp. 8748–8763.
5. Li, X.; Wen, C.; Hu, Y.; Yuan, Z.; Zhu, X.X. Vision-language models in remote sensing: Current progress and future trends. *IEEE Geosci. Remote Sens. Mag.* **2024**, *12*, 32–66. [CrossRef]
6. Wang, Z.; Prabha, R.; Huang, T.; Wu, J.; Rajagopal, R. Skyscript: A large and semantically diverse vision-language dataset for remote sensing. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; Volume 38, pp. 5805–5813.
7. Tao, L.; Zhang, H.; Jing, H.; Liu, Y.; Yan, D.; Wei, G.; Xue, X. Advancements in Vision–Language Models for Remote Sensing: Datasets, Capabilities, and Enhancement Techniques. *Remote Sens.* **2025**, *17*, 162. [CrossRef]
8. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
9. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In Proceedings of the Advances in Neural Information Processing Systems 28, Montreal, QC, Canada, 7–12 December 2015.
10. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.

11. Cai, Z.; Vasconcelos, N. Cascade r-cnn: Delving into high quality object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6154–6162.
12. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
13. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767. [CrossRef]
14. Li, Y.; Li, Q.; Pan, J.; Zhou, Y.; Zhu, H.; Wei, H.; Liu, C. Sod-yolo: Small-object-detection algorithm based on improved yolov8 for uav images. *Remote Sens.* **2024**, *16*, 3057. [CrossRef]
15. Yao, B.; Zhang, C.; Meng, Q.; Sun, X.; Hu, X.; Wang, L.; Li, X. SRM-YOLO for Small Object Detection in Remote Sensing Images. *Remote Sens.* **2025**, *17*, 2099. [CrossRef]
16. Hu, J.; Wei, Y.; Chen, W.; Zhi, X.; Zhang, W. CM-YOLO: Typical Object Detection Method in Remote Sensing Cloud and Mist Scene Images. *Remote Sens.* **2025**, *17*, 125. [CrossRef]
17. Law, H.; Deng, J. Cornernet: Detecting objects as paired keypoints. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 734–750.
18. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. Centernet: Keypoint triplets for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6569–6578.
19. Zhou, X.; Zhuo, J.; Krahenbuhl, P. Bottom-up object detection by grouping extreme and center points. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 850–859.
20. Yang, Z.; Liu, S.; Hu, H.; Wang, L.; Lin, S. Reppoints: Point set representation for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9657–9666.
21. Kong, T.; Sun, F.; Liu, H.; Jiang, Y.; Li, L.; Shi, J. Foveabox: Beyound anchor-based object detection. *IEEE Trans. Image Process.* **2020**, *29*, 7389–7398. [CrossRef]
22. Tian, Z.; Shen, C.; Chen, H.; He, T. Fcos: Fully convolutional one-stage object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9627–9636.
23. Zhou, J.; Xiao, C.; Peng, B.; Liu, Z.; Liu, L.; Liu, Y.; Li, X. DiffDet4SAR: Diffusion-based aircraft target detection network for SAR images. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 4007905. [CrossRef]
24. Bell, S.; Zitnick, C.L.; Bala, K.; Girshick, R. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2874–2883.
25. Tang, X.; Du, D.K.; He, Z.; Liu, J. Pyramidbox: A context-assisted single shot face detector. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 797–813.
26. Cui, L.; Lv, P.; Jiang, X.; Gao, Z.; Zhou, B.; Zhang, L.; Shao, L.; Xu, M. Context-aware block net for small object detection. *IEEE Trans. Cybern.* **2020**, *52*, 2300–2313. [CrossRef] [PubMed]
27. Zhang, M.; Yue, K.; Li, B.; Guo, J.; Li, Y.; Gao, X. Single-frame infrared small target detection via gaussian curvature inspired network. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5005013. [CrossRef]
28. Wu, S.; Xiao, C.; Wang, Y.; Yang, J.; An, W. Sparsity-Aware Global Channel Pruning for Infrared Small-target Detection Networks. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5615011. [CrossRef]
29. Zhang, S.; Zhu, X.; Lei, Z.; Shi, H.; Wang, X.; Li, S.Z. Faceboxes: A cpu real-time face detector with high accuracy. In Proceedings of the 2017 IEEE International Joint Conference on Biometrics (IJCB), Denver, CO, USA, 1–4 October 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1–9.
30. Xu, C.; Ding, J.; Wang, J.; Yang, W.; Yu, H.; Yu, L.; Xia, G.S. Dynamic coarse-to-fine learning for oriented tiny object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 7318–7328.
31. Wang, J.; Chen, K.; Yang, S.; Loy, C.C.; Lin, D. Region proposal by guided anchoring. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2965–2974.
32. Sun, P.; Zhang, R.; Jiang, Y.; Kong, T.; Xu, C.; Zhan, W.; Tomizuka, M.; Li, L.; Yuan, Z.; Wang, C.; et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 14454–14463.
33. Sun, P.; Zhang, R.; Jiang, Y.; Kong, T.; Xu, C.; Zhan, W.; Tomizuka, M.; Yuan, Z.; Luo, P. Sparse R-CNN: An end-to-end framework for object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 15650–15664. [CrossRef] [PubMed]
34. Zhang, S.; Zhu, X.; Lei, Z.; Shi, H.; Wang, X.; Li, S.Z. S3fd: Single shot scale-invariant face detector. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 192–201.
35. Zhang, S.; Chi, C.; Yao, Y.; Lei, Z.; Li, S.Z. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 9759–9768.

36. Zhu, B.; Wang, J.; Jiang, Z.; Zong, F.; Liu, S.; Li, Z.; Sun, J. Autoassign: Differentiable label assignment for dense object detection. *arXiv* **2020**, arXiv:2007.03496. [CrossRef]
37. Xu, C.; Wang, J.; Yang, W.; Yu, H.; Yu, L.; Xia, G.S. RFLA: Gaussian receptive field based label assignment for tiny object detection. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Cham, Switzerland, 2022; pp. 526–543.
38. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
39. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
40. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems 30, Long Beach, CA, USA, 4–9 December 2017.
41. Qiu, C.; Yu, A.; Yi, X.; Guan, N.; Shi, D.; Tong, X. Open self-supervised features for remote-sensing image scene classification using very few samples. *IEEE Geosci. Remote Sens. Lett.* **2022**, *20*, 2500505. [CrossRef]
42. Basso, L.D. CLIP-RS: A Cross-Modal Remote Sensing Image Retrieval Based on CLIP. Ph.D. Thesis, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA, 2022.
43. Bazi, Y.; Al Rahhal, M.M.; Mekhalfi, M.L.; Al Zuair, M.A.; Melgani, F. Bi-modal transformer-based approach for visual question answering in remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4708011. [CrossRef]
44. Sun, X.; Wang, P.; Lu, W.; Zhu, Z.; Lu, X.; He, Q.; Li, J.; Rong, X.; Yang, Z.; Chang, H.; et al. RingMo: A remote sensing foundation model with masked image modeling. *IEEE Trans. Geosci. Remote Sens.* **2022**, *61*, 5612822. [CrossRef]
45. Chen, K.; Liu, C.; Chen, H.; Zhang, H.; Li, W.; Zou, Z.; Shi, Z. RSPrompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4701117. [CrossRef]
46. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.Y.; et al. Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 4015–4026.
47. Chen, J.; Huang, Z.; Xia, R.; Wu, B.; Sheng, L.; Sun, L.; Yao, B. Large-scale multi-class SAR image target detection dataset-1.0. *J. Radars* **2022**, *14*, 1488.
48. Wei, S.; Zeng, X.; Qu, Q.; Wang, M.; Su, H.; Shi, J. HRSID: A high-resolution SAR images dataset for ship detection and instance segmentation. *IEEE Access* **2020**, *8*, 120234–120254. [CrossRef]
49. Wang, J.; Yang, W.; Guo, H.; Zhang, R.; Xia, G.S. Tiny object detection in aerial images. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 3791–3798.
50. Razakarivony, S.; Jurie, F. Vehicle detection in aerial imagery: A small target detection benchmark. *J. Vis. Commun. Image Represent.* **2016**, *34*, 187–203. [CrossRef]
51. Li, Y.; Chen, Y.; Wang, N.; Zhang, Z. Scale-aware trident networks for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6054–6063.
52. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part I 14. Springer: Cham, Switzerland, 2016; pp. 21–37.
53. Ross, T.Y.; Dollár, G. Focal loss for dense object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2980–2988.
54. Zhou, Y.; Liu, H.; Ma, F.; Pan, Z.; Zhang, F. A sidelobe-aware small ship detection network for synthetic aperture radar imagery. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5205516. [CrossRef]
55. Ma, Y.; Guan, D.; Deng, Y.; Yuan, W.; Wei, M. 3SD-Net: SAR small ship detection neural network. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*. [CrossRef]
56. Xu, C.; Wang, J.; Yang, W.; Yu, L. Dot distance for tiny object detection in aerial images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 1192–1201.
57. Chen, S.; Sun, P.; Song, Y.; Luo, P. Diffusiondet: Diffusion model for object detection. In Proceedings of the IEEE/CVF international conference on computer vision, Paris, France, 2–6 October 2023; pp. 19830–19843.
58. Cai, Z.; Liu, S.; Wang, G.; Ge, Z.; Zhang, X.; Huang, D. Align-detr: Improving detr with simple iou-aware bce loss. *arXiv* **2023**, arXiv:2304.07527.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.