remote sensing

MDPI

Article

Relevance Pooling Guidance and Class-Balanced Feature Enhancement for Fine-Grained Oriented Object Detection in Remote Sensing Images

Yu Wang, Hao Chen *, Ye Zhang and Guozheng Li

Department of Information Engineering, Harbin Institute of Technology, Harbin 150000, China; 20b905065@stu.hit.edu.cn (Y.W.); zhye@hit.edu.cn (Y.Z.); 23b905059@stu.hit.edu.cn (G.L.)
* Correspondence: hit_hao@hit.edu.cn

Abstract: Fine-grained object detection in remote sensing images is highly challenging due to class imbalance and high inter-class indistinguishability. The strategies employed by most existing methods to resolve these two challenges are relatively rudimentary, resulting in suboptimal model performance. To address these issues, we propose a fine-grained object-oriented detection method based on relevance pooling guidance and class balance feature enhancement. Firstly, we propose a global attention mechanism that dynamically retains spatial features pertinent to objects through relevance pooling during down-sampling, thereby enabling the model to acquire more discriminative features. Next, a class balance correction module is proposed to alleviate the class imbalance problem. This module employs feature translation and a learnable reinforcement coefficient to highlight the boundaries of tail class features while maintaining their distinctiveness. Furthermore, we present an enhanced contrastive learning strategy. By dynamically adjusting the contribution of inter-class samples and intra-class similarity measures, this strategy not only constrains inter-class feature distances but also facilitates tighter intra-class clustering, making it more suitable for imbalanced datasets. Evaluation on the FAIR1M and MAR20 datasets demonstrates that our method is superior compared to other methods in object detection and achieves 46.44 and 85.05% mean average precision, respectively.

Keywords: remote sensing; fine-grained object detection; relevance pooling; class balance correction; enhanced contrast learning

check for updates

Citation: Wang, Y.; Chen, H.; Zhang, Y.; Li, G. Relevance Pooling Guidance and Class-Balanced Feature Enhancement for Fine-Grained Oriented Object Detection in Remote Sensing Images. *Remote Sens.* **2024**, *16*, 3494. https://doi.org/10.3390/rs16183494

Academic Editor: Jocelyn Chanussot

Received: 31 July 2024
Revised: 13 September 2024
Accepted: 18 September 2024
Published: 20 September 2024

Copyright: © 2024 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

Fine-Grained Object Detection (FGOD) aims to automatically locate and distinguish highly confusable subcategories (e.g., different models of aircraft or types of automobiles) with subtle differences within a single category. With the advancement of high-resolution remote sensing satellites, the spatial information in Remote Sensing Image (RSI) has become increasingly rich and the texture details more distinct, which provides favorable conditions for FGOD.

Fine-grained classes typically exhibit smaller inter-class differences and lower intra-class similarity, posing a significant challenge to existing object detectors. Recently, deep learning-based techniques have led to significant breakthroughs in the field of computer vision, especially in fine-grained object classification within natural images. Current methods for fine-grained object classification can generally be divided into two types. Region-based localization uses a detection subnet to locate discriminative regions and, subsequently, detect objects based on these identified regions [1,2]. Given the challenges in representing or even defining discriminative regions for certain object classes, feature encoding methods that extract fine-grained features through the encoding of highly parameterized feature representations [3,4] have achieved favorable results.

However, the performance of most methods is often significantly reduced when applied to RSIs, primarily due to the inherent differences between RSIs and natural scene

images, as illustrated in Figure 1. Specifically: (1) The aerial perspective of RSIs often leads to severe occlusions and a reduction in discriminative regions. (2) Objects in RSIs demonstrate considerable scale diversity and arbitrary orientations, and occupy fewer pixels compared to those in natural images. (3) Due to factors such as brightness and resolution, RSIs are challenged by small inter-class differences and large intra-class variability. These issues introduce additional complexities to current detection methods.

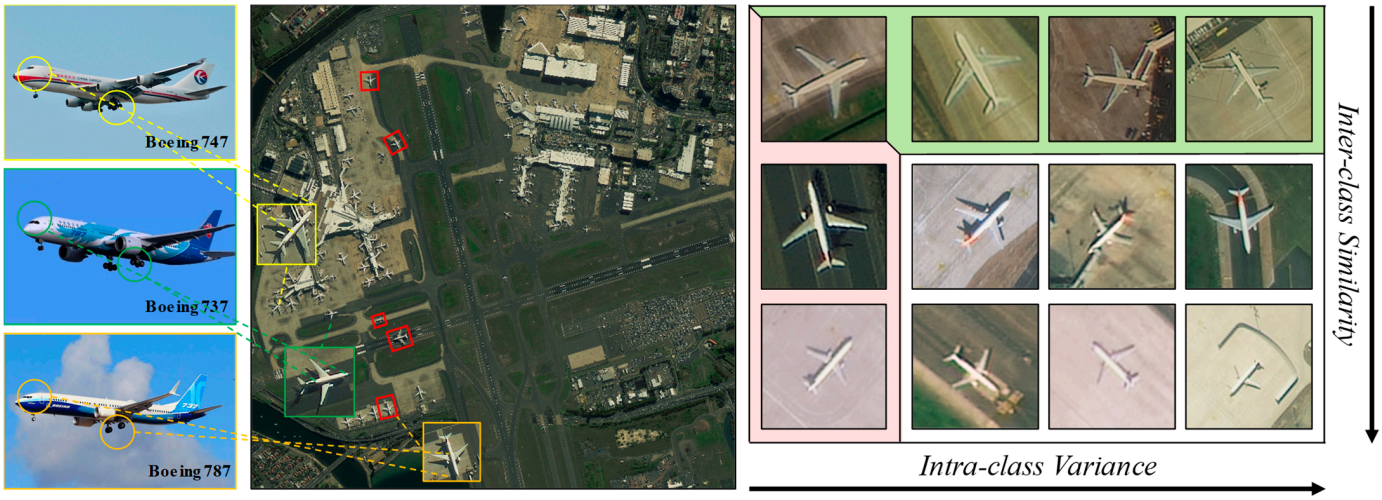


Figure 1. The characteristics of objects in RSI render FGOD tasks particularly challenging. (i) Severe occlusions (e.g., aircraft tires are not visible in the RSI, as shown in the left image). (ii) Diverse scales and orientations with smaller pixel coverage (marked by red boxes in the figure). (iii) Small inter-class differences and large intra-class variability (e.g., large variations among same-model aircraft in the right green row, and similar appearances among different-model aircraft in the pink column).

The most intuitive approach to addressing these challenges is to optimize discriminative feature representations in the feature space, thereby facilitating the extraction of subtle differences between fine-grained classes. Han et al. [5] proposed an information reuse network that combines dense feature fusion and dual mask attention to enhance inter-class differences, and they introduce super-classes to explore intra-class relationships. Chen et al. [6] integrated a dual-branch network with an ensemble module to decorrelate images and aggregate them into subclasses, thereby alleviating intra-class variability and inter-class similarity issues. Ouyang et al. [7] introduced a multi-granularity self-attention network that leverages both global and local features to explore fine-grained information and improve classification performance. These methods learn more discriminative feature representations through detectors while performing oriented detection to reduce redundancy in background regions, thereby enabling fine-grained object detection. Despite the improvements in object feature representation achieved by the aforementioned methods, the issue of high inter-class indistinguishability remains unresolved due to the lack of explicit constraints on inter-class feature distances.

In addition to the discriminative capability of networks, class imbalance is also a critical factor limiting the performance of existing FGOD models for RSIs. During the model learning process, the scarcity of reference labels for certain classes leads the network parameters to be predominantly optimized based on the losses of the majority classes (head classes), resulting in significantly decreased detection precision for minority classes (tail classes). Many studies address this issue by either resampling the training data or adjusting the weights of minority groups [8–10]. However, these methods may lead to decreased detection performance due to various reasons. On one hand, the limited samples in the tail classes may result in the reweighting methods overfitting these samples or continued bias towards head class samples, failing to achieve the intended balance. On the other hand, head classes often exhibit more complex semantic substructures, such as multiple

high-density regions in data distribution. Therefore, simply reducing head class samples and treating them equally can easily lead to the loss of crucial information.

To address these issues, we propose a Fine-Grained Object-Oriented Detection (FGO^{2}D) method based on relevance pooling guidance and class-balanced feature enhancement. We propose a Relevance-based Global Attention Mechanism (RGAM) to facilitate the model's acquisition of more discriminative features. Relevance pooling is employed to replace the down-sampling operation in the global attention mechanism, which dynamically weights high-rank feature components by capturing multi-dimensional channel information, thereby retaining spatial features relevant to the object and preventing the loss of detailed information typically associated with down-sampling pooling. In addition, a Category Balance Correction (CBC) module is proposed. The CBC module leverages feature translation to highlight the boundaries of the tail class feature space, and a learnable feature reinforcement coefficient is proposed to highlight the distinctiveness of tail class features, thereby alleviating the issue of class imbalance. Finally, an Enhanced Contrastive Learning (ECL) approach is proposed to impose spatial constraints on intra-class and inter-class distances. Specifically, similarity measure control weights and class balance coefficients are incorporated to optimize the contrastive learning method, thereby improving its suitability for imbalanced datasets. The contributions are summarized as follows:

1. We propose RGAM that replaces traditional pooling with relevance pooling. This approach mitigates the loss of details caused by down-sampling and enhances the feature representation.
2. We propose a CBC strategy that directly operates within the feature space to highlight the boundaries of tail class features and utilizes a learnable reinforcement mechanism to dynamically amplify these features, alleviating the long-tail problem.
3. We propose an ECL model that constrains inter-class distances while increasing intra-class compactness, thereby significantly enhancing the effectiveness of contrastive learning for imbalanced datasets.

The remainder of this paper is organized as follows: Section 2 reviews related work. Section 3 elaborates on the principles of the proposed method. Section 4 presents experimental evaluations and analysis. Finally, Section 5 summarizes our findings in this work.

2. Related Work

2.1. Fine-Grained Oriented Object Detection in RSI

In recent years, significant advancements have been achieved in fine-grained object classification in natural images, as exemplified by methods such as Cross-X [11] and P-CNN [12]. This paper focuses on studying FGO^{2}D in RSIs.

In contrast to natural images, the objects are often small in size and exhibit diverse orientations in RSIs. Consequently, FGO^{2}D needs to first achieve accurate localization of the objects to minimize extensive background noise and overlap between objects. This precision is crucial for enhancing the fine-grained representation capabilities of objects. It has been found that two-stage models are often more effective when using Oriented Bounding Box (OBB) for object detection. In this work, we construct a two-stage object detector that employs an oriented region proposal network to generate oriented proposals, thereby mitigating the interference of background noise.

In addition to the challenges in oriented object detection, FGO^{2}D also faces issues of high intra-class variance and low inter-class discrimination during fine-grained classification. Existing methods predominantly employ visual attention mechanisms or feature encoding techniques to enhance the feature representation capabilities of models. Cheng et al. [13] proposed a network with independent feature refinement, wherein two CNN-based branches execute task-specific feature refinements to achieve fine-grained classification and oriented localization. Song et al. [14] proposed a refined balanced feature pyramid network, which integrates features from different layers to provide high-quality semantic information for FGO^{2}D. Zhou et al. [15] introduced attention-based group feature enhancement and sub-significant feature learning to drive deep feature representation

for fine-grained classification. Zeng et al. [16] proposed a prototype detector based on contrastive learning, which optimizes feature space distribution to increase inter-class distances and reduce intra-class distances, thereby enhancing the discriminative capability of fine-grained objects.

From the aforementioned studies, it is evident that FGOD methods require the extraction of more granular details for each object. Therefore, in our method, we consider integrating attention mechanisms and contrastive learning to fully extract fine-grained features for each object and constrain inter-class feature distances. This approach enhances the discriminative ability of the model.

2.2. Discriminant Feature

In fine-grained classification, effective feature discrimination can enhance the inter-class separability of the model, leading to more compact feature representations for samples within the same class and more dispersed features for samples across different classes. Methods employed to achieve this include contrastive learning [17], attention mechanisms [18], and loss strategies [19], among others.

Attention mechanisms can assist in extracting more effective fine-grained features, improving the discriminative power and feature representation capabilities of the model, thereby effectively addressing the issue of high inter-class indistinguishability. Bao et al. [20] proposed a sparse attention mechanism that adaptively aggregates discriminative local features across multiple scales using fine-grained attention, aiming to improve the classification of lower-level aircraft. Zhu et al. [21] proposed a dual cross-attention learning algorithm to better learn subtle feature embeddings for recognizing fine-grained objects. Wang et al. [22] proposed a multi-scale attention network to embed low-layer texture features with high-level semantic features to train the re-identification model. In this work, we extended the global attention model and enhanced the interaction between global and local information through a hybrid attention module.

Contrastive learning can further optimize discriminative features by maximizing inter-class distance and minimizing intra-class distance. For example, Li et al. [23] proposed a dual similarity network that employs two similarity metrics to learn more discriminative and similarity-biased features. Sun et al. [24] trained a supervised contrastive head on proposal features to promote instance-level compactness and inter-class variance. While these methods focus on increasing inter-class variance and intra-class consistency of feature representations under certain metrics, they overlook the impact of positive and negative sample imbalance on learning. In this work, we further refine the learning model to ensure that, under conditions of sample imbalance, the features of samples within the same class are more compact, while those of different classes are more dispersed.

2.3. Class Imbalanced Learning

An intuitive solution to addressing class imbalance is to rebalance the data distribution, which can be achieved through two main approaches: resampling [25,26] and reweighting [27,28].

The resampling methods aim to achieve a more balanced data distribution by rebalancing the class prior distributions during training. These methods primarily employ oversampling (promoting the replication of minority samples) [29], sampling techniques (focusing on the reduction in majority samples) [30], or a combination of both [31]. However, repeatedly accessing the limited samples in tail classes can easily lead to model overfitting, resulting in suboptimal learning of the classifier or feature representations. On the other hand, the head classes often possess more complex semantic substructures, so simply reducing the number of head class samples and treating them as equal can easily lead to the loss of critical information.

The reweighting method balances class distribution by adjusting the weights of training instances in the loss function. Specifically, it assigns higher weights to tail classes and lower weights to head classes. Within reweighting methods, there are cost-sensitive learn-

ing (with higher cost for misclassifying minority samples) [32], boundary transfer methods (which push decision boundaries towards head classes) [33], and classifier ensembles [34]. However, research has found that reweighting methods struggle to handle large-scale long-tail data, often resulting in optimization difficulties [35].

In addition, existing research has also explored other methods to address class imbalance, including transfer learning [36], self-supervised learning [37], metric learning [38], and data completion [39], which have been proven to be effective in long-tailed learning scenarios.

In this work, we highlight the feature boundaries and enhance the representations of tail classes within the feature space, avoiding head class dominance and ensuring effective learning of tail class features.

3. Methods

The proposed method is shown in Figure 2. In the feature extraction stage, we propose RGAM, which uses relevance pooling to retain spatial information relevant to the object and employs channel attention to learn global feature responses. RGAM, along with local attention units, are embedded into a hybrid attention module to extract more discriminative fine-grained features. Additionally, we propose a CBC module that effectively highlights the boundaries of tail class features through feature translation and learns feature reinforcement to highlight the distinctiveness of tail class features. In the feature learning stage, an ECL approach is proposed, which constrains inter-class features while promoting intra-class features to form tighter clusters in the feature space.

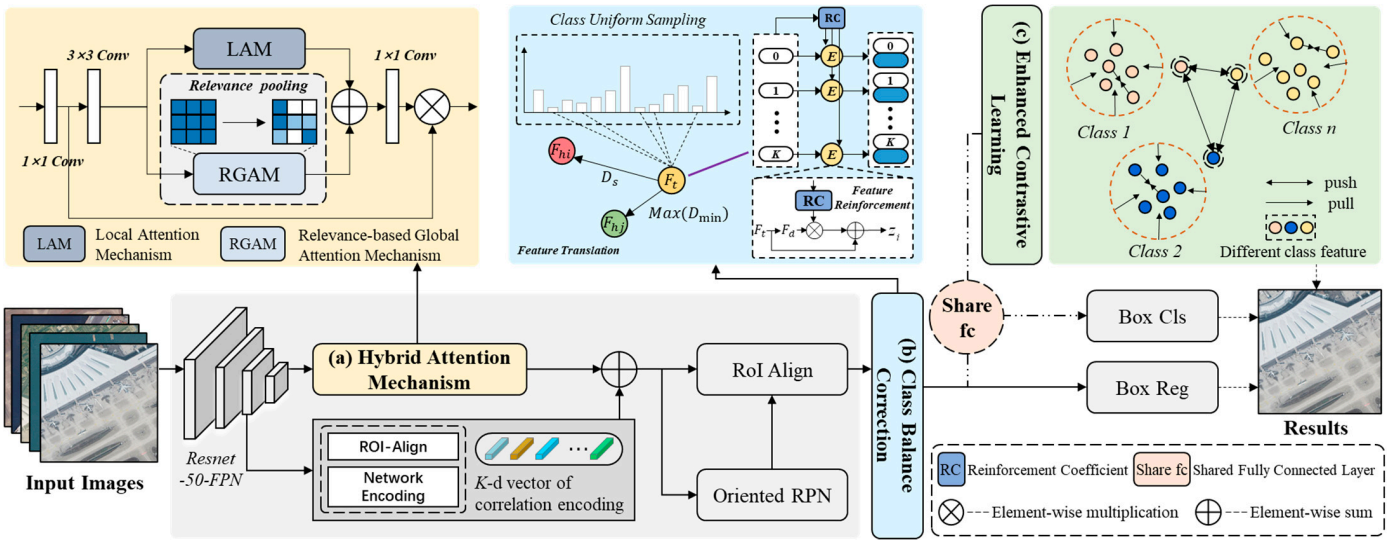


Figure 2. Framework of proposed method. The method is a two-stage fine-grained object detector: (**a**) Hybrid Attention Mechanism: retains correlated features and fuses local–global features to enhance feature representation. (**b**) Class Balance Correction: highlights feature boundaries and reinforces distinctive features. (**c**) Enhanced Contrastive Learning: constrains intra-class and inter-class feature distances.

3.1. Relevance-Based Global Attention Mechanism

Most existing attention mechanisms employ spatial pooling down-sampling to compress spatial information, thereby achieving different scales of receptive fields and reducing memory consumption. However, the features retained by conventional pooling operations may not necessarily be discriminative for FGOD, leading to the loss of details and potentially resulting in suboptimal learning models. This effect is particularly pronounced for small-sized objects in RSIs.

To meet the pooling operation requirements of the network framework, we propose a global attention mechanism embedded with an object-relevance pooling layer, termed RGAM. Our motivation is driven by the inherent loss of information during the feature

pooling process, which necessitates that careful consideration be given to which features are sampled, ensuring that they have sufficient discriminative power. In this work, spatial information is conditionally sampled based on the importance and richness of the feature map information, with global and local attention integrated to enhance the fine-grained feature representation of the network.

The output of convolutional layers typically consists of a high-dimensional feature map containing information at various levels. For the input feature map $f(I)$, we retain the high-rank components of the features based on the concept of singular value decomposition. This step aims to select and retain the feature components that encapsulate the most informative content. The high-rank feature map $\mathbf{f}_{\text{high},\,k}(I)$ is defined as:

$$\mathbf{f}_{\text{high},\,k}(I) = \text{reshape}(\mathbf{U}_k \cdot \text{diag}([\sigma_1, \sigma_2, \ldots, \sigma_k]) \cdot \mathbf{V}_k^T, C, H, W) \tag{1}$$

$$k = \min\left\{k \middle| \left(\sum_{l=1}^{k} \sigma_l / \sum_{l=1}^{n} \sigma_l\right) \geq v\right\} \tag{2}$$

where k represents the number of retained high-rank components, n is the rank of the feature map, and the threshold v is set to 0.95. The partial singular vectors $\mathbf{U}_k$ and $\mathbf{V}_k$ are computed based on the top k singular values, and parameters C, H, and W represent the number of channels, height, and width of the feature map, respectively. Through singular value decomposition, the most informative portions of the feature map $f(I)$ are effectively retained by the model, thereby mitigating noise and irrelevant information to a certain extent.

Although the selected high-rank components are distinctive, not every component is necessarily significant for detection. Our objective is to aggregate useful information to enhance the ability of the model to leverage and recognize critical features. The deep multi-dimensional channels corresponding to a high-rank component are interrelated. After performing global average pooling on these multi-dimensional channels, the resulting average value $z_{c,k}$ can represent the richness of a high-rank component. Based on the computed average information content, an anti-relevance weighting coefficient $\alpha_{c,k}$ is defined, reflecting the relative importance of information contained in corresponding deep multi-dimensional channels. Generally, a smaller coefficient $\alpha_{c,k}$ is associated with high-rank components that exhibit a higher degree of information richness. The anti-relevance weighting function $\alpha_{c,k}$ is defined as follows:

$$\alpha_{c,k} = 1 - \frac{z_{c,k} - \min(z_{c,k})}{\max(z_{c,k}) - \min(z_{c,k})} \tag{3}$$

where $\max(z_k)$ and $\min(z_k)$ denote the maximum and minimum values of the global average pooling across all deep multi-dimensional channels, respectively. The anti-relevance weighting function $\alpha_{c,k}$ dynamically adjusts the importance of various high-rank features during the feature representation process, enabling a more precise capture of the contribution of different high-rank components to the overall task, thereby enhancing the robustness and generalization capability of the model.

Finally, each high-rank component $\mathbf{f}_{\text{high},\,k}(I)$ is combined with the corresponding $\alpha_{c,k}$, focusing on features that are highly relevant to the object by weighting the high-rank feature maps. This process is aimed at retaining and enhancing the features that are relevant to the object, thereby improving the ability of the model to perceive key features. The relevance pooling is defined as:

$$\mathbf{f}_{\text{final}}(\mathbf{I}) = \sum_{c,k} \alpha_{c,k} \cdot \mathbf{f}_{\text{high},k,c}(\mathbf{I}) \tag{4}$$

By weighting the high-rank feature maps, the anti-relevance weighting function α_k effectively controls the importance of features. Our pooling model places greater emphasis on object-relevant features, reducing the sensitivity of the model to noise and irrelevant information, thereby mitigating the risk of overfitting and enhancing the robustness and

reliability of the model. It is worth noting that our method is agnostic to specific data structures or model architectures, making it applicable to various domains and tasks.

Due to the relatively small size of objects in RSI, both global and local information are equally important for FGOD. Retaining useful spatial information aids in the extraction of more discriminative features through global attention. Therefore, the relevance pooling module is integrated into the global attention unit to replace global average pooling, and the RGAM is further integrated with the local attention unit within the hybrid attention mechanism, as illustrated in Figure 3. The hybrid attention module, which combines global and local attention units, is stacked into the deep network, dynamically generating attention matrices based on the local and global information collected within the module, thereby enhancing the representational power of the network.

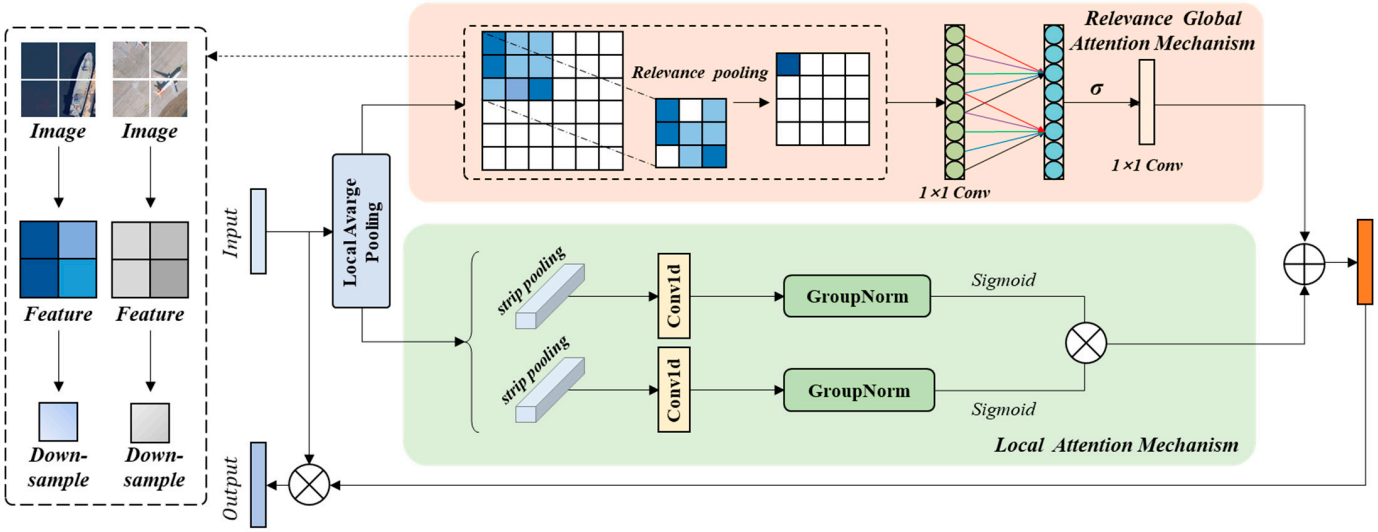


Figure 3. Hybrid attention mechanism: The mechanism consists of a blend of RGAM and local attention units, aimed at endowing the network with enhanced representational power.

The deep features may lose some spatial information in the feature extraction channels. Initially, we employ bilinear interpolation to aggregate low-level features corresponding to the proposals. Subsequently, these low-level features are encoded by the network encoder for each scale and concatenated into a fixed-dimensional feature vector. We fuse low-level features and deep semantic features along the feature dimension axis, resulting in deep fusion features that encapsulate rich semantic and spatial information shared by both the RPN and the detection network.

3.2. Class Balance Correction

In FGOD for RSIs, the long-tail problem is pronounced due to the scarcity of reference labels for tail classes. Consequently, model parameters are predominantly optimized based on head class losses, leading to the dominance of head class features and insufficient learning of tail class features. Moreover, since tail class samples often share some common characteristics with similar head class samples, during the learning process, tail class features tend to shift along the feature axis towards the feature space of these similar head classes, ultimately being learned as those similar head classes.

From the perspective of feature correction, we propose a novel CBC strategy. We translate tail class features along inter-class variances to optimize their boundaries in the feature space. Additionally, the distinctiveness of tail class features is enhanced through feature reinforcement, with the aim of minimizing the dominance effect of head class features.

Firstly, feature vectors for the tail and head class samples are extracted using a pre-trained backbone network. Let $\mathbf{f}_{ti}$ represent the feature vector for the i-th tail class sample,

and $\mathbf{f}_{hj}$ represent the feature vector for the j-th head class sample. We define the feature space distance $\mathbf{d}(\mathbf{f}_{ti}, \mathbf{f}_{hj})$ between tail class $\mathbf{f}_{ti}$ and head class $\mathbf{f}_{hj}$, as follows:

$$\mathbf{d}(\mathbf{f}_{ti}, \mathbf{f}_{hj}) = \left\| \mathbf{f}_{ti} - \mathbf{f}_{hj} \right\|_2^2, t \in S_t, h \in S_h \tag{5}$$

where S_h and S_t denote the sample sets for the head classes and tail classes, respectively. In FGOD, subcategories often share multiple common features with subtle differences. Consequently, in the feature space, a tail class may exhibit shifts towards the features of multiple head classes, implying similarities with several head classes. To mitigate feature shifts, we perform feature translation on tail class features within the feature space. Our core idea is to establish inter-class feature distance relationships to define the translation loss for each tail class, ensuring clarity of boundaries within tail class feature spaces. This approach aims to highlight boundaries between tail classes and similar head classes, thereby preventing feature shifts. For each tail class feature $\mathbf{f}_{ti}$, a translation loss function is learned to adjust its position in the feature space, which is formulated as follows:

$$L_{\mathrm{Cbc}} = \frac{1}{\min \mathbf{d}(\mathbf{f}_{ti}, \mathbf{f}_{hj}) + \frac{1}{\rho} \sum \mathbf{d}(\mathbf{f}_{ti}, \mathbf{f}_{hj})} \tag{6}$$

where ρ is a hyperparameter used to balance the overall distance and the minimum distance. This process involves an iterative optimization step governed by the translation loss function. During this process, our goal is to increase the distance between the feature representation of the tail class and those of several similar head classes to reduce confusion and improve classification accuracy. Simultaneously, we focus on maintaining the maximum possible distance between the tail class and the most similar head class to ensure that the tail class is distinctly separated from the closest competing class. This strategy is crucial for improving the detection performance of underrepresented or minority classes in fine-grained object detection tasks. This configuration facilitates the establishment of clear boundaries between minority and head classes, thereby effectively differentiating the feature space relationships among different classes.

Secondly, due to the scarcity of training samples in tail classes, feature extractors often learn incomplete representations. To address this issue, tail class features are selectively enhanced based on inter-class similarity, aiming to accentuate their distinctiveness and ensure effective learning of tail class features by the model. We first define the distinctive features $\mathbf{F}_{di}$ of tail class relative to other head classes as follows:

$$\mathbf{F}_{di} = \frac{1}{m} \sum_{\mathbf{f}_{hj} \in \{g \mid \|\mathbf{f}_{ti} - g\|_2^2 \leq d_{av}\}} \mathbf{d}(\mathbf{f}_{ti}, \mathbf{f}_{hj}) \tag{7}$$

where d_{av} represents the average spatial distance between samples of the tail class and all head class samples, and g denotes the collection of head classes feature sets for samples that meet the specified conditions $\|\mathbf{f}_t - \mathbf{f}_h\|_2^2 \leq d_{av}$. This average spatial distance d_{av} serves as a criterion: When the spatial distance between a tail class sample and a head class sample falls below this threshold, it indicates that tail class features exhibit greater similarity to those of the head class. In such instances, reinforcing tail class features becomes essential. The distinctiveness of tail class features $\mathbf{f}_{ti}$ relative to $\mathrm{j}-\mathrm{th}$ individual head class features is defined as the feature dissimilarity between these two sets of features. Ultimately, the distinctive feature $\mathbf{F}_{di}$ is quantified as the feature dissimilarity between the tail class features $\mathbf{f}_{ti}$ and the multiple similar head class features that meet the threshold condition.

For each tail class, a reinforcement coefficient is computed to enhance the original features $\mathbf{f}_{ti}$. This coefficient is determined by two key factors: the frequency of occurrence of the class within the dataset, which reflects the proportion of samples belonging to this class in the overall dataset, and the degree of similarity between this class and the head classes. Specifically, when the frequency of the tail class is low and its similarity to the

head classes is high, the reinforcement coefficient is amplified, thereby enhancing the distinctiveness of the tail class features and preventing them from being overshadowed by head class features during training. In this way, the reinforcement coefficient ensures a clear representation of tail class features within the feature space, facilitating more effective learning and recognition of tail classes by the model when handling imbalanced datasets. On this basis, the enhanced features are defined as follows:

$$\mathbf{z}_i = \mathbf{f}_{ti} + \frac{N_{\max} - n_i}{N_{\max} - N_{\min}} \cdot \partial_i \cdot \mathbf{F}_{di} \tag{8}$$

$$\partial_i = \frac{1}{\Gamma} \sum_{Q} \max(d_{av} - \mathbf{d}(\mathbf{f}_{ti}, \mathbf{f}_{hj}), 0) \tag{9}$$

where $(N_{\max} - n_i)/(N_{\max} - N_{\min}) \cdot \partial_i$ denotes the reinforcement coefficient, and parameters $N_{\max}$ and $N_{\min}$, respectively, represent the number of the most frequently occurring and the rarest classes in the training dataset. Parameter n_i is the number of samples in the $\mathrm{i} - \mathrm{th}$ class, and Q represents all the classes in the training set. Variable Γ represents the number of head classes that satisfy $\mathbf{d}(\mathbf{f}_{ti}, \mathbf{f}_{hj}) - d_{av} \leq 0$. From Equation (7), it is evident that if the overall similarity between a tail class and all head classes is low, the parameter ∂ approaches 0 and the tail class features can largely be retained, which helps prevent them from being updated to resemble irrelevant head classes. Conversely, when the overall similarity between a tail class and most head classes is high and the sample size is small, the distinctiveness of the tail class features is significantly amplified. The distinctiveness of the tail class features is thus retained by this amplification, thereby assisting the classifier in learning the features effectively.

As outlined above, the CBC method delineates the boundaries of tail class sample features and enhances the independence of tail class features. This methodology suppresses the shift of tail class features towards similar head classes, enhances the representation of tail class features, and mitigates the dominance of head classes. Consequently, this ensures that the model effectively learns the tail class features.

3.3. Enhanced Contrast Learning

Fine-grained objects often exhibit high feature similarity. Our goal is to increase intra-class similarity and maximize inter-class separation. Contrastive learning extracts discriminative features by maximizing inter-class distances and minimizing intra-class distances. Inspired by this approach, we incorporate contrastive learning into supervised learning to enhance inter-class separability in FGOD models for RSI.

In self-supervised contrastive learning, due to the lack of class labels, positive and negative samples are generated from the image itself using various data augmentation techniques. All augmented versions of different images are considered 'negative' samples, which hinders the training of the model. Given that labels in the training phase of supervised object detection tasks are accessible, we integrate contrastive learning into supervised learning. Specifically, cosine similarity is employed to measure the similarity between input features and the features of other samples, with the inter-class feature distribution being optimized by maximizing the discriminative power between different samples. The contrastive learning loss function for the training batch is formulated as follows:

$$L_{\mathrm{Cl}} = \frac{1}{N} \sum_{i=1}^{I} \frac{-1}{\left|\widetilde{P}_i\right|} \log \frac{\sum_{p \in \widetilde{P}_i} \exp(\mathrm{sim}(\mathbf{z}_i, \mathbf{z}_p^+)/\tau)}{\sum_{a \in A(i)} \exp(\mathrm{sim}(\mathbf{z}_i, \mathbf{z}_a)/\tau)} \tag{10}$$

where $\widetilde{P}_i$ denotes the set of all positive samples in the batch that are different from i, sim represents the function used to compute the similarity between two samples, and temperature parameter τ is a scalar temperature parameter set to 0.1. The set $A(i)$ is the set of samples excluding i itself, parameter N denotes the number of samples in the batch, and I is the set of samples contained within the batch. Within this module, each input feature

z_i is encouraged to be closer to the corresponding positive samples while being distanced from samples of other classes.

Due to class imbalance, the number of negative samples for tail classes far exceeds that of positive samples. When there is a severe imbalance between positive and negative samples, the use of $\widetilde{P}_i$ as a coefficient may cause the model to favor the head classes and disregard the tail classes. In such cases, while the model imposes constraints on inter-class distances, it may disregard the attraction of intra-class distances, resulting in less compact feature representations within the same class. Moreover, if the similarity between samples in $\widetilde{P}_i$ and negative samples is insufficiently differentiated, the gradients may become very small, resulting in vanishing gradient problems. This issue can slow down the learning process of the model and potentially cause the model to converge to a local optimum early in training.

In response to these situations, we further adjusted the loss function as follows:

$$L_{\text{Ecl}} = \underbrace{\frac{1}{N}\sum_{i=1}^{I}\frac{1}{\left|\widetilde{P}_i\right|}(\log(\varepsilon+\sum_{p\in\widetilde{P}}\frac{\exp(-\text{sim}(\mathbf{z}_i,\mathbf{z}_p^+)\cdot v)}{\tau}))}_{\text{sample-pull}} + \underbrace{\frac{\beta_i}{N}\sum_{i=1}^{I}(\log(\varepsilon+\sum_{a\in A(i)}\frac{\exp(\text{sim}(\mathbf{z}_i,\mathbf{z}_a)}{\tau}))}_{\text{sample-push}} + L2 \quad (11)$$

$$\text{sim}(\mathbf{z}_i,\mathbf{z}_p^+)\cdot v = \frac{\mathbf{z}_i\mathbf{z}_p^+}{\|\mathbf{z}_i\|\left\|\mathbf{z}_p^+\right\|}\cdot\mathbb{N}((n_i\log(\delta+n_i))) \quad (12)$$

To address intra-class distance attraction, a similarity measure control weight v is constructed to dynamically adjust intra-class feature distance constraints based on the sample quantity ratio during training. The parameter n_i denotes the number of the corresponding sample within the entire dataset, while δ serves as a hyperparameter. To maintain numerical stability without significantly altering the scale of the loss function, the parameter is set to 10^{-8} for ε. The term $\mathbb{N}(\cdot)$ refers to the normalization operation. In this paper, balanced weights are assigned to positive samples from both head and tail classes, with $1/|\widetilde{P}_i|$ being combined with these balanced weights to constrain positive samples from head classes and enhance positive samples from tail classes. This adjustment explicitly emphasizes the similarity between positive sample pairs of tail classes during loss computation. It ensures that the model prioritizes the contribution of positive sample pairs during optimization, preventing their dilution by a large number of negative samples, thereby assisting the model in forming tighter clusters within the intra-class feature space.

For inter-class distance constraints, a class balance coefficient β_i/N is proposed to flexibly adjust the weights of positive and negative sample pairs, ensuring that equal contributions are made by each class to the loss computation. The parameter β_i denotes $\beta_i = 1/\sqrt{n_i}$. This approach guarantees that less frequent classes receive higher weights during loss computation so that the recognition ability of the model regarding rare classes is improved. Finally, an L2 regularization term is incorporated to help mitigate overfitting, thereby improving the performance and stability of the model.

The first component of L_{Ecl}, termed "sample-pull", is used to manage the similarity of positive sample pairs with the aim of encouraging more compact representations in the feature space. The second component, termed "sample-push", is employed to regulate the similarity across all samples, ensuring effective discrimination among samples from different classes is achieved. In summary, inter-class distances are simultaneously constrained while intra-class similarity is enhanced by our model, with the goal of dispersing feature representations across samples from different classes while compacting those of the same class. This approach thereby promotes the adaptability of contrastive learning to imbalanced datasets.

We further design a multi-task loss function to optimize all the subtasks, so that more precise and stable object detection results can be obtained.

$$L_{\text{Total}} = L_{\text{Reg}} + L_{\text{Cls}} + L_{\text{Cbc}} + \lambda L_{\text{Ecl}} \quad (13)$$

where L_{Cls} and L_{Reg} denote the cross-entropy loss and smoothL1 loss, respectively, which are used to supervise the classification and regression branches. The hyperparameter λ is used to balance the weights between the proposed enhanced contrastive loss and the detection loss, and it is set to 0.25.

4. Results

4.1. Datasets

In this paper, we evaluate our method on the FAIR1M dataset [40] and the MAR20 dataset [41].

(1) The FAIR1M dataset consists of 24,625 images, with 16,488 images allocated for training and 8137 images for testing. The spatial resolution of the image ranges from 0.3 m to 0.8 m, with each image ranging in size from 1000 × 1000 to 10,000 × 10,000 pixels. The dataset includes 5 coarse-grained classes and 37 fine-grained classes, encompassing objects of various scales, orientations, and shapes. All categories and the number of instances per category in the FAIR1M and MAR20 datasets are shown in Figure 4.

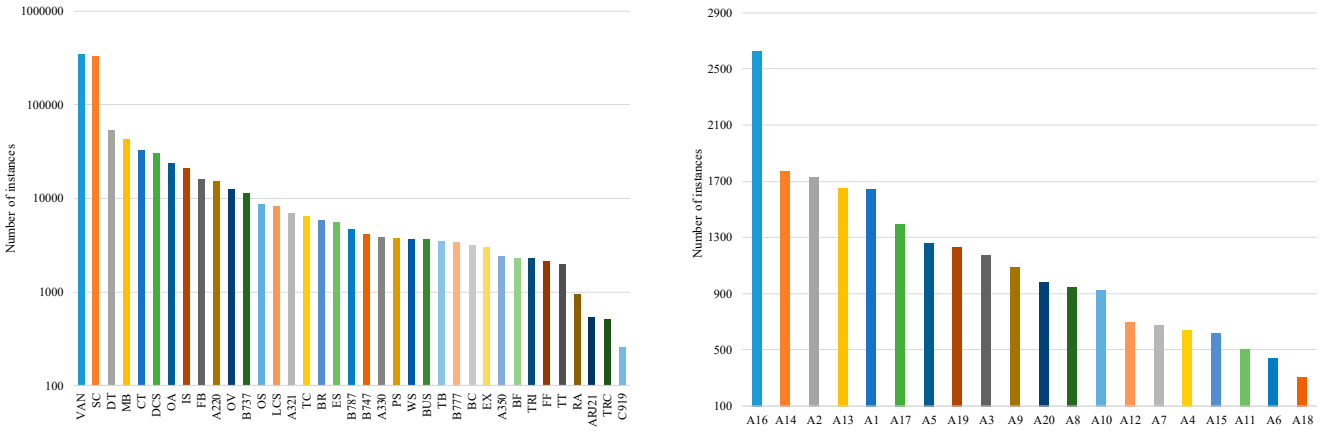


Figure 4. Sample distribution. (**a**) All categories and the number of instances per category in FAIR1M dataset; (**b**) All categories and the number of instances per category in MAR20 dataset.

(2) The MAR20 dataset is a high-resolution, fine-grained military aircraft object detection dataset that includes 3842 images and 22,341 instance objects. We identify 20 classes of fine-grained military aircraft objects, including SU-35, C-130, C-17, C-5, F-16, TU-160, E-3, B-52, P-3C, B-1B, E-8, TU-22, F-15, KC-135, F-22, FA-18, TU-95, KC-10, SU-34, and SU-24. These models are abbreviated as A1 to A20 in sequential order. We adopt the official dataset partitioning scheme of MAR20, comprising 1331 images and 7870 objects for training and 2511 images and 14,471 objects for testing.

4.2. Implementation Details

We employed Oriented R-CNN as the baseline, with ResNet50-FPN serving as the backbone network, optimizing model parameters using the Adam algorithm for stochastic gradient descent. The initial learning rate, momentum, and weight decay were set to 0.001, 0.9, and 0.0001, respectively. The model was trained for 12 epochs, with learning rate reductions scheduled at the 6th and 10th epochs. A batch size of 16 was used for the experiments, which were conducted exclusively on a single NVIDIA GeForce RTX 3060Ti GPU.

Our method is compared with the former methods, as shown in Tables 1 and 2. To assess the performance of object detection, we used average precision (AP) and mAP, where mAP represents the mean of the AP scores for all object classes.

Table 1. Quantitative experiment results on the FAIR1M dataset. (Red indicates optimal results, and blue represents suboptimal results).

Methods		FCOS-O [42]	RetinaNet-O [43]	S2A-Net [44]	Faster R-CNN-O [45]	Redet [46]	Oriented R-CNN [47]	RoI-Transformer [48]	Gliding Vertex [49]	SFRNet [13]	RB-FPN [14]	PCLDet [50]	Ours
Airplane	B737	32.43	35.24	36.06	37.11	36.51	35.17	40.14	36.50	40.63	39.16	43.51	46.21
	B747	79.97	74.90	84.33	84.23	87.03	85.17	84.92	81.88	84.32	86.70	87.37	88.51
	B777	13.76	10.54	15.62	15.64	22.45	14.57	15.39	14.06	17.64	16.17	28.73	31.55
	B787	46.85	38.49	42.34	46.95	50.68	47.68	49.17	44.60	48.88	49.16	55.37	58.32
	C919	1.18	0.96	1.95	9.88	10.82	11.68	19.73	10.16	21.25	13.60	20.04	26.25
	A220	45.87	41.39	44.09	48.15	47.92	46.55	50.46	46.43	48.59	49.18	51.65	53.15
	A321	65.38	63.51	68.00	66.46	71.70	68.18	70.31	65.47	71.18	66.91	73.06	72.18
	A330	61.39	46.28	63.84	68.86	72.45	68.60	71.42	67.73	72.97	71.43	74.52	73.52
	A350	67.25	63.42	70.00	69.33	77.57	70.21	72.62	68.31	74.05	74.34	79.86	79.44
	ARJ21	10.40	2.29	12.10	25.23	38.80	25.32	33.65	25.97	32.41	29.00	31.08	42.50
Ship	PS	10.02	5.78	8.82	9.08	18.38	13.77	13.21	10.01	17.39	18.38	18.21	20.92
	MB	51.99	23.00	48.03	49.44	61.71	60.42	56.54	51.63	60.55	68.41	62.88	68.02
	FB	8.19	2.52	6.79	4.55	11.77	9.10	6.82	5.26	8.46	10.55	12.05	11.61
	TB	31.67	24.87	34.01	31.85	35.87	36.83	35.71	34.00	34.92	38.27	35.17	41.78
	ES	11.31	6.98	7.49	9.31	13.35	11.32	9.96	10.34	12.60	11.83	13.52	15.93
	LCS	18.35	7.39	18.3	9.79	23.83	21.86	16.91	14.53	19.68	25.12	23.79	24.89
	DCS	38.26	21.75	37.62	25.93	41.56	38.22	36.01	33.10	38.22	38.41	42.26	46.76
	WS	20.90	2.75	22.96	10.42	34.36	22.67	17.30	12.45	21.44	34.72	37.45	40.02
Vehicle	SC	48.09	37.22	61.57	54.38	61.82	57.62	58.29	54.39	58.75	70.75	61.26	66.79
	BUS	11.54	4.25	11.76	18.76	19.96	24.40	28.02	28.63	32.77	36.14	29.13	37.02
	CT	29.66	18.38	34.28	35.85	41.74	40.84	40.55	36.90	41.05	44.94	42.81	43.18
	DT	20.70	12.59	36.03	41.29	47.33	45.20	45.97	42.16	46.63	50.20	47.99	47.27
	VAN	40.77	26.44	54.62	48.91	56.30	54.01	54.10	48.52	54.12	70.77	56.93	69.55
	TRI	7.50	0.01	3.47	8.25	13.60	15.46	11.82	13.38	15.70	16.75	14.57	19.56
	TRC	3.73	0.02	0.96	1.88	1.98	2.37	2.61	1.39	7.12	1.68	5.19	7.79
	EX	7.60	0.13	7.24	7.17	12.20	13.55	11.74	12.19	16.02	17.24	13.08	16.97
	TT	0.30	0.03	0.04	0.40	0.79	0.24	0.72	0.19	0.39	0.49	1.36	1.48
Court	BC	41.27	27.51	38.44	48.93	54.67	48.18	47.50	45.83	49.78	54.59	55.01	58.18
	TC	79.34	79.20	80.44	78.58	79.35	78.45	80.12	77.99	79.25	80.49	78.69	79.37
	FF	60.04	55.44	56.34	53.55	70.09	60.79	58.03	59.27	59.87	65.69	69.55	69.63
	BF	87.03	87.89	87.47	87.80	90.57	88.43	87.37	86.08	88.73	88.92	90.69	92.94
Road	IS	58.95	54.38	50.76	59.15	61.70	57.90	58.58	58.19	59.52	56.64	62.35	66.44
	RA	23.46	23.91	16.67	22.44	20.47	17.57	21.85	19.45	22.08	20.37	20.55	23.56
	BR	24.22	4.38	17.24	13.48	38.82	28.63	23.52	22.82	32.79	32.19	39.30	37.64
	mAP	**34.10**	**26.58**	**34.71**	**35.38**	**42.00**	**38.85**	**39.15**	**36.47**	**40.87**	**42.62**	**43.50**	**46.44**

Table 2. Quantitative experiment results on the MAR20 dataset. (Red indicates optimal results, and blue represents suboptimal results).

Category	One-Stage				Two-Stage					
	FCOS-O [42]	RetinaNet-O [43]	S2A-Net [44]	Faster R-CNN-O [45]	Double-Head-O [51]	Oriented R-CNN [47]	Gliding Vertex [49]	RoI-Transformer [48]	SFRNet [13]	Ours
A1	68.50	79.04	82.62	85.01	86.35	86.05	85.85	85.40	85.22	88.27
A2	79.70	84.31	81.59	81.63	80.85	81.73	81.53	81.53	82.04	84.29
A3	61.00	71.05	86.21	87.47	88.90	88.08	86.80	87.61	88.71	87.96
A4	52.34	54.72	80.75	70.68	82.54	69.57	76.35	78.33	79.73	86.19
A5	64.00	73.18	76.86	79.63	76.04	75.61	72.22	80.45	79.84	82.58
A6	83.30	86.59	90.00	90.58	90.06	89.92	89.90	90.49	90.68	90.14
A7	72.90	75.57	84.73	89.71	89.76	90.49	89.84	90.24	90.21	89.87
A8	82.28	85.51	85.70	89.82	87.28	89.54	89.38	87.58	89.62	91.58
A9	81.11	88.65	88.70	90.40	89.20	89.78	89.14	87.93	89.93	91.82
A10	84.55	85.84	90.84	90.89	90.78	90.91	90.77	90.89	90.77	90.75
A11	67.70	68.20	81.67	85.54	84.35	87.62	86.20	85.88	86.55	90.56
A12	78.77	73.22	86.09	88.08	86.18	88.39	87.45	89.29	88.39	88.72
A13	60.86	63.51	69.59	68.39	65.76	67.52	64.94	67.24	67.12	70.34
A14	81.26	79.72	85.25	88.27	87.42	88.50	88.28	88.20	88.08	87.14
A15	32.82	24.10	47.69	42.44	44.13	46.33	47.01	47.85	53.45	53.73
A16	81.84	84.85	88.10	88.86	87.49	88.27	87.84	89.11	88.56	91.74
A17	90.60	90.32	90.20	90.45	90.25	90.59	90.40	90.46	90.34	92.47
A18	51.06	49.19	61.98	62.23	56.40	70.50	64.94	74.59	76.14	77.47
A19	68.12	74.96	83.59	78.25	82.82	78.72	83.90	81.30	84.29	86.53
A20	71.10	76.07	79.84	77.71	76.69	80.25	76.83	80.00	72.67	78.94
mAP	**70.69**	**73.43**	**81.10**	**81.35**	**81.16**	**81.92**	**81.48**	**82.72**	**84.41**	**85.05**

4.3. Comparisons with Other Methods

(1) Quantitative results: The proposed method is compared with other advanced fine-grained object detection methods to analyze performance differences. These methods include FCOS-O [42], RetinaNet-O [43], S2A-Net [44], Faster R-CNN-O [45], Double-Head-O [51], Oriented R-CNN [47], RoI-Transformer [48], Gliding Vertex [49], RB-FPN [14], PCLDet [50], and Redet [46].

We conducted quantitative evaluations of our detection results, as shown in Tables 1 and 2. It can be seen that our method exhibits superior detection performance in terms of mAP compared to other methods in the FAIR1M and MAR20 datasets. Specifically, our method achieves 46.48 mAP (7.59 mAP ↑) and 85.05 mAP (3.13 mAP ↑) compared to the baseline methods on the FAIR1M and MAR20 datasets, respectively, while surpassing the second-best performing by 2.94 mAP and 2.33 mAP.

The superior performance of our model may stem from its thorough consideration of the requirements of RSI. By considering the unique characteristics of objects in RSI, relevance pooling was designed to retain spatial features pertinent to the objects, thereby enabling the attention module to provide more discriminative features for the model. Additionally, acknowledging the severe class imbalance prevalent in RSI, we addressed this issue from a feature perspective. By employing feature translation and tail feature reinforcement, we explicitly highlighted the boundaries of the tail class feature space, enhancing the distinctiveness of tail class features and alleviating the long-tail problem. This approach avoids potential overfitting and insufficient information issues that might arise from model resampling and reweighting. Lastly, we proposed supervised ECL, incorporating similarity measure control weights and class balance coefficients to impose spatial constraints on intra-class and inter-class distances, thereby adapting more effectively to imbalanced datasets.

Table 2 presents the detection results of our proposed method and other approaches across all fine-grained classes in MAR20. From the experimental data in Table 2, it is evident that models achieve higher average precision (AP) values for models such as A8, A9, A10, and A17, which possess more discriminative object features despite having fewer samples. Conversely, all methods exhibit lower recognition performance for the A15 category, primarily due to high inter-class similarity between A15 and other models like A5 coupled with significantly fewer samples in the A15 category compared to other object models. Nonetheless, our method demonstrates superior AP performance in this category compared to other comparative methods, attributable to effective strategies designed to address these challenges.

We note that our method is not superior to other methods in every class. This could be attributed to the increased number of parameters, leading to potential slight overfitting of the classifier to specific classes. But in general, our method demonstrates superior performance in most cases, particularly on tail-end samples.

(2) Visual results: We initially present the visual detection results of our proposed method on the FAIR1M and MAR20 datasets in Figures 5 and 6, respectively. The results demonstrate that our method performs better in fine-grained classification and exhibits robust detection performance across objects of varying scales and fine-grained classes. As depicted in Figures 5 and 6, our method effectively detects various object subclasses and performs well even on tail classes, such as A350, TB in FAIR1M, and A15, A11, and A18 in MAR20. Additionally, Figure 7 provides a comparison of fine-grained recognition between our method and other comparative methods on the FAIR1M dataset, including Oriented R-CNN and PCLDet. The results in Figure 7 illustrate that our method accurately locates objects and achieves precise fine-grained classification, as evidenced in the first column (A330 and OA) and the fifth column (A350). Moreover, our method significantly reduces false positives and missed detections, as shown in the third column (Van and SC) and the fourth column (MB).

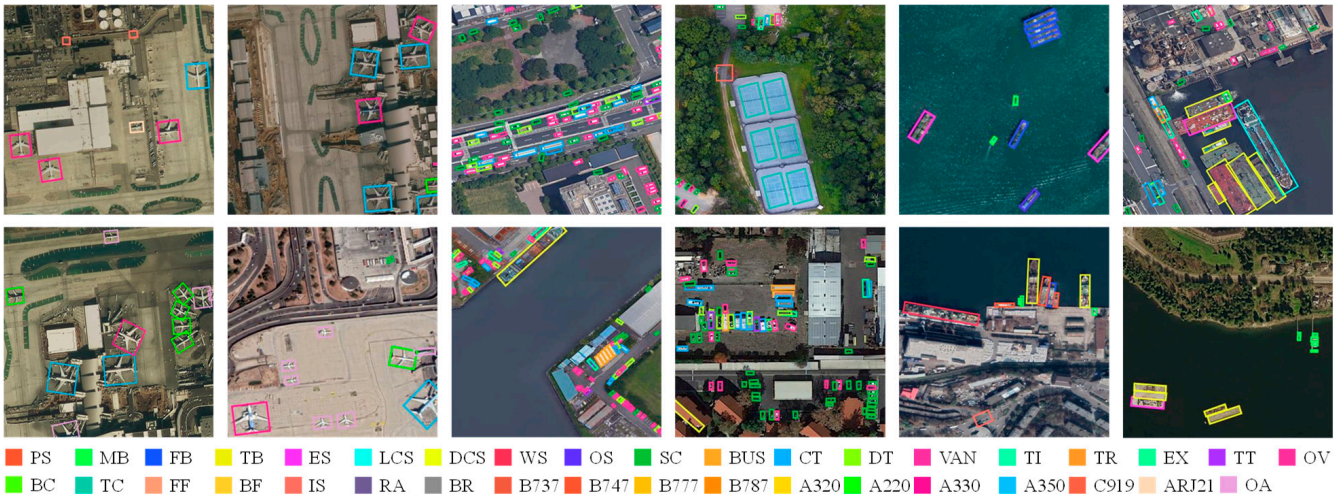


Figure 5. Results of our method for fine-grained oriented object detection on FAIR1M dataset.



Figure 6. Results of our method for fine-grained oriented object detection on MAR20 dataset.

In our method, we retain object-relevant features during down-sampling in the attention mechanism, providing high-quality fine-grained feature mappings to the model. Furthermore, we effectively mitigate the tail class problem through feature translation and reinforcement. Finally, supervised ECL was proposed to impose spatial constraints on intra-class and inter-class distances by using similarity measure control weights and class balance coefficients, thereby enabling the model to effectively recognize various subclasses within imbalanced datasets.



Figure 7. Visual comparisons on FAIR1M. The first line: Ground truth. The second line: Proposed method. The third line: Oriented R-CNN. The fourth line: PCLDet.

4.4. Ablation Study

(1) To validate the effectiveness of the submodule, we conducted ablation experiments on the FAIR1M and MAR20 datasets. We perform ablation experiments using different combinations of baseline: baseline alone, baseline + RGAM module, baseline + CBC, baseline + ECL, and our proposed method. Experimental results are presented in Table 3.

Table 3. Ablation experiment results on the FAIR1M and MAR20 datasets.

Baseline	Modules			mAP (%)		FLOPs (G)	Param (M)	Training Time (ms)
	RGAM	CBC	ECL	FAIR1M	MAR20			
√				38.85	81.92	134.52	41.14	558
√	√			39.87	82.34	134.57	41.14	587
√		√		44.08	84.37	134.74	41.46	654
√			√	41.83	83.02	134.66	41.34	628
√	√	√	√	**46.44**	**85.05**	**134.93**	**41.63**	**672**

Baseline configuration: We adopt Oriented R-CNN with direction prediction as our baseline, which is a two-stage object detection method. To ensure fairness in the comparison, all hyperparameters were uniformly set across all experiments.

It can be seen that, when all three modules are utilized simultaneously, the detection performance is improved by 7.59 mAP and 3.13 mAP compared to the baseline meth-

ods on the FAIR1M and MAR20 datasets, respectively. When only the CBC module is used, the Baseline + CBC module achieves an improvement of 5.23 mAP on FAIR1M and 2.45 mAP on MAR20. With only the RGAM module added, the performance of the model is improved by 1.02 mAP on FAIR1M and 0.42 mAP on MAR20. Compared to the baseline, the Baseline + ECL configuration achieves quantitative improvements, with 41.83 mAP (2.98 mAP ↑) on FAIR1M and 83.02 mAP (1.10 mAP ↑) on MAR20, surpassing the performance of the Baseline + RGAM module.

The results indicate that the CBC module exerts the most significant effect on model performance, followed by the ECL module, with the RGAM module having the least effect. The greatest improvement is observed when all three modules are used together.

The computational complexity and training costs of our method are primarily determined by the additional operations introduced by the CBC, RGAM, and ECL modules. To quantify these aspects, we compared the FLOPs (floating-point operations) and training time of the baseline with our proposed method.

It can be observed from Table 3 that both the submodules and the overall model achieve improvements with only negligible increases in FLOPs and model parameters. The introduction of new modules adds extra layers and operations, resulting in an approximately 114 ms increase in training time per batch compared to the baseline. However, this increase remains manageable within standard deep learning settings.

The experimental results demonstrate that the CBC module significantly enhances the performance of tail classes, while the RGAM effectively retains crucial spatial information within the attention units, thereby aiding the model in accurately identifying classes. Additionally, the ECL module imposes effective constraints on intra-class and inter-class feature distances, thereby optimizing contrastive learning for imbalanced datasets.

(2) To further validate the performance of the proposed submodules, we compared them with functionally similar modules as shown in Table 4. From the table, it is evident that, compared to the baseline modules with similar functionalities, the proposed submodules result in a significant enhancement of baseline performance. These results underscore the efficacy of the proposed submodules in facilitating FGOD in RSI.

Table 4. Functional comparison experiment results on the FAIR1M and MAR20 datasets.

Baseline	Submodules	Similar Modules	mAP (%)			
			FAIR1M	Gain	MAR20	Gain
√			38.85	0	81.92	0
√	RGAM		39.87	1.02 ↑	82.34	0.42 ↑
√		Global average pooling	39.01	0.16 ↑	82.11	0.19 ↑
√	CBC		44.08	5.23 ↑	84.37	2.45 ↑
√		Reweighting	41.31	2.46 ↑	82.72	0.80 ↑
√	ECL		41.83	2.98 ↑	83.02	1.20 ↑
√	-	Contrast learning	39.84	0.99 ↑	82.33	0.41 ↑

(3) Hyper-Parameter λ in multi-task loss function: In this section, we examine the impact of varying λ on model performance. As shown in Table 5, the object detection performance of the model exhibits different trends as λ increases, peaking at $\lambda = 0.25$. The differing degrees of class imbalance between the FAIR1M and MAR20 datasets may account for the varying effects of λ on mAP across the two datasets. Ultimately, we empirically select $\lambda = 0.25$ as the optimal value.

Table 5. Ablation experiment results on hyper-parameter λ in the multi-task loss function.

Hyper-Parameter λ		λ = 0.1	λ = 0.25	λ = 0.5	λ = 1
mAP (%)	**FAIR1M**	45.82	**46.44**	45.97	44.89
	MAR20	84.67	**85.05**	84.55	83.25

Overall, the aforementioned observations confirm the effectiveness of the submodules. On one hand, the RGAM module provides the model with more discriminative features by retaining more object-relevant features. On the other hand, CBC simultaneously highlights the boundaries of the tail class feature space and enhances the distinctiveness of tail class features through feature translation and tail feature reinforcement, thereby alleviating the long-tail problem. Lastly, ECL imposes spatial constraints on intra-class and inter-class distances by providing similarity measure control weights and class balance coefficients, making it more suitable for imbalanced datasets.

5. Discussion

The experimental results underscore the effectiveness of our proposed FGO^{2}D method in tackling the intrinsic challenges of fine-grained object detection in remote sensing images, particularly concerning class imbalance and high inter-class similarity. The RGAM module demonstrated its proficiency in preserving crucial spatial information by dynamically concentrating on object-relevant features during the down-sampling process. This capability markedly enhanced the ability of the model to distinguish subtle differences among fine-grained classes, as evidenced by the superior mAP scores attained on the FAIR1M and MAR20 datasets.

Additionally, the CBC module successfully mitigated the long-tail problem through the introduction of a learnable reinforcement coefficient and feature translation, ensuring that tail class features retained their distinctiveness and were not overshadowed by the more prevalent head class features. This sophisticated approach to managing class imbalance was a critical factor in the superior detection performance observed.

The ECL strategy further bolstered the robustness of our method by optimizing feature distributions across classes. Through the dynamic adjustment of the influence of inter-class and intra-class samples, this strategy facilitated tighter intra-class clustering while maintaining sufficient inter-class separation. This balance is critical in fine-grained detection tasks where classes often exhibit high visual similarity.

Despite the promising results, there remain areas for further investigation. The computational complexity introduced by additional modules, such as RGAM and ECL, may affect the scalability of the method for larger datasets or real-time applications. Future research could explore optimization techniques to reduce this computational overhead without compromising detection accuracy. Moreover, expanding the evaluation to a more diverse array of remote sensing datasets with varying characteristics could yield deeper insights into the generalizability of our approach.

In summary, our proposed method significantly advances the state-of-the-art in fine-grained object detection within remote sensing images by effectively addressing the dual challenges of class imbalance and high inter-class similarity. Nonetheless, ongoing research is essential to further refine the method and broaden its applicability across different domains and dataset conditions.

6. Conclusions

This paper presents an FGO^{2}D method based on relevance pooling guidance and class balance feature enhancement to enhance the detection performance of the network for fine-grained objects in RSI. To improve the discriminative feature extraction ability of the model, we propose a novel RGAM to retain and learn useful features relevant to the object. Furthermore, a CBC module is proposed to highlight tail class feature boundaries and maintain feature distinctiveness, thereby mitigating the long-tail problem. To constrain

inter-class distances and optimize intra-class feature distributions in the context of learning from imbalanced datasets, we devise an ECL approach. Experimental results on benchmark datasets demonstrate the superior performance of our method in FGOD.

Author Contributions: Conceptualization, Y.W.; methodology, Y.W.; software, Y.W.; validation, G.L.; formal analysis, H.C.; investigation, Y.W.; resources, H.C.; data curation, Y.W.; writing—original draft preparation, Y.W.; writing—review and editing, Y.W.; visualization, Y.W. and G.L.; supervision, Y.Z. and H.C.; project administration, Y.Z. and H.C.; funding acquisition, H.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: These data were derived from the following resources available in the public domain: [FAIR1M Website: https://gaofen-challenge.com/ and MAR20: https://gcheng-nwpu.github.io/].

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

References

1. Fu, J.; Zheng, H.; Mei, T. Look Closer to See Better: Recurrent Attention Convolutional Neural Network for Fine-Grained Image Recognition. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4476–4484.
2. Yang, Z.; Luo, T.; Wang, D.; Hu, Z.; Gao, J.; Wang, L. Learning to navigate for fine-grained classification. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
3. Lee, S.; Moon, W.; Heo, J.-P. Task Discrepancy Maximization for Fine-grained Few-Shot Classification. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 5321–5330.
4. Ding, Y.; Zhou, Y.; Zhu, Y.; Ye, Q.; Jiao, J. Selective Sparse Sampling for Fine-Grained Image Recognition. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6598–6607.
5. Han, Y.; Yang, X.; Pu, T.; Peng, Z. Fine-grained recognition for oriented ship against complex scenes in optical remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5612318. [CrossRef]
6. Chen, J.; Chen, K.; Chen, H.; Li, W.; Zou, Z.; Shi, Z. Contrastive learning for fine-grained ship classification in remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4707916. [CrossRef]
7. Ouyang, L.; Fang, L.; Ji, X. Multigranularity Self-Attention Network for Fine-Grained Ship Detection in Remote Sensing Images. *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.* **2022**, *15*, 9722–9732. [CrossRef]
8. Yu, S.; Guo, J.; Zhang, R.; Fan, Y.; Wang, Z.; Cheng, X. A Re-Balancing Strategy for Class-Imbalanced Classification Based on Instance Difficulty. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 70–79.
9. Qin, Y.; Zheng, H.; Yao, J.; Zhou, M.; Zhang, Y. Class-Balancing Diffusion Models. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 18434–18443.
10. Alshammari, S.; Wang, Y.-X.; Ramanan, D.; Kong, S. Long-Tailed Recognition via Weight Balancing. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 6887–6897.
11. Luo, W.; Yang, X.; Mo, X.; Lu, Y.; Davis, L.; Li, J.; Yang, J.; Lim, S. Cross-X Learning for Fine-Grained Visual Categorization. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8241–8250.
12. Han, J.; Yao, X.; Cheng, G.; Feng, X.; Xu, D. P-CNN: Part-Based Convolutional Neural Networks for Fine-Grained Visual Categorization. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 579–590. [CrossRef]
13. Cheng, G.; Li, Q.; Wang, G.; Xie, X.; Min, L.; Han, J. SFRNet: Fine-Grained Oriented Object Recognition via Separate Feature Refinement. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5610510. [CrossRef]
14. Song, J.; Miao, L.; Ming, Q.; Zhou, Z.; Dong, Y. Fine-Grained Object Detection in Remote Sensing Images via Adaptive Label Assignment and Refined-Balanced Feature Pyramid Network. *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.* **2023**, *16*, 71–82. [CrossRef]
15. Zhou, Y.; Wang, S.; Zhao, J.; Zhu, H.; Yao, R. Fine-Grained Feature Enhancement for Object Detection in Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6508305. [CrossRef]
16. Zeng, L.; Guo, H.; Yang, W.; Yu, H.; Yu, L.; Zhang, P.; Zou, T. Instance Switching-Based Contrastive Learning for Fine-Grained Airplane Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5633416. [CrossRef]
17. Zhuang, P.; Wang, Y.; Qiao, Y. Learning attentive pairwise interaction for fine-grained classification. In Proceedings of the AAAI Conference on Artificial Intelligence, Austin, TX, USA, 30 July–3 August 2020; Volume 34, pp. 13130–13137.

18. Zheng, H.; Fu, J.; Zha, Z.-J.; Luo, J. Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 5007–5016.
19. Shi, W.; Gong, Y.; Tao, X.; Cheng, D.; Zheng, N. Fine-Grained Image Classification Using Modified DCNNs Trained by Cascaded Softmax and Generalized Large-Margin Losses. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 683–694. [CrossRef]
20. Bao, W.; Hu, J.; Huang, M.; Xu, Y.; Ji, N.; Xiang, X. Detecting Fine-Grained Airplanes in SAR Images With Sparse Attention-Guided Pyramid and Class-Balanced Data Augmentation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 8586–8599. [CrossRef]
21. Zhu, H.; Ke, W.; Li, D.; Liu, J.; Tian, L.; Shan, Y. Dual Cross-Attention Learning for Fine-Grained Visual Categorization and Object Re-Identification. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 4682–4692.
22. Wang, Y.; Peng, J.; Wang, H.; Wang, M. Progressive learning with multi-scale attention network for cross-domain vehicle re-identification. *Sci. China Inf. Sci.* **2022**, *65*, 160103. [CrossRef]
23. Li, X.; Wu, J.; Sun, Z.; Ma, Z.; Cao, J.; Xue, J.-H. BSNet: Bi-Similarity Network for Few-shot Fine-grained Image Classification. *IEEE Trans. Image Process.* **2021**, *30*, 1318–1331. [CrossRef] [PubMed]
24. Sun, B.; Li, B.; Cai, S.; Yuan, Y.; Zhang, C. FSCE: Few-shot object detection via contrastive proposal encoding. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 7348–7358.
25. Byrd, J.; Lipton, Z. What is the effect of importance weighting in deep learning? In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 10–15 June 2019; pp. 872–881.
26. Li, J.; Lin, X.; Zhang, W.; Tan, X.; Li, Y.; Han, J.; Ding, E.; Wang, J.; Li, G. Gradient-based Sampling for Class Imbalanced Semi-supervised Object Detection. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 16344–16354.
27. Zhang, Z.; Pfister, T. Learning Fast Sample Re-weighting Without Reward Data. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 705–714.
28. Tan, J.; Wang, C.; Li, B.; Li, Q.; Ouyang, W.; Yin, C.; Yan, J. Equalization loss for long-tailed object recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11662–11671.
29. Zou, Y.; Yu, Z.; Kumar, B.V.; Wang, J. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 289–305.
30. Buda, M.; Maki, A.; Mazurowski, M.A. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Netw.* **2018**, *106*, 249–259. [CrossRef]
31. 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. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 181–196.
32. Khan, S.H.; Hayat, M.; Bennamoun, M.; Sohel, F.A.; Togneri, R. Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE Trans. Neural Netw. Learn. Syst.* **2018**, *29*, 3573–3587. [CrossRef]
33. Datta, A.; Ghosh, S.; Ghosh, A. Combination of clustering and ranking techniques for unsupervised band selection of hyperspectral images. *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.* **2015**, *8*, 2814–2823. [CrossRef]
34. Galar, M.; Fernandez, A.; Barrenechea, E.; Bustince, H.; Herrera, F. A review on ensembles for the class imbalance problem: Bagging- boosting- and hybrid-based approaches. *IEEE Trans. Syst. Man Cybern. C Appl. Rev.* **2012**, *42*, 463–484. [CrossRef]
35. Zhou, B.; Cui, Q.; Wei, X.-S.; Chen, Z.-M. BBN: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 9719–9728.
36. Yin, X.; Yu, X.; Sohn, K.; Liu, X.; Chandraker, M. Feature transfer learning for face recognition with under-represented data. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
37. Li, T.; Wang, L.; Wu, G. Self supervision to distillation for long-tailed visual recognition. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021.
38. Cui, J.; Zhong, Z.; Liu, S.; Yu, B.; Jia, J. Parametric contrastive learning. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021.
39. Wang, H.; Yao, M.; Chen, Y.; Xu, Y.; Liu, H.; Jia, W.; Fu, X.; Wang, Y. Manifold-based Incomplete Multi-view Clustering via Bi-Consistency Guidance. *IEEE Trans. Multimed.* **2024**. early access. [CrossRef]
40. Sun, X.; Wang, P.; Yan, Z.; Xu, F.; Wang, R.; Diao, W.; Chen, J.; Li, J.; Feng, Y.; Xu, T.; et al. FAIR1M: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *184*, 116–130. [CrossRef]
41. Yu, W.Q.; Cheng, G.; Wang, M.J.; Yao, Y.Q.; Xie, X.X.; Yao, X.X.; Han, J.W. MAR20: A benchmark for military aircraft recognition in remote sensing images. *National Remote Sensing Bulletin* **2023**, *27*, 2688–2696. [CrossRef]
42. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: Fully Convolutional One-Stage Object Detection. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27–28 October 2019; pp. 9626–9635.

43. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2999–3007.
44. Han, J.; Ding, J.; Li, J.; Xia, G.-S. Align Deep Features for Oriented Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–11. [CrossRef]
45. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
46. Han, J.; Ding, J.; Xue, N.; Xia, G.-S. ReDet: A rotation-equivariant detector for aerial object detection. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 2785–2794.
47. Xie, X.; Cheng, G.; Wang, J.; Yao, X.; Han, J. Oriented R-CNN for object detection. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 3520–3529.
48. 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 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 2849–2858.
49. Xu, Y.; Fu, M.; Wang, Q.; Wang, Y.; Chen, K.; Xia, G.-S.; Bai, X. Gliding Vertex on the Horizontal Bounding Box for Multi-Oriented Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 1452–1459. [CrossRef] [PubMed]
50. Ouyang, L.; Guo, G.; Fang, L.; Ghamisi, P.; Yue, J. PCLDet: Prototypical Contrastive Learning for Fine-Grained Object Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5613911. [CrossRef]
51. Wu, Y.; Chen, Y.; Yuan, L.; Liu, Z.; Wang, L.; Li, H.; Fu, Y. Rethinking Classification and Localization for Object Detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10183–10192.

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.