

Article

Semantic Consistency and Uncertainty-Driven Small-Object Detection for Class Imbalance

Nuo Chen , Peng Zhao * and Shouquan Hou

College of Operational Support, Rocket Force University of Engineering, Xi'an 710025, China

* Correspondence: zpxhh@163.com

Highlights

What are the main findings?

- This study proposes a semantic consistency and uncertainty-driven training framework, named SCUD, to address class imbalance in UAV aerial small-object detection.
- The proposed method improves detection performance on both the DOTA-v1.0 and DIOR datasets, achieving mAP50 values of 70.7% and 88.1%, respectively, with notable improvements for rare categories and small-scale targets.

What are the implications of the main findings?

- The proposed framework provides an effective solution for improving detection robustness in class-imbalanced remote sensing scenarios.
- By jointly considering feature consistency, scale-aware resampling, and uncertainty-driven instance selection, the method can reduce missed detections and improve the practical applicability of aerial image object detection.

Abstract

In aerial image small-object detection, complex imaging perspectives, arbitrary object orientations, and long-tailed category distributions jointly exacerbate sample imbalance, which significantly degrades detection stability and leads to frequent misclassification of minority categories. To address these challenges, this paper proposes a novel training framework termed SCUD. Specifically, in the label noise suppression strategy (LNSS), a contrastive learning mechanism based on semantic consistency is introduced to constrain the aggregation of similar samples in the feature space, thereby reducing the adverse impact of noisy samples on model optimization. In addition, a scale-aware resampling strategy (SARS) is designed to alleviate noise amplification and overfitting caused by excessive repetition of small objects during training. Furthermore, an adaptive instance selection mechanism (AISM) is developed by jointly modeling prediction uncertainty and global statistical priors, enabling the model to dynamically emphasize learning from informative samples. Extensive experiments are conducted on two publicly available unmanned aerial vehicle (UAV) aerial image datasets to validate the effectiveness of the proposed approach. The proposed method achieves an mAP50 of 70.7% on the DOTA-v1.0 dataset and 88.1% on the DIOR dataset. Notably, the detection accuracy of several rare categories is significantly improved, further demonstrating the effectiveness of the proposed method in addressing sample imbalance in aerial image small-object detection.



Academic Editor: Zhenwei Shi

Received: 27 May 2026

Revised: 25 June 2026

Accepted: 3 July 2026

Published: 5 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.**Keywords:** class imbalance; contrastive learning; uncertainty-driven learning; small-object detection

1. Introduction

Aerial images offer wide-field-of-view and fixed downward perspectives, often containing numerous targets oriented in various directions within a single frame, making them an important data source in observation fields [1]. Leveraging these advantages, aerial images have been widely applied in diverse scenarios such as national defense and security [2], autonomous driving [3], and urban traffic management [4]. However, small targets in aerial images are randomly oriented and easily disturbed by high-frequency background clutter. Related research has addressed this by modeling the high-frequency differences between targets and backgrounds through dynamic high-frequency convolution to enhance discriminative representations [5]. Moreover, image collection is typically concentrated in specific regions, which objectively exacerbates the problem of class imbalance—often more severe than in traditional natural scene datasets. Consequently, the limited number of samples for rare classes presents an unavoidable practical challenge in aerial object detection research.

In recent years, semi-supervised learning and unsupervised learning methods have gradually attracted attention [6]. These methods have alleviated the decline in detection performance to some extent due to the insufficiency of labeled samples [7]. However, most of these methods rely on random labeling strategies and fail to fully consider the particularities of aerial images in terms of target distribution and category structure, which may introduce redundant samples. Sparse annotation object detection [8] has improved the above problems to some extent through object-level random labeling. However, due to the highly uneven category distribution in aerial images, this method tends to focus more on frequently occurring head categories during training, thereby suppressing the learning effect of rare categories. This not only exacerbates the redundancy of intra-class samples but also to some extent inherits and amplifies the original category imbalance problem.

To address the issue of class imbalance, some studies have introduced contrastive learning-based object detection methods [9,10]. These methods enhance the overall detection performance in aerial images by modeling similarities between objects with massive unlabeled data and designing more effective training strategies based on sample relationships [11]. However, existing contrastive learning methods mostly focus on unsupervised pre-training or clustering tasks. Due to the lack of explicit label constraints, they struggle to effectively distinguish fine-grained category differences. Moreover, many methods conduct contrastive learning at the image level, treating different images as negative samples and randomly augmented versions of the same image as positive samples [12]. Although introducing a large number of negative samples through memory banks helps learn more discriminative feature representations, such methods still face inherent limitations in class-imbalanced scenarios: most classes lack sufficient effective negative samples, while only a few classes fail to acquire adequate positive samples, leading to degraded model performance on small-sample classes [13]. Therefore, class imbalance remains a key factor restricting the performance improvement of small-object detection methods in aerial images, especially in real-world scenarios with limited sample sizes, where its adverse impact is particularly prominent.

Therefore, this paper addresses the widely existing class imbalance problem in real-world small-object detection by proposing the SCUD training framework. First, a semantic consistency-based contrastive learning approach is employed to counteract gradient bias caused by noisy labels. Second, at the data level, a scale-aware resampling strategy is introduced to suppress the repeated amplification of small objects. Finally, during training, an uncertainty-driven adaptive instance selection mechanism dynamically guides the learning focus toward scarce samples that have not yet been sufficiently modeled. The proposed method constructs a unified training framework from three perspectives, namely

feature constraint, data distribution, and instance selection, enabling systematic modeling of the class imbalance problem. The contributions of this work can be summarized as follows:

1. A label noise suppression strategy is proposed to explicitly enhance intra-class compactness and inter-class separability in the feature space by constructing weak-label-coupling semantic constraints and to employ stable contrastive gradients to counteract the gradient shift caused by noisy supervision.
2. A scale-aware resampling strategy is proposed, which normalizes the target scale via statistical modeling and introduces both a saturation upper bound for small-object sampling and batch-wise scale-ratio constraints to prevent local gradient dominance caused by repeated sampling.
3. An adaptive instance selection mechanism, based on prediction uncertainty and category scarcity prior, is introduced to adjust sample participation in the training process probabilistically, thereby avoiding redundant samples from repeatedly participating in optimization.

2. Related Works

2.1. Small-Object Detection

Compared to traditional object detection tasks in natural scenes, small-object detection presents significant differences, primarily due to the fact that targets in aerial images often appear with arbitrary orientations [14]. Therefore, numerous methods tailored for small-object detection in aerial images have been proposed [15,16]. In recent years, small-object detection approaches have gradually evolved toward more efficient network architectures and simplified training processes. Wu et al. [17] designed a multi-branch convolutional module to capture multi-scale spatial structures and extract global spatial information. Subsequently, RoI Transformer [18] introduced an RoI learner that converts horizontal RoIs into oriented RoIs, alleviating the issue of massive redundant computation caused by RRPN. DRN [19] attempted to eliminate bounding box constraints by proposing a new representation of oriented bounding boxes through sliding vertices, further optimizing the regression path. Recently, Li et al. [20] leveraged motion consistency across multiple frames to design a directional encoding temporal module, providing effective temporal discriminative information for small-object detection.

However, all the aforementioned methods adopt a two-stage detection framework, which is suboptimal in terms of detection speed. To address this, S2ANet [21] introduces deformable convolutional networks to explicitly align features with anchors, thus improving detection accuracy. Moreover, DeepPro [22] and DFBSNet [23] both adopt a refined detection pipeline to better balance detection accuracy and inference speed. Although these methods have significantly advanced small-object detection performance, they heavily depend on large-scale, meticulously annotated samples; annotating rotated bounding boxes is substantially more labor-intensive than annotating axis-aligned bounding boxes. Consequently, label-efficient object detection has garnered growing attention in recent research.

2.2. Class Imbalance

Class imbalance is a common issue in object detection datasets [24]. Although it has been extensively studied in classification tasks, it has not received the same level of attention in object detection scenarios. A large body of research indicates that class imbalance can significantly affect deep learning models, particularly in terms of their generalization ability. To address this issue, mainstream approaches typically adopt strategies such as resampling [25] or reweighting [26] to rebalance different categories, thereby alleviating performance degradation caused by uneven data distributions.

Therefore, to address the class imbalance problem in object detection tasks, Zhang et al. [27] proposed an enhanced class-balanced loss that adjusts loss weights according to class information content, thereby suppressing the undesirable positive correlation between IoU-based classification scores and EEN. Wang et al. [28] introduced a unified model-aware margin and weighting function, incorporating inter-sample difference comparisons to mitigate both inter-class and intra-class imbalances. Additionally, Qezelbash et al. [29] integrated the emerging Kolmogorov–Arnold Network with focal loss, replacing fixed neural activations with trainable univariate spline expansions to tackle class imbalance. Kiobya et al. [30] proposed a novel enhanced focal loss formulation that inherently incorporates a weighting factor to address class imbalance. Similarly, Wang et al. [31] employed gated recurrent units and cost-sensitive loss functions to improve minority-class detection accuracy degraded by skewed data distributions. Xue et al. [32] introduced an asymmetric weighted loss to dynamically assign continuous weights to positive and negative samples, thereby enhancing the detector’s classification discrimination capability.

2.3. Contrastive Learning

In recent years, contrastive learning has emerged as a promising self-supervised learning paradigm and has attracted widespread attention due to its ability to learn effective feature representations from unlabeled data [33]. The core idea of contrastive learning is to learn representative features from different types of samples, thereby mitigating the tendency of models to incorrectly group heterogeneous features into the same category [34]. This approach helps strengthen the representation of informative features and improves detection accuracy.

Therefore, some studies have explored contrastive learning for object detection tasks. Dong et al. [9] introduced self-supervised contrastive learning into the salient object detection framework, employing a token-to-token vision transformer and a vision graph neural network backbone to suppress background noise interference. Fang et al. [10] proposed a multi-level contrastive learning module that pulls positive sample pairs together and pushes negative sample pairs apart at both global and local levels. Chen et al. [35] extracted multi-scale modality-specific cues and heterogeneous cross-modal correlations by optimizing their proposed multi-scale cross-modal contrastive similarity loss. Additionally, Juyal et al. [36] proposed a joint training multiple-instance learning framework that progressively transitions from feature representation learning to classifier optimization to address low detection performance induced by class imbalance. Zhang et al. [37] integrated multi-source domain supervision with contrastive feature learning to enhance discriminative feature learning via fine-grained sample relationships. Similarly, Li et al. [38] adopted a self-guided contrastive learning strategy to train the model in a fully label-free manner, thereby mitigating label-induced bias. However, existing methods predominantly rely on single-scale feature perception, hindering comprehensive modeling of nonlinear inter-feature relationships and limiting effective learning of nonlinear representations for small aerial target feature vectors.

3. Materials and Methods

To address the data imbalance issue in small-object detection tasks, this paper proposes an SCUD model, whose overall architecture is illustrated in Figure 1. First, a contrastive learning mechanism based on semantic consistency is introduced to effectively suppress gradient bias induced by noise caused by class imbalance from an optimization perspective. Second, a resampling strategy that accounts for scale is designed to prevent unreliable small-object samples from being excessively amplified at the data level, thereby improving the quality of training data. Finally, an adaptive selection mechanism is employed to guide

the model to prioritize categories with fewer samples during training. The core idea of this method lies in unifying contrastive learning, resampling, and uncertainty modeling within a single training iteration framework, establishing a progressive optimization process that evolves from feature constraints to distribution calibration and then to instance selection. As a result, the method systematically alleviates imbalance issues and enhances the overall performance of small-object detection.

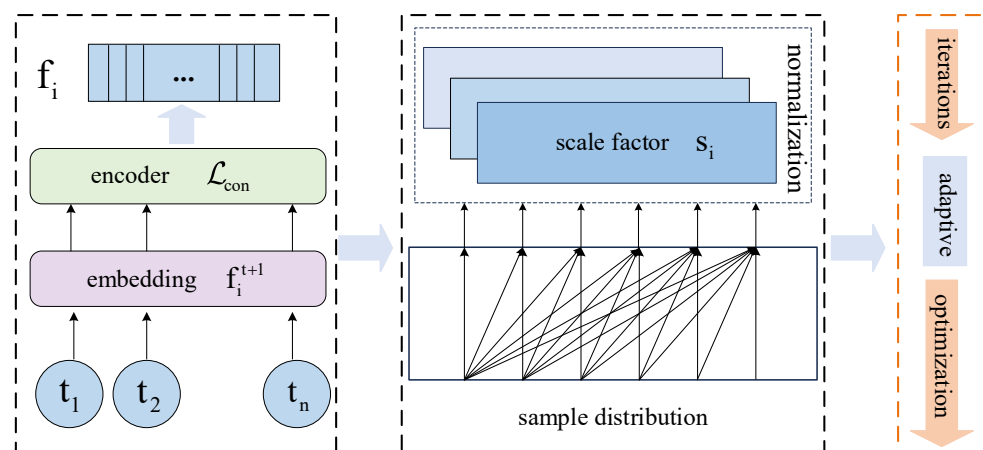


Figure 1. Overall Network Structure.

3.1. Label Noise Suppression Strategy

Object detection tasks in large-scale real-world scenarios are often significantly affected by label noise, including category mislabeling, inaccurate bounding box localization, and label omission [39]. Traditional supervised learning typically relies on cross-entropy loss to optimize classifiers, and its gradient formulation is:

$$\mathcal{L}_{CE} = - \sum_c y_c \log p_c \quad (1)$$

$$\frac{\partial \mathcal{L}_{CE}}{\partial z_c} = p_c - y_c \quad (2)$$

where $\mathcal{L}_{CE}$ represents the cross-entropy loss function, z_c denotes the output of the classifier before applying the softmax operation, c indicates the class index, p_c represents the probability predicted by the network that the sample belongs to class c , and y_c represents the corresponding ground-truth label.

When label errors occur, the noisy label $\tilde{y}_c$ directly alters the gradient direction, causing the model parameters to update toward the incorrect semantic space and resulting in gradient bias in the features. Further derivation yields:

$$\Delta g = g_{\text{noise}} - g_{\text{clean}} = y_c - \tilde{y}_c \quad (3)$$

where g_{noise} represents the gradient generated by the incorrect labels, and g_{clean} represents the gradient computed under the true labels. As shown in the above equation, noisy labels continuously provide unreliable supervisory signals toward the incorrect class and contaminate the overall feature space.

To alleviate this issue, we introduce a contrastive learning mechanism. By imposing semantic consistency constraints, we establish a weakly supervised contrastive learning framework grounded in instance-augmentation consistency and integrated with weak label information for semantic regularization. LNSS constructs positive and negative sample pairs within the feature space and computes the contrastive loss, thereby providing stable semantic-structural constraints at the optimization level. As illustrated in Figure 2, for each

instance, a set of positive pairs is constructed, including augmented views with different perspectives or scales; instance-level features are extracted from the multi-scale feature maps output by the neck. At the same time, negative pairs are constructed to impose separation constraints on samples that are not semantically consistent. The optimization objective is given by the InfoNCE loss:

$$s_{\text{pos}} = \frac{\langle f_i, f_i^+ \rangle}{\tau} \quad s_{\text{neg},j} = \frac{\langle f_i, f_j^- \rangle}{\tau} \quad (4)$$

$$\mathcal{L}_{\text{con}} = -\log \frac{\exp(s_{\text{pos}})}{\exp(s_{\text{pos}}) + \sum_j \exp(s_{\text{neg},j})} \quad (5)$$

where f_i denotes the feature vector of the current sample; f_i^+ denotes the semantically consistent positive feature; f_j^- denotes the j -th negative feature; τ is the temperature parameter controlling softmax smoothness, set to 0.07 [40]; s_{pos} denotes the similarity between the anchor and the positive sample; and $s_{\text{neg},j}$ denotes the similarity score between the anchor and the j -th negative sample. From a gradient perspective, the contrastive loss updates the feature representation as:

$$\frac{\partial \mathcal{L}_{\text{con}}}{\partial f_i} \propto -\left(f_i^+ - \sum_j w_j f_j^-\right) \quad (6)$$

where f_i^+ denotes the positive sample feature, and f_j^- denotes the j -th negative sample feature. The positive sample term f_i^+ exerts an attractive force on the anchor feature, while the weighted negative sample term $\sum_j w_j f_j^-$ exerts a repulsive force on the corresponding negative samples via the weight w_j .

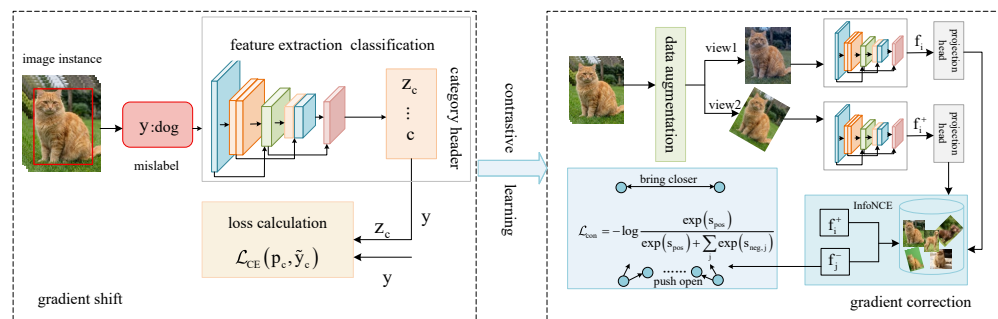


Figure 2. Weak-label-coupled instance-level contrastive learning framework.

Contrastive learning explicitly shapes intra-class compactness and inter-class separability in the feature space, providing a stable structural prior that is independent of label noise at the mechanism level; its gradient always points in the direction of bringing semantically consistent features closer together and pushing semantically inconsistent features apart. Therefore, the gradient direction g_{con} provided by contrastive learning can counteract the offset gradient Δg caused by noisy labels, making the updated features satisfy the following formula:

$$f_i^{t+1} = f_i^t - \gamma(g_{\text{noise}} + g_{\text{con}}) \quad (7)$$

where f_i^t denotes the feature representation at the t -th iteration; $g_{\text{noise}} + g_{\text{con}}$ denotes the final actual update direction; and γ denotes the contrastive learning loss weight. This operation mainly introduces a correction term at the optimization level to suppress the long-term contamination of the feature distribution by the noise gradient, thereby yielding a more discriminative and noise-robust feature representation.

Through this mechanism, even if the labels are incorrect, the model can still leverage semantic consistency constraints to learn stable feature representations, thereby achieving automatic correction of the feature space and improved robustness to label noise. In other words, contrastive learning constructs a structured semantic space during optimization: it repels noisy samples toward semantic boundaries and encourages foreground samples to aggregate more tightly, thus reducing the interference of label noise on the detection model.

3.2. Scale-Aware Resampling Strategy

In the previous section, by introducing a contrastive learning mechanism, we constructed a semantic consistency constraint that is weakly coupled with labels, making feature learning more robust to label noise. However, relying solely on feature-level constraints is still insufficient to address the inherent class imbalance problem in object detection, especially as small objects are prone to being overly amplified during resampling, leading to overfitting.

Therefore, in this section, we introduce SARS at the sample distribution level, as shown in Figure 3, to perform scale normalization modeling and probability resampling on the input samples. Meanwhile, we impose scale-ratio constraints during the mini-batch construction process to form structurally balanced and noise-controlled training batches, thereby suppressing the excessive repetition of small targets. Specifically, to explicitly model the differences among targets at different scales in the sampling stage, we first quantify the scale of each target instance and normalize it into the image-scale space. The normalized scale factor of target i is defined as:

$$s_i = \left[(WH)^{-1} (w_i h_i) \right]^{\frac{1}{2}} \quad (8)$$

where w_i and h_i denote the width and height of the preprocessed bounding box, respectively, while W and H denote the width and height of the preprocessed image, respectively, and s_i is computed based on candidate instances across the entire training set after scale partitioning is completed.

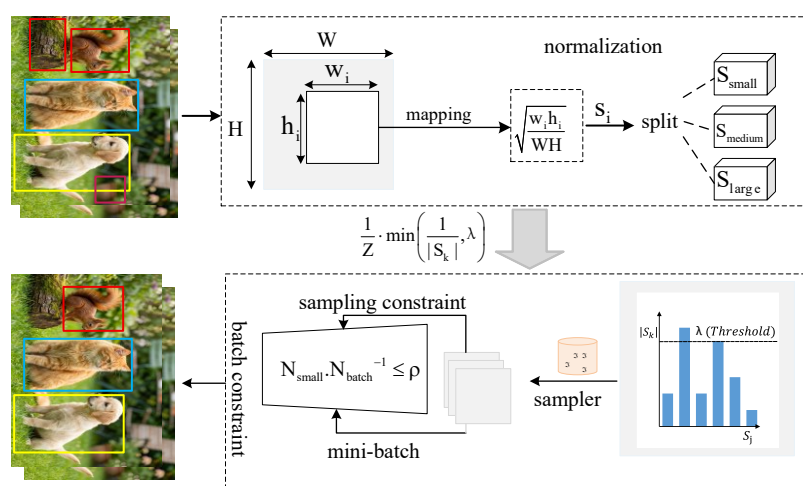


Figure 3. Scale-aware resampling strategy diagram.

Based on the scale factor s_i , we divide the targets into three categories: small ($s < 0.02$), medium ($0.02 \leq s < 0.05$), and large ($s \geq 0.05$). An empirical normalization scale threshold is determined based on the target scale distribution of the training set. Small targets suffer the most significant degradation in detection performance, as their features are prone to being lost during down-sampling. Therefore, scale-dependent sampling probabilities are

assigned to targets at different scales during the sampling stage. To prevent excessive resampling of small targets, a saturation constraint is introduced into their sampling probabilities. The sampling probability of target i is defined as:

$$P(i) = \frac{1}{Z} \cdot \max\left(\min\left(\frac{1}{|S_k|}, \lambda\right), \varepsilon\right), \quad i \in S_k \quad (9)$$

where S_k denotes the scale interval of the target, and $|S_k|$ is the number of targets within that interval. The parameter λ , set to 0.001 [41], defines the upper bound of the small-target sampling probability to avoid excessive resampling. The parameter ε is the lower bound of the sampling weight, preventing severe under-sampling when $|S_k|$ is large. The normalization constant z ensures sufficient observation for small targets.

Although sample-level resampling can alleviate the problem of excessive repetition of small targets, small targets may still dominate the gradients within local batches during mini-batch construction. To address this, we further impose scale-ratio constraints at the batch level:

$$N_{\text{small}} \cdot N_{\text{batch}}^{-1} \leq \rho \quad (10)$$

where N_{small} is the number of small objects in the current mini-batch, N_{batch} is the total number of objects in the batch, and ρ is a predefined upper bound on the proportion of small objects.

Inspired by the foreground and background sampling ratio setting for mini-batches in Fast R-CNN [42], the upper bound ρ on the proportion of small objects in each mini-batch is set to 0.25. This constraint ensures scale diversity within each mini-batch, guaranteeing that the gradient update reflects multi-scale discrimination requirements. By combining scale-aware resampling with semantic consistency constraints, the proposed framework achieves robustness at both the data distribution and feature representation levels.

3.3. Adaptive Instance Selection Mechanism

The first two stages respectively alleviate the imbalance problem in small-object detection from the perspectives of feature-representation robustness and sample-scale distribution. To further enhance the model's learning efficiency for key samples, we introduce an adaptive instance selection mechanism based on uncertainty during training, as shown in Figure 4. AISM calculates instance weights over the pre-formed candidate positive sample set using prediction uncertainty and category- and scale-specific statistical priors, and probabilistically selects a subset to participate in the loss computation. The statistical prior is precomputed based on the entire training set before training begins and remains fixed throughout the training process. The specific implementation is detailed in Algorithm 1.

During the training process, we utilize the model's current prediction response to characterize the predictive reliability of individual instances. Specifically, for instance, for i in the t -th iteration, its uncertainty measure is defined as:

$$u_i^{(t)} = \left(1 - p_{\theta}^{(t)}(y_i | x_i) \cdot q_i^{(t)}\right)^{\eta} \quad (11)$$

where $p_{\theta}^{(t)}(y_i | x_i)$ denotes the model's predicted probability for the instance's ground-truth class; $q_i^{(t)}$ is the output of an auxiliary prediction branch that reflects localization quality. η is a hyperparameter that controls the degree of nonlinearity in the uncertainty response.

Specifically, q_i denotes the localization quality estimation score of the i -th candidate positive instance, and its value lies in $[0, 1]$. It is predicted by a lightweight auxiliary quality branch that shares the backbone and neck features with the detector and is placed in parallel with the classification and regression branches.

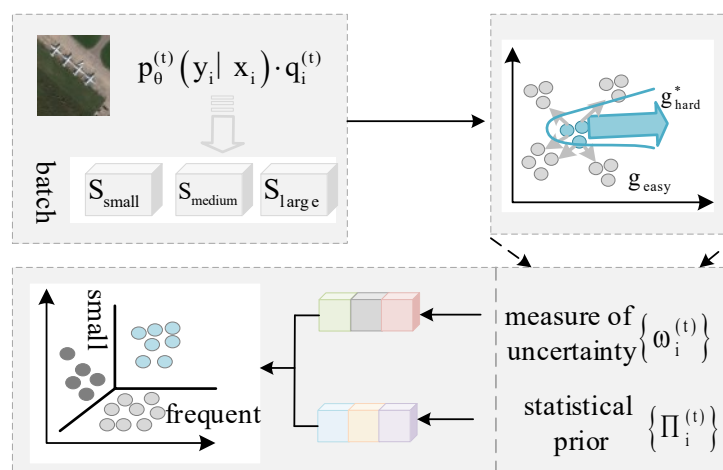


Figure 4. Adaptive instance selection structure.

Algorithm 1 Adaptive Instance Selection Mechanism for Uncertainty

Input: Candidate positives in a mini-batch $S = \{(x_i, y_i, s_i, p_i, q_i)\}$; hyperparams η, α, β ; sample size B

Output: Selected subset S_{sample} for loss computation

```

1: for each training step  $t$  do
2:   Count class/scale frequencies  $N_y, N_s$  over  $S$ 
3:   for each instance  $i \in S$  do
4:     Compute uncertainty score  $u_i$  from  $(p_i, q_i, \eta)$ 
5:     Compute weight  $\omega_i$  combining  $u_i$  with  $(N_{y_i}, N_{s_i}, \alpha, \beta)$ 
6:   end for
7:   Normalize weights to probabilities  $\Pi_i = \omega_i / \sum_j \omega_j$ 
8:   Sample indices  $I$  from  $S$  with probs  $\{\Pi_i\}$ , sample size  $B$ ,
   without replacement
9:   Form  $S_{\text{sample}} = \{(x_i, y_i, s_i, p_i, q_i) \mid i \in I\}$ 
10:  Compute loss on  $S_{\text{sample}}$  and update model parameters
11: end for

```

Based on the uncertainty measure, we further incorporate a statistical prior derived from the global sample distribution to explicitly characterize the relative scarcity of instances across categories and scales. For instance i , its comprehensive importance weight is defined as:

$$\omega_i^{(t)} = u_i^{(t)} \cdot \left(\frac{1}{N_{y_i}} \right)^\alpha \cdot \left(\frac{1}{N_{s_i}} \right)^\beta \quad (12)$$

where N_{y_i} and N_{s_i} respectively represent the numbers of samples belonging to the instance's category and to the scale interval, while α and β are hyperparameters that control the influence strength of the statistical prior on sample selection.

Inspired by the class-imbalance reweighting mechanism in Focal Loss [43], we adopt a smooth reweighting strategy based on the square root function, setting α to 0.5 and β to 0.5. This weight design explicitly assigns higher importance to instances that exhibit both higher prediction uncertainty and relative scarcity in the data distribution during training.

Based on the instance importance weights defined above, we perform adaptive instance selection on the candidate positive sample set S in each mini-batch, ensuring that instances contributing more to model optimization are more likely to be selected for training. The selection probability is defined as:

$$\Pi_i^{(t)} = \left(\sum_{j \in S} \omega_j^{(t)} \right)^{-1} \omega_i^{(t)} \quad (13)$$

where $\omega_i^{(t)}$ denotes the weight of instance i at the t -th iteration, and $\omega_j^{(t)}$ denotes the weight of instance j in the candidate set S .

In terms of loss function design, we introduce a contrastive learning loss in addition to the original detection loss and perform weighted optimization on the subset selected by AISM. The final joint loss function is:

$$L_{\text{total}} = L_{\text{det}}(S_{\text{sel}}) + L_{\text{con}}(S_{\text{sel}}) \quad (14)$$

where S_{sel} denotes the subset of instances selected by AISM, L_{det} denotes the standard detection loss, and L_{con} denotes the contrastive loss constructed by LNSS.

Under this mechanism, the model no longer processes all samples in a static or uniform manner, but dynamically adjusts the frequency of instance participation in optimization based on its own prediction status and the characteristics of the sample distribution. This process effectively suppresses the redundant involvement of well-learned samples while strengthening the supervision of small-scale instances.

4. Experimental Results

4.1. Experimental Details

The experiments were conducted on an Ubuntu 20.04 operating system, using PyTorch 2.0 as the deep learning framework and Python 3.8 as the programming language, with a GPU acceleration environment built on CUDA 11.8. The hardware platform consists of an Intel(R) Xeon(R) Platinum 8352 V processor for the CPU and an NVIDIA RTX 4090 graphics card with 24 GB of video memory. This software and hardware configuration provides efficient and stable computational support for model training and inference, while ensuring good reproducibility and reliability throughout the experimental process. Specific details of the experimental environment are shown in Table 1.

Table 1. Experimental environment.

Configurations	Details
Operating System	Ubuntu 20.04
Language	Python 3.8
Framework	PyTorch 2.0
CPU	Intel(R) Xeon(R) Platinum 8352 V
GPU	NVIDIA RTX 4090
GPU Memory	24 G

The experimental setup is as follows: The batch size is set to 8, and the input image resolution is 640×640 pixels. The dataset is split into training and validation sets. The model is trained for 200 epochs. Specific training parameters are summarized in Table 2.

Table 2. Training parameters.

Parameters	Values
Batch size	8
Epoch	200
Optimizer	SGD
Learning rate	0.01
Momentum	0.937
Input size	640×640

4.2. Datasets

DOTAv1.0 includes 2806 aerial images and more than 188,000 annotated object instances belonging to 15 categories, including plane, ship, storage tank, baseball diamond, tennis court, basketball court, bridge, and vehicle. Unlike traditional horizontal bounding box annotations, the dataset is characterized by large-scale variations, dense object distributions, arbitrary object orientations, and complex backgrounds, making it highly challenging for object detection tasks.

The DIOR dataset consists of 23,463 high-resolution optical remote sensing images and 192,472 annotated object instances across 20 object classes: airplane, airport, baseball field, basketball court, bridge, chimney, dam, expressway service area, expressway toll station, golf field, ground track field, harbor, overpass, ship, stadium, storage tank, tennis court, train station, vehicle, and windmill. The dataset contains approximately 1200 images with no annotated objects; all images are 800×800 pixels in size and have a spatial resolution ranging from 0.5 to 30 m.

4.3. Evaluation Metrics

In this study, precision, recall, mAP, and parameter count were adopted as the primary evaluation metrics to comprehensively assess detection performance and model complexity. The corresponding formulas are defined as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (15)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (16)$$

where TP denotes the number of true positive samples, TN denotes the number of true negative samples, FP represents the number of false positive samples, and FN represents the number of false negative samples. Precision measures the accuracy of positive predictions, while recall measures the proportion of correctly detected positive samples.

$$AP = \int_0^1 \text{Precision}(\text{Recall}) d(\text{Recall}) \quad (17)$$

$$\text{mAP} = \frac{1}{N} \sum_{t=1}^N AP_t \quad (18)$$

AP represents the average precision for a single class, while mAP represents the mean of AP values across all classes. AP_t denotes the average precision of the t -th class, and N denotes the number of classes.

4.4. Comparative Experiments

4.4.1. Experimental Results on DOTAv1.0 Dataset

To verify the effectiveness of the proposed method on the category-imbalanced aerial image dataset, YOLOv8n was selected as the baseline due to its balance between detection accuracy and efficiency. The proposed method was compared with several representative and advanced detectors on the DOTAv1.0 dataset, and the results are shown in Table 3. As shown in the table, the proposed method achieves the best overall performance, with an mAP_{50} of 70.7%, outperforming YOLOv8 by 4.6 percentage points and slightly surpassing A2S-DET by 0.1 percentage points. In terms of category-level performance, the proposed method obtains the highest accuracy in the BR category, reaching 73.9%. It also achieves competitive results in BD and GTF, with accuracies of 76.5% and 65.3%, respectively. These results indicate that the proposed method improves the overall detection capability under category-imbalanced conditions while maintaining strong performance on several

challenging categories. Therefore, the experimental results verify the effectiveness of the proposed method for aerial image object detection.

Table 3. Performance evaluation on the DOTAv1.0 dataset.

Methods	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP50
YOLOv8	87.4	72.4	65.9	62.7	67.4	61.8	60.7	85.7	75.8	77.4	52.3	59.7	55.3	54.1	54.2	66.1
YOLOv9	91.2	74.5	48.3	58.8	49.5	75.3	59.2	88.9	64.6	62.7	64.1	65.9	74.4	55.9	59.8	66.2
YOLOv10	90.8	63.6	73.8	50.8	71.2	50.6	76.9	63.7	90.8	65.7	64.8	59.7	61.3	71.5	52.4	67.2
RT-DETR	92.0	74.9	51.6	73.5	53.8	76.4	63.7	91.5	66.3	64.9	64.3	59.3	72.5	52.1	60.5	67.8
LO-DET [44]	89.2	66.1	31.3	55.9	70.0	71.0	84.2	90.7	75.0	81.2	44.6	59.3	59.9	65.1	48.4	66.1
A2S-DET [45]	89.5	77.8	46.3	56.4	75.8	74.8	86.0	90.5	81.0	83.7	50.2	60.9	65.2	69.7	50.9	70.6
STPANet [46]	91.5	72.2	46.8	58.2	68.4	85.4	87.7	93.5	64.9	68.6	52.0	59.3	82.7	62.1	54.7	69.9
Mamba-YOLO [47]	94.2	75.7	52.4	75.4	61.2	75.6	74.3	92.5	60.3	62.5	62.8	61.4	71.4	59.4	57.8	69.1
WSOOD-GAEN [48]	89.3	71.1	43.8	54.0	72.0	81.4	88.4	90.7	71.6	74.1	62.2	61.5	73.9	50.7	49.9	69.0
Ours	90.5	76.5	73.9	65.3	71.3	64.0	80.2	89.6	81.8	82.6	55.6	63.1	56.6	51.4	57.7	70.7

Table 4 compares different methods on the DOTAv1.0 dataset in terms of precision, recall, model parameters, and computational cost. Besides its advantage in addressing class imbalance, the proposed method also demonstrates clear superiority in overall detection performance and model efficiency. The proposed method achieves the highest precision of 76.6% and recall of 67.1% among all compared methods, outperforming YOLOv8n, LRRS, and BFRI-YOLO in detection accuracy and robustness. In addition, the proposed method contains only 3.0 M parameters, which is fewer than LRRS with 7.0 M parameters and BFRI-YOLO with 3.4 M parameters, indicating a more lightweight network structure. Although maintaining a low computational cost of 11.5 GFLOPs, the proposed method still achieves better detection performance than other methods with higher computational complexity. Compared with YOLOv8n, the proposed method improves precision and recall significantly while reducing parameter redundancy. These results demonstrate that the proposed method not only alleviates the impact of class imbalance but also achieves advantages in detection accuracy, lightweight design, and computational efficiency.

Table 4. Performance comparison of methods on DOTAv1.0 dataset.

Method	Precision	Recall	Params	GFLOPs
YOLOv8n	45.2	34.1	3.2	8.7
STPANet [46]	75.7	64.0	7.0	16.1
BFRI-YOLO [49]	69.3	41.7	3.4	18.6
Ours	76.6	67.1	3.0	11.5

4.4.2. Experimental Results on DIOR Dataset

According to Table 5, the proposed method achieves the best performance on the DIOR dataset, with an mAP50 of 88.1%. Compared with the second-best method, DASSF, which obtains 86.7%, the proposed method improves the mAP50 by 1.4 percentage point. It also outperforms Gold-YOLO by 2.1 percentage points and RTMDET by 2.8 percentage points, demonstrating clear overall superiority. In terms of category-wise performance, the proposed method achieves leading results in several representative categories. For example, it obtains 99.3% in the airplane category, exceeding DASSF at 97.0% and RT-DETR at 82.2%. In the airport category, it reaches 80.9%, which is higher than DN-DETR at 74.2%. In the bridge category, the proposed method achieves 98.1%, outperforming DASSF at 96.6%. In the storage tank category, it reaches 97.2%, showing strong detection capability for dense and small-scale objects. Overall, the proposed method achieves superior detection accuracy and exhibits stronger robustness and generalization ability in complex remote sensing scenarios.

Table 5. Performance evaluation of some representative categories of the DIOR dataset.

Model	AL	AP	BF	BC	BG	CM	DAM	EA	ES	GC	GF	HB	OP	SP	SD	ST	TC	TS	VH	WM	mAP50
RT-DETR	82.2	85.8	84.7	90.2	48.7	74.7	61.8	87.8	78.5	77.0	84.4	60.7	61.5	91.0	72.0	75.7	90.4	59.7	59.1	93.5	76.0
DINO-SE [50]	71.9	89.0	78.7	87.7	51.9	81.5	68.9	89.2	78.9	79.9	83.5	53.4	64.6	76.4	76.9	73.6	87.9	67.5	54.6	90.4	75.3
RTMDET [51]	96.2	94.2	94.4	94.3	56.7	93.6	78.4	96.3	86.0	90.8	90.8	77.8	71.9	77.7	96.5	79.2	96.9	78.1	65.0	92.0	85.3
DN-DETR [52]	85.8	74.2	78.9	89.8	46.1	77.8	60.5	65.1	65.3	75.6	75.7	59.9	58.6	89.7	71.9	78.7	89.5	55.4	56.4	78.1	71.6
LERT-DETR [53]	88.2	85.2	87.1	89.5	50.2	76.2	68.7	88.9	80.8	81.6	84.7	62.0	64.3	91.7	78.5	77.7	92.0	62.3	61.5	93.0	78.2
Gold-YOLO [54]	96.2	93.2	95.8	92.9	57.2	89.9	79.8	96.8	85.9	88.5	90.9	74.7	72.2	94.2	97.0	86.5	97.4	76.2	61.5	94.2	86.0
MTGS-YOLO [55]	97.4	82.1	95.2	91.6	55.7	85.6	67.5	90.1	82.7	79.9	87.1	75.7	67.2	93.6	94.3	90.9	96.0	63.5	66.7	90.6	82.6
DASSF [56]	97.0	92.8	96.6	91.7	56.6	90.0	80.7	96.7	87.9	89.9	91.0	76.0	69.9	95.3	97.3	90.6	97.9	70.7	69.7	95.0	86.7
Ours	99.3	80.9	98.1	92.1	63.0	95.1	68.2	94.6	91.4	90.0	92.9	76.3	79.1	98.2	97.8	97.2	98.6	83.1	80.6	94.9	88.1

As shown in Table 6, the proposed method demonstrates a clear advantage on the DIOR dataset. Compared with DINO-SE, the proposed method reduces the parameter count to 3.0 million and the GFLOPs to 10.8, achieving notable improvements in overall efficiency. Compared with DASSF, it increases the recall rate from 79.4% to 83.9%, delivering faster detection with fewer parameters and lower computational cost. Compared with MTGS-YOLO, it improves precision from 84.6% to 87.6% and recall from 77.8% to 83.9%, while reducing parameters to 3.0 million and achieving an inference speed of 100.1 FPS, striking a superior balance among accuracy, efficiency, and model lightweights. Compared to FCDet, which employs feature-level contrastive learning, our method achieves better performance in terms of parameter efficiency and real-time detection. Overall, through comprehensive optimization, the proposed method maintains leading precision and recall rates while significantly reducing model complexity and increasing inference speed, demonstrating enhanced practicality and competitiveness.

Table 6. Performance comparison of methods on DIOR dataset.

Method	Precision	Recall	Params	GFLOPs	FPS
DINO-SE [50]	–	–	51.0	191.0	6.2
MTGS-YOLO [55]	84.6	77.8	–	77.0	86.0
DASSF [56]	87.4	79.4	–	–	46.7
FCDet [57]	–	–	41.3	134.4	38.4
Ours	87.6	83.9	3.0	10.8	100.1

4.5. Ablation Experiments

4.5.1. Ablation Experiment of Each Module

This section systematically analyzes the contribution of each submodule to overall performance through ablation studies to verify the effectiveness of the proposed method and determine the optimal configuration. All experiments were conducted on the DOTA1.0 dataset to ensure consistency in training epochs and evaluation protocols. Table 7 presents the detection performance comparison under different module combinations.

From the baseline model, it can be seen that without introducing any modules, the model achieves 66.0%, 59.4%, 61.7%, and 40.2% in Precision, Recall, mAP50, and mAP50-95 respectively, indicating that the overall detection performance is still affected by class imbalance and insufficient feature representation. After introducing LNSS alone, Precision and Recall increase to 67.4% and 60.9% respectively, while mAP50-95 rises to 40.7%, demonstrating that the semantic consistency constraint can effectively enhance feature discrimination ability and improve detection stability. After introducing SARS alone, Precision and Recall increase to 67.2% and 60.2%, respectively, and mAP50-95 reaches 41.0%; this indicates that scale-aware resampling can enhance scale-sensitive object feature representations by integrating fine-grained spatial information with high-level semantic information. Even when the target scale falls below the small-object range ($s < 0.02$), the proposed method can still preserve richer discriminative information. When only AISM is introduced, Precision increases to 68.4%, Recall reaches 61.3%, and mAP50-95 improves to 41.2%, verifying the effectiveness of the uncertainty-based instance selection mechanism in highlighting hard samples and improving localization quality.

In the case of multi-module combination, performance improvement is more significant. After combining LNSS and SARS, Precision and Recall increase to 72.1% and 64.3%, respectively, and mAP50-95 reaches 43.1%, indicating that semantic constraints and scale distribution correction exhibit strong complementarity at both the feature and data levels. When LNSS and AISM are combined, mAP50 increases to 66.1% and mAP50-95 reaches 44.6%, further demonstrating that uncertainty-guided instance selection enhances

the model's ability to learn from difficult and scarce samples. When SARS and AISM are combined, mAP50 increases to 65.4%, reflecting the synergistic effect between scale-aware resampling and dynamic instance selection. Ultimately, when all three modules are introduced simultaneously, the model achieves optimal overall performance, with Precision at 76.6%, Recall at 67.1%, mAP50 at 70.7%, and mAP50-95 at 47.7%. This result indicates that the modules collaboratively optimize the model across three dimensions—feature robustness, scale balance, and instance-level adaptability—thereby effectively enhancing both the overall detection performance and robustness for small objects.

Table 7. Ablation experiment design for each module.

LNSS	SARS	AISM	Precision	Recall	mAP50	mAP50-95
			66.0	59.4	61.7	40.2
✓			67.4	60.9	62.3	40.7
	✓		67.2	60.2	63.2	41.0
		✓	68.4	61.3	62.7	41.2
✓	✓		72.1	64.3	64.4	43.1
✓		✓	72.9	66.5	66.1	44.6
	✓	✓	73.2	65.1	65.4	43.9
✓	✓	✓	76.6	67.1	70.7	47.7

As shown in Table 8, the detection performance of both SCUD w/o LNSS and SCUD w/LNSS decreases as the proportion of random label noise increases. However, the performance degradation is significantly mitigated when LNSS is introduced. In the absence of label noise, SCUD w/LNSS achieves mAP50 and mAP50-95 values of 70.7% and 47.7%, respectively. When the noise ratio increases to 10%, SCUD w/LNSS still maintains mAP50 and mAP50-95 values of 68.6% and 45.9%. Under the 20% noise setting, the mAP50 of SCUD w/o LNSS drops to 56.8%, and its mAP50-95 decreases to 36.7%, whereas SCUD w/LNSS still achieves 65.1% and 43.4%, respectively. These results demonstrate that LNSS can better preserve detection performance under increasing label noise, indicating that its semantic consistency constraint effectively alleviates the interference of incorrect labels with feature learning and gradient-based optimization, thereby improving the robustness of the model under noisy supervision.

Table 8. Robustness comparison under different random label noise ratios.

Noise Ratio	Model	mAP50	mAP50-95
0%	SCUD w/o LNSS	65.4	43.9
0%	SCUD w/LNSS	70.7	47.7
10%	SCUD w/o LNSS	61.6	40.5
10%	SCUD w/LNSS	68.6	45.9
20%	SCUD w/o LNSS	56.8	36.7
20%	SCUD w/LNSS	65.1	43.4

4.5.2. The Weights γ and Hyperparameters η

This paper systematically analyzes the impact of η in the proposed AISM module and the contrastive learning weight γ on the detection model's performance. The results in Tables 9 and 10 show that different parameter settings significantly affect model performance. From Table 9, when η is 0.1, the model achieves an mAP50 of 69.5% and an mAP50-95 of 46.2%, but precision remains relatively low at 75.2%. When η increases to 0.2, the model achieves the best overall performance, with precision rising to 76.4%, recall reaching 66.6%, mAP50 increasing to 70.4%, and mAP50-95 improving to 47.0%, indicating that moderate adjustment of the adaptive selection parameter can effectively enhance the

model's comprehensive detection capability. However, as η further increases to 0.3 and 0.4, both mAP50 and recall gradually decline. Specifically, mAP50 decreases to 69.6% and 68.9%, while mAP50-95 drops to 46.5% and 45.9%, respectively. These results suggest that an excessively large η may introduce bias in the adaptive sample selection process, thereby weakening the overall optimization effect and reducing detection stability.

Table 9. Performance comparison under different η values.

η	Precision	Recall	mAP50	mAP50-95
0.1	75.2	65.8	69.5	46.2
0.2	76.4	66.6	70.4	47.0
0.3	76.1	66.2	69.6	46.5
0.4	75.9	65.7	68.9	45.9

Additionally, from Table 10, it can be observed that the contrastive learning weight γ also affects the model performance. When γ is 0.1, the overall performance of the model is relatively low, with mAP50 and mAP50-95 being 69.8% and 46.9% respectively. When γ increases to 0.2, mAP50 rises to 70.3%, and mAP50-95 increases to 47.4%. Further increasing to 0.3, Precision, Recall, mAP50, and mAP50-95 reach 75.9%, 66.8%, 70.7%, and 47.5% respectively, achieving the best overall performance, indicating that moderately enhancing the contrastive learning constraint can improve the feature discrimination ability and localization quality. When γ continues to increase to 0.4, mAP50 drops to 69.2%, and mAP50-95 decreases to 46.8%, suggesting that overly strong contrastive learning constraints may affect the optimization balance of the detection branch. Based on the comprehensive analysis of the two sets of experimental results, this paper ultimately selects η as 0.2 and γ as 0.3 as the preferred parameter settings, as this combination demonstrates more stable performance in terms of detection accuracy, recall rate, mAP50, and mAP50-95, while improving feature representation and maintaining stable detection branch optimization.

Table 10. Performance comparison under different γ values.

γ	Precision	Recall	mAP50	mAP50-95
0.1	75.5	65.9	69.8	46.9
0.2	65.2	66.0	70.3	47.4
0.3	75.9	66.8	70.7	47.5
0.4	75.8	65.6	69.2	46.8

4.6. Visualization

To further validate the detection performance of the proposed SCUD, qualitative visual comparison experiments were conducted across multiple scenarios, as illustrated in Figure 5.

Compared with the baseline method, the proposed approach demonstrates superior detection completeness and stability in scenarios involving dense objects, small-scale targets, and complex backgrounds. In scenes such as densely parked vehicles, closely arranged aircraft, harbor ships, and small objects around sports fields, the proposed method achieves more accurate localization, produces more complete bounding box coverage, and significantly reduces missed detections. In contrast, the baseline method exhibits multiple missed detections, particularly in regions with small-object scales, dense distributions, or strong background interference.

In conjunction with the SCUD framework described in the paper, the improvements can be attributed to three key components. The semantic-consistency-based contrastive learning enhances feature discriminability, the scale-aware resampling strategy mitigates

the over-amplification of small-object samples, and the uncertainty-driven adaptive instance selection mechanism strengthens the learning of critical samples. These designs collectively improve the model's capability to detect minority classes, small objects, and hard examples.



Figure 5. Qualitative comparison on DOTA-v1.0. (a) Original images. (b) SCUD results. (c) Baseline results with missed detections. (d) SCUD results on the same regions.

5. Discussion

The experimental results demonstrate that the proposed SCUD framework achieves consistent improvements on the DOTAv1.0 and DIOR datasets, indicating its effectiveness in alleviating class imbalance in aerial small-object detection. These gains mainly come from the joint optimization of feature representation, sample distribution, and instance selection.

From the perspective of feature learning, LNSS introduces semantic-consistency contrastive constraints to reduce the influence of noisy labels on gradient optimization. By enhancing intra-class compactness and inter-class separability, it improves feature discrimination and detection robustness, especially under noisy supervision. At the data level, SARS addresses the excessive repetition of small objects during resampling. By modeling target scale and constraining the proportion of small objects, it helps maintain scale diversity and reduces overfitting caused by local gradient dominance. Furthermore, AISM improves the training focus by selecting samples according to prediction uncertainty and global category-scale priors. This mechanism allows the model to pay more attention to scarce, difficult, and informative instances while reducing redundant optimization on well-learned samples. Compared with existing methods that mainly rely on reweighting, resampling, or feature enhancement alone, SCUD provides a unified framework for handling class imbalance from multiple complementary perspectives. The ablation results further confirm that each module contributes to performance improvement, while their combination achieves the best overall results.

SCUD provides an effective solution for detecting small aerial targets in imbalanced categories, achieving a favorable balance between accuracy, robustness, and efficiency.

Future work should validate the method under more diverse remote sensing scenarios, imaging conditions, and detection backbones.

6. Conclusions

This paper addressed the class imbalance problem in small-object detection by proposing a training framework tailored for UAV aerial scenarios. First, we introduced a semantic-consistency contrastive learning constraint that enhances the coherence of same-class samples in feature space and suppresses the interference of noisy samples on gradient direction. Second, we designed a scale-aware resampling strategy that reasonably caps the sampling frequency of small objects and constrains the scale distribution within mini-batches, mitigating overfitting and training instability caused by the frequent resampling of small objects. Third, we proposed an adaptive instance selection mechanism that dynamically focuses training on high-uncertainty, relatively scarce critical samples and reduces the interference of redundant samples in the optimization process. Extensive experimental results demonstrate that the proposed method achieves stable and consistent performance improvements on UAV aerial datasets, with particularly pronounced advantages on few-shot categories and small-scale targets.

Author Contributions: N.C.: conceptualization, methodology, writing—original draft, software, formal analysis, investigation; P.Z.: supervision, project administration; S.H.: visualization. All authors have read and agreed to the published version of the manuscript.

Funding: This research did not receive funding.

Data Availability Statement: Data will be made available on request.

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

Abbreviations

The following abbreviations are used in this manuscript:

PL	plane	BF	baseball field
BR	bridge	BG	bridge
GTF	ground track field	CM	chimney
SV	small vehicle	EA	Expressway-Service-area
LV	large vehicle	ES	Expressway-toll-station
SH	ship	GC	ground track field
TC	tennis court	GF	golf field
BC	basketball court	HB	harbor
ST	storage tank	OP	overpass
SBF	soccer ball field	SP	ship
RA	roundabout	SD	stadium
HA	harbor	SP	swimming pool
HC	helicopter	TS	train station
AL	airport	VH	vehicle
AP	airplane	WM	windmill
BD	baseball diamond		

References

1. Gill, M.U.; Rajendran, P. Attention-integrated one-stage detector with weighted feature fusion for small object detection in aerial images. *Eng. Appl. Artif. Intell.* **2026**, *165*, 113429. <https://doi.org/10.1016/j.engappai.2025.113429>.
2. Sun, J.; Liu, Z.; Xie, M.; Wang, Y.; Yuan, R.; Wu, J.; Duan, S.; Wang, L. DP-CIADet: Detail perception and critical information aggregation compact detection network for aerial optical remote sensing images. *Opt. Laser Technol.* **2026**, *193*, 113519. <https://doi.org/10.1016/j.optlastec.2025.113519>.

3. Jiang, W.; Zhao, D.; Wang, C.; Yu, X.; Arun, P.V.; Asano, Y.; Xiang, P.; Zhou, H. Hyperspectral video object tracking with cross-modal spectral complementary and memory prompt network. *Knowl.-Based Syst.* **2025**, *330*, 114595. <https://doi.org/10.1016/j.knosys.2025.114595>.
4. Ji, Y.; Song, K.; Wen, H.; Xue, X.; Yan, Y.; Meng, Q. UAV applications in intelligent traffic: RGBT image feature registration and complementary perception. *Adv. Eng. Inform.* **2025**, *63*, 102953. <https://doi.org/10.1016/j.aei.2024.102953>.
5. Li, R.; Xiao, C.; Yin, Q.; An, W.; Chen, N.; Ying, X.; Li, M.; Wang, Y. Dynamic High-Frequency Convolution for Infrared Small Target Detection. *IEEE Trans. Circuits Syst. Video Technol.* **2026**, *36*, 7676–7680. <https://doi.org/10.1109/TCSVT.2026.3661285>.
6. Zhao, D.; Asano, Y.; Gu, L.; Sato, I.; Zhou, H. City-Scale Distance Sensing via Bispectral Light Extinction in Bad Weather. *Remote Sens.* **2020**, *12*, 1401. <https://doi.org/10.3390/rs12091401>.
7. Jiang, Z.; Chen, J.; Yuan, Y. Structural consistency learning for unsupervised domain adaptive object detection. *Neural Netw.* **2025**, *191*, 107767. <https://doi.org/10.1016/j.neunet.2025.107767>.
8. Tao, H.; Yang, Z.; Liu, J.; Qiu, H.; Li, X.; Gao, L. Iterative model pruning with sparsity learning for infrared rotary-wing UAV detection. *Expert Syst. Appl.* **2026**, *295*, 128755. <https://doi.org/10.1016/j.eswa.2025.128755>.
9. Dong, H.; Wu, J.; Xing, C.; Xi, H.; Cui, H.; Zhu, J. Treasure in the background: Improve saliency object detection by self-supervised contrast learning. *Expert Syst. Appl.* **2025**, *267*, 126244. <https://doi.org/10.1016/j.eswa.2024.126244>.
10. Fang, X.; Wang, X.; Zhu, J.; Chen, Q.; Chen, Z. MGCNet: Multiple group-wise correlation network with hierarchical contrastive learning for co-salient object detection. *Knowl.-Based Syst.* **2025**, *309*, 112852. <https://doi.org/10.1016/j.knosys.2024.112852>.
11. Guo, Y.; Ma, L.; Luo, X.; Xie, S. DP-DDCL: A discriminative prototype with dual decoupled contrast learning method for few-shot object detection. *Knowl.-Based Syst.* **2024**, *297*, 111964. <https://doi.org/10.1016/j.knosys.2024.111964>.
12. Chen, H.; Shen, F.; Zhang, L.; Deng, Y.; Wang, P. Modality-balanced contrastive pre-training for RGB-D salient object detection. *Inf. Fusion* **2026**, *126*, 103576. <https://doi.org/10.1016/j.inffus.2025.103576>.
13. Chen, Q.; Xu, H.; He, W.; Chen, X.; Bian, M.; Gao, H. Segmentation-enhanced Medical Visual Question Answering with mask-prompt alignment using contrastive learning and multitask object grounding. *Eng. Appl. Artif. Intell.* **2026**, *163*, 112866. <https://doi.org/10.1016/j.engappai.2025.112866>.
14. Zhao, D.; Gu, L.; Qian, K.; Zhou, H.; Yang, T.; Cheng, K. Target Tracking from Infrared Imagery via an Improved Appearance Model. *Infrared Phys. Technol.* **2020**, *104*, 103116. <https://doi.org/10.1016/j.infrared.2019.103116>.
15. Lyu, X.; Tian, L.; Teng, S. ECDet: Efficient oriented object detection on the aerial image with cross-layer attention. *J. Real-Time Image Process.* **2025**, *22*, 42. <https://doi.org/10.1007/s11554-024-01617-3>.
16. Chen, H.; Chen, T.; Zhang, H. PointOOD: Point-supervised remote sensing oriented object detection via canonical feature fusion and dynamic radius sampling. *Digit. Signal Process.* **2026**, *168*, 105665. <https://doi.org/10.1016/j.dsp.2025.105665>.
17. Wu, L.; You, M.; Zhao, D.; Xu, X.; Yao, Y.; Pattathal Vijayakumar, A.; Cheng, K.; Xiang, P. DSFNet: Dual-stream spatial-spectral fusion network for hyperspectral anomaly detection. *Digit. Signal Process.* **2026**, *175*, 106019. <https://doi.org/10.1016/j.dsp.2026.106019>.
18. Ding, J.; Xue, N.; Long, Y.; Xia, G.S.; Lu, Q. Learning RoI Transformer for Oriented Object Detection in Aerial Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2019; pp. 2844–2853. <https://doi.org/10.1109/CVPR.2019.00296>.
19. Pan, X.; Ren, Y.; Sheng, K.; Dong, W.; Yuan, H.; Guo, X.; Ma, C.; Xu, C. Dynamic Refinement Network for Oriented and Densely Packed Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2020; pp. 11204–11213. <https://doi.org/10.1109/CVPR42600.2020.01122>.
20. Li, R.; An, W.; Xiao, C.; Li, B.; Wang, Y.; Li, M.; Guo, Y. Direction-Coded Temporal U-Shape Module for Multiframe Infrared Small Target Detection. *IEEE Trans. Neural Netw. Learn. Syst.* **2025**, *36*, 555–568. <https://doi.org/10.1109/TNNLS.2023.3331004>.
21. Han, J.; Ding, J.; Li, J.; Xia, G.S. Align Deep Features for Oriented Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5602511. <https://doi.org/10.1109/TGRS.2021.3062048>.
22. Li, R.; An, W.; Wang, Y.; Ying, X.; Dai, Y.; Wang, L.; Li, M.; Guo, Y.; Liu, L. Probing Deep into Temporal Profile Makes the Infrared Small Target Detector Much Better. *IEEE Trans. Pattern Anal. Mach. Intell.* **2026**, Early access. <https://doi.org/10.1109/TPAMI.2026.3683258>.
23. Yao, Y.; Wang, Q.; Zhao, D.; You, M.; Xiang, P.; Asano, Y.; Yu, X.; Wang, C.; Zhou, H.; Ren, J. DFBSNet: Dual frequency-domain branch fusion and selection network for hyperspectral anomaly detection. *Pattern Recognit.* **2026**, *180*, 113967. <https://doi.org/10.1016/j.patcog.2026.113967>.
24. Chen, Y.; Liu, Y.; Gao, Z.; Yan, J. A novel semi-supervised learning framework for large-scale imbalanced remote sensing object detection. *Eng. Appl. Artif. Intell.* **2026**, *163*, 113116. <https://doi.org/10.1016/j.engappai.2025.113116>.
25. Mahajan, D.; Girshick, R.; Ramanathan, V.; He, K.; Paluri, M.; Li, Y.; Bharambe, A.; van der Maaten, L. Exploring the Limits of Weakly Supervised Pretraining. *arXiv* **2018**, arXiv:1805.00932. <https://doi.org/10.48550/arXiv.1805.00932>.

26. Cui, Y.; Jia, M.; Lin, T.Y.; Song, Y.; Belongie, S. Class-Balanced Loss Based on Effective Number of Samples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2019; pp. 9260–9269. <https://doi.org/10.1109/CVPR.2019.00949>.
27. Zhang, H.; He, T.; Li, Z.; Deng, Y.; Feng, Y.; Yuan, D.; Yang, Y. E2IGB: Enhanced effective-information-guided class-balanced loss for long-tailed object recognition. *Pattern Recognit.* **2026**, *173*, 112756. <https://doi.org/10.1016/j.patcog.2025.112756>.
28. Wang, W.; Chen, Z.; Zhao, Z.; Su, F. PDL: Progressive Discriminative Feature Learning for Long-Tailed Recognition. *Neurocomputing* **2025**, *65*, 130463. <https://doi.org/10.1016/j.neucom.2025.130463>.
29. Qezelbash-Chamak, J.; Hicklin, K.; Kim, M. KANBalance: Kolmogorov-Arnold network mitigates class imbalance. *Pattern Recognit.* **2026**, *171*, 112325. <https://doi.org/10.1016/j.patcog.2025.112325>.
30. Kiobya, T.; Zhou, J.; Maiseli, B. A multi-scale semantically enriched feature pyramid network with enhanced focal loss for small-object detection. *Knowl.-Based Syst.* **2025**, *310*, 113003. <https://doi.org/10.1016/j.knosys.2025.113003>.
31. Wang, X.; Zhang, Y.; Bai, N.; Yu, Q.; Wang, Q. Class-imbalanced time series anomaly detection method based on cost-sensitive hybrid network. *Expert Syst. Appl.* **2024**, *238*, 122192. <https://doi.org/10.1016/j.eswa.2023.122192>.
32. Xue, Y.; Zhang, Y.; Liu, Y.; Qian, X. Overlapping object detection with adaptive Gaussian sample division and asymmetric weighted loss. *Knowl.-Based Syst.* **2024**, *293*, 111685. <https://doi.org/10.1016/j.knosys.2024.111685>.
33. Guo, C.; Huang, H. Enhancing camouflaged object detection through contrastive learning and data augmentation techniques. *Eng. Appl. Artif. Intell.* **2025**, *141*, 109703. <https://doi.org/10.1016/j.engappai.2024.109703>.
34. An, Y.; Sun, Y.; Li, Y.; Yang, Y.; Yu, J.; Zhu, Z. Sample imbalance remote sensing small target detection based on discriminative feature learning and imbalanced feature semantic enrichment. *Expert Syst. Appl.* **2025**, *273*, 126753. <https://doi.org/10.1016/j.eswa.2025.126753>.
35. Chen, H.; Chen, Z.; Wu, Y.; Chen, H. Spatial-Aware Multi-Modal Contrastive Learning for RGB-D Salient Object Detection and Beyond. *Inf. Fusion* **2025**, *124*, 103362. <https://doi.org/10.1016/j.inffus.2025.103362>.
36. Juyal, D.; Shingi, S.; Javed, S.A.; Padigela, H.; Shah, C.; Sampat, A.; Khosla, A.; Abel, J.; Taylor-Weiner, A. SC-MIL: Supervised Contrastive Multiple Instance Learning for Imbalanced Classification in Pathology. *arXiv* **2023**, arXiv:2303.13405. <https://doi.org/10.48550/arXiv.2303.13405>.
37. Zhang, Y.; Chen, X.; Lai, J.; Zhang, B.; Chen, Z.; Jiang, F. Global-local contrastive learning: A multi-operating-condition guided approach for few-shot cross-domain bearing fault diagnosis. *Eng. Appl. Artif. Intell.* **2026**, *165*, 113375. <https://doi.org/10.1016/j.engappai.2025.113375>.
38. Li, Y.; Zhao, X.; Zeng, Y.; Feng, H.; Chai, J.; Xie, H.; Fu, S.; Zhang, S. Node transfer with graph contrastive learning for class-imbalanced node classification. *Neural Netw.* **2025**, *190*, 107674. <https://doi.org/10.1016/j.neunet.2025.107674>.
39. You, M.; Yao, Y.; Zhao, D.; Zhao, Z.; Arun, P.V.; Wang, Y.; Zhou, H.; Chi, R. S3CRAD: Superpixel-guided background inpainting and spatial-spectral constrained representation for hyperspectral anomaly detection. *Opt. Lasers Eng.* **2026**, *201*, 109657. <https://doi.org/10.1016/j.optlaseng.2026.109657>.
40. Wu, Z.; Xiong, Y.; Yu, S.; Lin, D. Unsupervised Feature Learning via Non-Parametric Instance-level Discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2018; pp. 3733–3742. <https://doi.org/10.1109/CVPR.2018.00393>.
41. Gupta, A.; Dollár, P.; Girshick, R. LVIS: A Dataset for Large Vocabulary Instance Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2019; pp. 5351–5359. <https://doi.org/10.1109/CVPR.2019.00550>.
42. Girshick, R. Fast R-CNN. In *Proceedings of the International Conference on Computer Vision, Santiago, Chile, 2015*; IEEE: Piscataway, NJ, USA, 2015; pp. 1440–1448. <https://doi.org/10.1109/ICCV.2015.169>.
43. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In *Proceedings of the International Conference on Computer Vision*; IEEE: Piscataway, NJ, USA, 2017; pp. 2999–3007. <https://doi.org/10.1109/ICCV.2017.324>.
44. Huang, Z.; Li, W.; Xia, X.G.; Wang, H.; Jie, F.; Tao, R. LO-Det: Lightweight Oriented Object Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5603515. <https://doi.org/10.1109/TGRS.2021.3067470>.
45. Xiao, Z.; Wang, K.; Wan, Q.; Tan, X.; Xu, C.; Xia, F. A2S-Det: Efficiency Anchor Matching in Aerial Image Oriented Object Detection. *Remote Sens.* **2021**, *13*, 73. <https://doi.org/10.3390/rs13010073>.
46. Xu, Z.; Wang, X.; Huang, K.; Chen, R. Low resolution remote sensing object detection with fine grained enhancement and swin transformer. *Sci. Rep.* **2025**, *15*, 24183. <https://doi.org/10.1038/s41598-025-10286-6>.
47. Wang, Z.; Li, C.; Xu, H.; Zhu, X.; Li, H. Mamba YOLO: A Simple Baseline for Object Detection with State Space Model. *arXiv* **2024**, arXiv:2406.05835. <https://doi.org/10.48550/arXiv.2406.05835>.
48. Zhu, Y.; Hong, J.; Wang, T. Weakly Supervised Oriented Object Detection in Remote Sensing via Geometry-Aware Enhancement Network. *Remote Sens.* **2026**, *18*, 1253. <https://doi.org/10.3390/rs18081253>.
49. Zeng, X.; Fang, S.; Sun, Q. BFRI-YOLO: Harmonizing Multi-Scale Features for Precise Small Object Detection in Aerial Imagery. *Electronics* **2026**, *15*, 297. <https://doi.org/10.3390/electronics15020297>.

50. Yang, F.; Chen, G.; Duan, J. Skip-Encoder and Skip-Decoder for Detection Transformer in Optical Remote Sensing. *Remote Sens.* **2024**, *16*, 2884. <https://doi.org/10.3390/rs16162884>.
51. Lyu, C.; Zhang, W.; Huang, H.; Zhou, Y.; Wang, Y.; Liu, Y.; Zhang, S.; Chen, K. RTMDet: An Empirical Study of Designing Real-Time Object Detectors. *arXiv* **2022**, arXiv:2212.07784. <https://doi.org/10.48550/arXiv.2212.07784>.
52. Li, F.; Zhang, H.; Liu, S.; Guo, J.; Ni, L.M.; Zhang, L. DN-DETR: Accelerate DETR Training by Introducing Query DeNoising. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 2239–2251. <https://doi.org/10.1109/TPAMI.2023.3335410>.
53. Su, J.; Chang, Y.H.; Yang, M.F.; Zhang, B.Y.; Meng, J.T. The light-weight RT-DETR improvement algorithm. *Microelectron. Comput.* **2025**, *42*, 103–113. <https://doi.org/10.19304/J.ISSN1000-7180.2024.0502>.
54. Wang, C.; He, W.; Nie, Y.; Guo, J.; Liu, C.; Han, K.; Wang, Y. Gold-YOLO: Efficient Object Detector via Gather-and-Distribute Mechanism. *arXiv* **2023**, arXiv:2309.11331. <https://doi.org/10.48550/arXiv.2309.11331>.
55. Jin, Z.; Duan, J.; Qiao, L.; He, T.; Shi, X.; Yan, B. MTGS-YOLO: A task-balanced algorithm for object detection in remote sensing images based on improved YOLO. *J. Supercomput.* **2025**, *81*, 542. <https://doi.org/10.1007/s11227-025-07003-5>.
56. Li, H.; Qu, H. DASSF: Dynamic-Attention Scale-Sequence Fusion for Aerial Object Detection. *arXiv* **2024**, arXiv:2406.12285. <https://doi.org/10.48550/arXiv.2406.12285>.
57. Li, Z.; Hu, X.; Qian, J.; Zhao, T.; Xu, D.; Wang, Y. Self-Supervised Feature Contrastive Learning for Small Weak Object Detection in Remote Sensing. *Remote Sens.* **2025**, *17*, 1438. <https://doi.org/10.3390/rs17081438>.

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.