

Article

Weakly Supervised Fine-Grained Aircraft Detection in Remote Sensing Based on Prior-Knowledge Prototype Learning

Peng Chen ^{1,†}, Yang Qu ^{2,†} , Ruye Li ¹ and Zicong Zhu ^{1,*}

¹ School of Electronic and Information Engineering, Soochow University, Suzhou 215006, China; pchen1416146@stu.suda.edu.cn (P.C.); 20255228008@stu.suda.edu.cn (R.L.)

² State Key Laboratory of Cryospheric Science and Frozen Soil Engineering, Northwest Institute of Eco-Environment and Resources, Chinese Academy of Sciences, Lanzhou 730000, China; quyang@nieer.ac.cn

* Correspondence: zczhu2023@suda.edu.cn

† These authors contributed equally to this work.

Highlights

What are the main findings?

- The two-stage weakly supervised framework enables accurate fine-grained classification and localization with limited labels, using the Prior-knowledge-based Soft-label Generator (PSG) and Text Prototype Alignment (TPA) to replace large-scale annotations.
- The Feature Purification Module (FPM) reduces feature noise from complex backgrounds and target structures by embedding differentiable morphological operations into deep networks.

What are the implications of the main findings?

- Reducing annotation dependence lowers costs and boosts the practicality of fine-grained detection in remote sensing.
- The integration of soft-label generation, text alignment, and feature purification offers an effective approach to adapting Vision-Language Models for remote sensing with scarce labels, thereby enhancing generalization and discrimination.

Abstract

The weakly supervised aircraft fine-grained detection task remains challenging due to the lack of fine-grained category labels. In addition, the high inter-class similarity among aircraft categories and complex background interference further hinders discriminative feature learning. To address these challenges, we propose a weakly supervised fine-grained aircraft detection framework based on prior-knowledge prototype learning. Specifically, a Prior-knowledge-based Soft-label Generator (PSG) is designed to exploit geometric and semantic priors, enabling the propagation of fine-grained category annotations from a few annotated samples to large-scale unlabeled data through reliable soft-label generation. To alleviate category ambiguity caused by highly similar aircraft categories, a Text Prototype Alignment (TPA) strategy is introduced to align vision-language semantic knowledge with the remote sensing domain and construct discriminative semantic prototypes. Furthermore, a Feature Purification Module (FPM) is developed to suppress background noise through morphology-guided feature purification and enhance aircraft-specific representations. Experiments on the FAIR1M-Aircraft dataset demonstrate that the proposed method achieves 30.5% mAP with only 50 fine-grained annotated images, outperforming the baseline by 10.1 mAP points.



Academic Editor: Akira Iwasaki

Received: 2 June 2026

Revised: 27 June 2026

Accepted: 6 July 2026

Published: 17 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](#)

[Attribution \(CC BY\)](#) license.

Keywords: weakly supervised learning; fine-grained classification; aircraft detection; visual-textual alignment; remote sensing

1. Introduction

In recent years, the rapid development of high-resolution remote sensing (RS) optical imagery and deep learning technologies has brought revolutionary progress to the field of earth observation [1–4]. Among them, aircraft detection is an important fundamental task. It plays a crucial role in applications such as aviation surveillance, intelligent airport management, military situational awareness, and disaster emergency response [5–7]. However, traditional coarse-grained detection only recognizes the general category “aircraft” and can no longer meet the increasing practical needs of the real world. Accurately locating aircraft and identifying their specific models (such as the Boeing 737 and the Airbus 320) is not only a prerequisite for fine situational monitoring but also a key challenge that urgently needs to be addressed in RS vision tasks [8,9]. This has led to a more challenging task: fine-grained aircraft detection and recognition [10–14].

Despite significant progress in existing fine-grained detection methods, their practical application in RS scenes is still severely constrained by several key bottlenecks. Firstly, supervised learning methods rely heavily on a large number of high-quality, fine-grained annotated samples [1,2]. However, different aircraft models typically exhibit highly similar visual appearances. For example, the Boeing 737 and Airbus A320 display extremely subtle differences in top view, which makes manual annotation not only highly dependent on professional knowledge but also extremely labor-intensive. Secondly, the single-mode feature learning has inherent semantic limitations. Existing detection models [15–17] struggle to effectively distinguish highly similar subclasses using only visual features, and incorporating text features with rich semantic information warrants exploration. Finally, the aircraft’s stable apparent structure further complicates the extraction of discriminative features [13,18]. Due to the aircraft’s cross-shaped or swept-wing structure, the proposal regions provided by the existing region proposal network (RPN) [19] are mottled. These features are mixed with a large amount of background noise, such as apron markings and boarding bridges. This further blurs the distinctions between the target’s fine-grained classes.

To alleviate the high cost of annotation, WSL offers a promising alternative for object detection. However, current weakly supervised methods in RS mainly focus on the spatial localization of targets. The existing mainstream methods can usually be divided into two paradigms: one uses low-cost horizontal bounding-box annotations to learn object-oriented detectors [20,21], and the other uses low-cost point supervision schemes [22–25]. Although these methods effectively reduce the annotation cost of spatial localization, WSL on fine-grained classification subtasks remains a challenge.

It is crucial to clarify the distinction between our setting and conventional few-shot learning. First, the labeling criteria for fine-grained categories in RS often exhibit low correlation with the actual physical differences between targets (e.g., visually similar aircraft models). This makes purely supervised mapping from a few samples highly ambiguous. Second, even if a model were to cluster fine-grained features without explicit labels, the lack of a deterministic mapping from its emergent clusters to the ground-truth categories would hinder objective performance evaluation. Therefore, unlike few-shot learning, which aims to learn a classifier from a handful of labeled examples per novel class, our proposed framework operates under a weakly supervised setting. It leverages a large number of samples with coarse-grained category annotations (e.g., “airplane”), complemented by

an extremely small set (e.g., 1%) of fine-grained annotated samples. The primary role of the few fine-grained samples is not to learn a complete classifier, but to establish a reliable semantic mapping between the model's predicted feature space and the real-world category labels. Meanwhile, the abundance of coarse-labeled data provides essential structural and distributional priors. This paradigm significantly reduces the annotation burden while enabling controllable and valuable fine-grained recognition.

Recently, several methods have attempted to introduce contrastive learning into RS fine-grained object detection by using learnable prototype parameters or category-averaged features to represent category centers [26,27]. However, these methods remain largely limited to optimizing a single mode and rely heavily on learning from large numbers of samples. This significantly degrades the model's recognition performance under extremely limited training samples. The rapid development of VLMs [28–31] provides a new perspective for overcoming this challenge. Their cross-modal alignment ability enables semantically guided visual learning from textual information. However, the lack of expert knowledge in RS leads to significant domain gaps in its application. This further leads to insufficient differentiation among fine-grained visual features of RS that are aligned with text embeddings. Therefore, effectively integrating multi-source prior knowledge with extremely limited fine-grained annotations remains a challenge.

Based on the above analysis, we summarize three major challenges in weakly supervised fine-grained aircraft detection in RS images: (1) lack of fine-grained classification labels; (2) high inter-class similarity among categories; and (3) complex background interference in RS scenes. Based on this, we have designed and proposed a novel two-stage weakly supervised fine-grained object detection framework. This framework aims to significantly reduce annotation costs while achieving precise fine-grained recognition of aircraft categories. Specifically, our framework consists of three key innovations.

For the first challenge, we design a Prior-knowledge-based Soft-label Generator (PSG). PSG leverages geometric information and cross-modal semantic similarity to construct Gaussian distribution-based category prototypes, deriving reliable fine-grained information for large-scale coarse-labeled samples from only a handful of annotated instances. For the second, we introduce a Text Prototype Alignment (TPA) strategy for fine-grained representation learning. It adapts the text embeddings produced by a vision-language model to the RS domain and aligns them with RoI visual features, enabling more discriminative feature learning across highly similar aircraft categories. For the last, we further propose a morphology-guided Feature Purification Module (FPM). By integrating morphological suppression [32] with local enhancement convolution, FPM suppresses background interference while preserving discriminative target structures, producing cleaner and more informative RoI representations.

The main contributions of this paper can be summarized as follows:

- We propose a novel two-stage weakly supervised fine-grained object detection framework. The proposed framework effectively eliminates the dependence on large-scale fine-grained annotations while achieving fine-grained aircraft classification and localization with significantly reduced annotation costs.
- We design a Prior-knowledge-based Soft-label Generator (PSG) based on multi-source prior knowledge. It constructs Gaussian-distributed class prototypes by mining geometric information and cross-modal semantic similarity. Reliable soft labels are derived from large-scale unlabeled samples, thereby providing supervision.
- We design a Text Prototype Alignment (TPA) strategy. It adapts VLM-derived text prototypes for the RS domain using a learnable transformation. Next, the contrastive learning method is applied to align RoI features with the adapted text prototypes, thereby improving the distinguishability of fine-grained features.

- We propose a background-suppressed Feature Purification Module (FPM). By innovatively introducing morphological operations into deep networks, the proposed module effectively suppresses feature noise arising from complex RS backgrounds and specialized target structures.

2. Related Work

2.1. Remote Sensing Aircraft Detection

RS aircraft detection plays an important role in airport surveillance, military reconnaissance, and airspace management [7,9]. In recent years, with the rapid development of deep learning techniques, extensive research has focused on aircraft structural modeling, the utilization of prior knowledge, and feature extraction. Early studies mainly focused on aircraft detection based on convolutional neural networks (CNNs). Wu et al. [33] combined candidate target regions with CNNs to achieve fast aircraft detection in satellite RS images. Subsequently, Zhang et al. [34] proposed a weakly supervised aircraft detection method based on coupled convolutional neural networks, which effectively identified aircraft targets in high-resolution RS imagery. Li et al. [35] further integrated reinforcement learning with CNNs to construct an aircraft detection framework for RS scenarios, thereby improving detection performance. As research progressed, increasing attention has been paid to aircraft structural information and key component features. Zuo et al. [36] introduced an aircraft model recognition method based on deep convolutional neural network segmentation. By detecting critical components, including the nose, tail, and two wingtips, along with optimized contour masks, the method enabled finer-grained modeling of aircraft structure. Subsequently, Xiong et al. [37] proposed NLFFNet, which integrated non-local feature enhancement, component-level feature fusion, and an adaptive loss function to mitigate hard-sample ambiguity. By modeling long-range spatial dependencies and exploiting discriminative aircraft parts, the method achieved high-precision recognition across 47 aircraft categories. To address the high similarity among fine-grained aircraft categories, Jia et al. [38] designed ACDN, which introduces additional discriminative modules to accurately distinguish the unique tail features of confusing aircraft types, such as the ARJ21. Building on these studies, several methods have further improved aircraft detection performance through improved feature representation and semantic modeling. Cai et al. [39] proposed an end-to-end network, OE-FCN, which captures intra-class variations in RS images through a feature mining mechanism, thereby enhancing aircraft detection capability. Wei et al. [40] introduced X-LineNet, which transforms the aircraft detection task into predicting and combining intersecting line segments, enabling the network to learn richer, more discriminative visual semantic information. Subsequently, Zeng et al. [41] proposed a top-down aircraft detection framework for large-scale RS scenes, in which U-Net [42] was first used to detect airport regions, followed by an FEF-R-CNN for aircraft detection, achieving promising performance in complex scenarios. In addition, Liu et al. [43] proposed S2CGNet, which enhanced the perception of aircraft structures and component semantics by introducing the SAM [44] and IAG modules. Qian et al. [45] further proposed AR-Net, which incorporated knowledge graphs and expert-derived priors into an end-to-end detection framework to improve aircraft detection and recognition performance.

Overall, existing methods have made significant progress in aircraft structure modeling, feature extraction, and the utilization of prior knowledge. However, most mainstream approaches still heavily rely on large amounts of precise annotations. Under weak semantic supervision or in few-shot optimization scenarios, insufficient supervision and high similarity among fine-grained categories severely limit detection performance.

2.2. Weakly Supervised Object Detection

Weakly supervised object detection aims to achieve object classification and localization solely through image-level labels, thereby significantly reducing data annotation costs. In the general field of computer vision, this paradigm is typically based on multi-instance learning frameworks: the early representative work, WSDDN, generated candidate region scores via parallel classification and localization branches. However, these methods often suffer from the “local advantage”: detectors tend to focus on the most discriminative parts of the object while ignoring its overall structure. Subsequent research, such as OICR [46], aimed to alleviate this issue by iteratively improving pseudo labels. Although these methods have made significant progress on general datasets, directly transferring them to RS scenarios presents unique challenges. The complex background interference, arbitrary object orientations, and significant scale changes from an overhead perspective exacerbate the risk of failure of traditional weakly supervised methods in RS scenes. In recent years, H2RBox [20] has been a representative work in this direction, using only horizontal bounding box annotations to train a rotating object detector. The core idea was to mine object-oriented information from horizontal boxes through self-supervised learning. H2RBox-v2 [21] further introduced rotation-equivariant constraints, achieving performance close to that of a fully supervised rotation detector using only horizontal box annotations. Another paradigm was point-supervised object detection, which incurred lower annotation costs. P2RBox [22] used point annotations to guide segmentation models in generating candidate masks, used MIL networks to filter high-quality masks, and then converted the masks into rotated boxes through SAE modules. Point2RBox [23] proposed the first end-to-end oriented object detection scheme with single-point supervision. Point2RBox-v2 [24] further explored the spatial layout information between instances by introducing Gaussian overlap loss and Voronoi watershed loss, and performed particularly well in dense scenes. Point2RBox-v3 [25] addressed the two main problems of low pseudo-label quality and low utilization in existing methods by combining the complementary advantages of SAM [44] in sparse scenes and watershed in dense scenes.

Overall, existing weakly supervised and point-supervised oriented object detection methods have achieved significant progress in reducing annotation costs and improving object localization accuracy. However, fine-grained object detection requires not only precise object localization but also accurate discrimination among fine-grained categories. Most existing methods focus on object localization, whereas WSL for fine-grained category discrimination remains underexplored.

2.3. Prototype Learning for Object Detection

Prototype learning, a key branch of contrastive learning, aims to learn discriminative representations by pulling same-class samples closer and pushing different-class samples apart [26,27,47,48]. While instance-level contrastive methods like MoCo [49] effectively addressed negative sample scarcity and encoder inconsistency, they treated all instances as negatives by default, ignoring potential semantic correlations and thus limiting high-level representation learning. To address this, PCL [47] first introduced prototype learning into contrastive learning, generating class prototypes via clustering under an Expectation-Maximization framework. This paradigm has since been widely adopted in Few-Shot Object Detection (FSOD). Early FSOD methods struggled to extract robust prototypes from limited samples of novel classes. Wu et al. [50] proposed “Universal Prototypes” to learn shared cross-category features, enhancing generalization to unseen classes. In RS FSOD, methods such as P-CNN [51] and DPL-Net [48] have introduced prototype-guided proposal generation and dynamic hard-sample partitioning to address challenges such as complex backgrounds and intra-class diversity. Prototypical contrastive learning has also

been extended to fine-grained RS detection. PCLDet [26] built a class prototype bank to maximize inter-class distance. However, its learnable prototype vectors struggled with the limited number of categories per batch, resulting in inaccurate estimates. APCL [27] addressed this via exponential moving average-based prototype updates, significantly improving fine-grained ship recognition.

Despite these advances, most existing methods construct prototypes solely from visual features, underutilizing cross-modal semantic priors. Moreover, the semantic alignment between fine-grained visual features and prototypes remains underexplored in RS scenarios.

3. Preliminaries

(1) Annotation Form:

In supervised-learning-based aircraft detection tasks, training data requires not only Oriented Bounding Boxes (OBBs) but also fine-grained categories. For each aircraft, the OBB is represented using a five-parameter formulation $B = (x, y, w, h, \theta)$, where (x, y) denotes the center coordinates, w and h represent the width and height respectively, and $\theta \in [-\pi/2, \pi/2)$ is the rotation angle. Category information adopts one-hot encoding $c \in \{0, 1\}^C$, where C is the total number of categories, with only the ground-truth category set to 1 and all others to 0. The above precise localization and fine-grained category labeling of aircraft provide detailed supervisory information for the model but also entail high labor costs.

In contrast, WSL means weaker semantic annotation information. Our method aims to achieve WSL of the model on sub-tasks of fine-grained category recognition. Specifically, for each unlabeled instance, we generate soft labels $Y = [y_1, y_2, \dots, y_C]$, $y_c \in [0, 1]$ that form probability distributions across all categories, satisfying $\sum_{c=1}^C y_c = 1$. These soft labels serve as supervisory signals for unlabeled samples, effectively propagating knowledge from a limited set of annotated samples to the entire dataset.

(2) Training Pipeline:

Oriented object detectors typically adopt a two-stage framework. In the first stage, the oriented RPN [15] slides predefined anchor boxes over the feature map. For each anchor box $A = (x_a, y_a, w_a, h_a, \theta_a)$, the network predicts regression offsets $\Delta = (\Delta x, \Delta y, \Delta w, \Delta h, \Delta \theta)$ to generate candidate proposals $P = (x, y, w, h, \theta)$, where:

$$\begin{cases} x = x_a + \Delta x \cdot w_a, \\ y = y_a + \Delta y \cdot h_a, \\ w = w_a \cdot \exp(\Delta w), \\ h = h_a \cdot \exp(\Delta h), \\ \theta = \theta_a + \Delta \theta \end{cases} \quad (1)$$

In the second stage, Rotated RoI Align [52] extracts rotation-invariant features from the candidate proposals. Given a proposal P , it is rotated to a horizontal orientation via an affine transformation. The transformation matrix is defined as:

$$T = \begin{bmatrix} \cos \theta & -\sin \theta & x \\ \sin \theta & \cos \theta & y \\ 0 & 0 & 1 \end{bmatrix} \quad (2)$$

For any point $p = (u, v)$ within the proposal, its corresponding position in the original feature map is computed as $[x', y', 1]^T = T \cdot [u - w/2, v - h/2, 1]^T$, followed by bilinear interpolation to obtain the feature F_{RoI} . Finally, the classification head outputs category

probability distribution $p_c \in \mathbb{R}^N$, and the regression head outputs refinement offsets $\Delta' \in \mathbb{R}^5$, yielding the final detection box:

$$B_{\text{final}} = (x + \Delta'_x w, y + \Delta'_y h, w \cdot \exp(\Delta'_w), h \cdot \exp(\Delta'_h), \theta + \Delta'_\theta) \quad (3)$$

(3) Loss Function:

Based on the previous description of weakly supervised fine-grained detection tasks, we define variable M as the number of samples provided for fine-grained category labeling, and N as the number of samples in the dataset containing coarse-grained labeling. Specifically, let $M \leq N/100$; $M \leq 4 \cdot C$ to meet the requirement of a few samples provided. For supervised-learning-based fine-grained detectors, M fine-grained annotated samples are used to train the model. And our method is also based on M fine-grained samples, which are used to construct category-specific prototypes. And further expand it to N coarse-grained samples by generating fine-grained soft labels for model training. The classification loss functions used by both are standard cross-entropy:

$$\mathcal{L}_{cls} = -\frac{1}{N_{\text{pos}}} \sum_{i=1}^{N_{\text{pos}}} \sum_{c=1}^C \hat{y}_{i,c}^+ \log(p_{i,c}) - \frac{1}{N_{\text{neg}}} \sum_{j=1}^{N_{\text{neg}}} \hat{y}_{j,bg}^- \log(p_{j,bg}), \quad (4)$$

where N_{pos} denotes the positive sampled RoI samples, and N_{neg} denotes negative. For positive RoI samples, $\hat{y}_{i,c}^+$ is generated by the proposed PSG to provide fine-grained category supervision, while negative RoI samples are assigned with the background one-hot label $\hat{y}_{i,c}^-$ following the original detection strategy.

For localization loss, the smooth L1 distance is usually selected:

$$\mathcal{L}_{\text{reg}} = \frac{1}{N_{\text{pos}}} \sum_i \begin{cases} 0.5(B_i - \hat{B}_i)^2, & \text{if } |B_i - \hat{B}_i| < 1 \\ |B_i - \hat{B}_i| - 0.5, & \text{otherwise} \end{cases} \quad (5)$$

The total loss for the detector is $\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_{\text{reg}} \mathcal{L}_{\text{reg}}$. By jointly optimizing the classification and regression losses, the detector learns to leverage knowledge from the limited annotated samples while effectively utilizing abundant unlabeled data.

4. Methodology

4.1. Overview

In this section, we detail the proposed weakly supervised fine-grained aircraft detection framework. As illustrated in Figure 1, the framework consists of two stages: prior knowledge preprocessing and weakly supervised training. This two-stage strategy enables the model to efficiently transfer discriminative knowledge from a very limited number of fine-grained annotated samples to massive unlabeled data.

(1) Stage I: Prior Knowledge Preprocessing.

In the first stage, we mainly completed the generation of soft labels and the migration of text prototypes. Specifically, for fairness, we first extract and construct reliable prior prototypes from M fine-grained sample annotations. Specifically, for each fine-grained category, we extract geometric and cross-modal semantic similarity features and build Gaussian-prior prototypes based on them. For any unlabeled sample, soft labels are generated by calculating its similarity to each previous prototype. In addition, to fully leverage the semantic knowledge of VLMs, we extracted text embeddings and transferred them to text prototypes via parameter-efficient fine-tuning, thereby better aligning them with prior knowledge in RS. These text prototypes provide highly discriminative semantic information for feature alignment in the next stage.

(2) Stage II: Weakly Supervised Training.

We introduced a contrastive learning head equipped with the TPA strategy in the detector. It maps RoI features to the semantic space of the text prototype, bringing similar features closer and separating dissimilar features. In addition, we embedded FPM into the RoI Head's output to further suppress background noise and enhance feature representation. The entire model is trained end-to-end under the joint supervision of classification, regression, and contrastive loss.

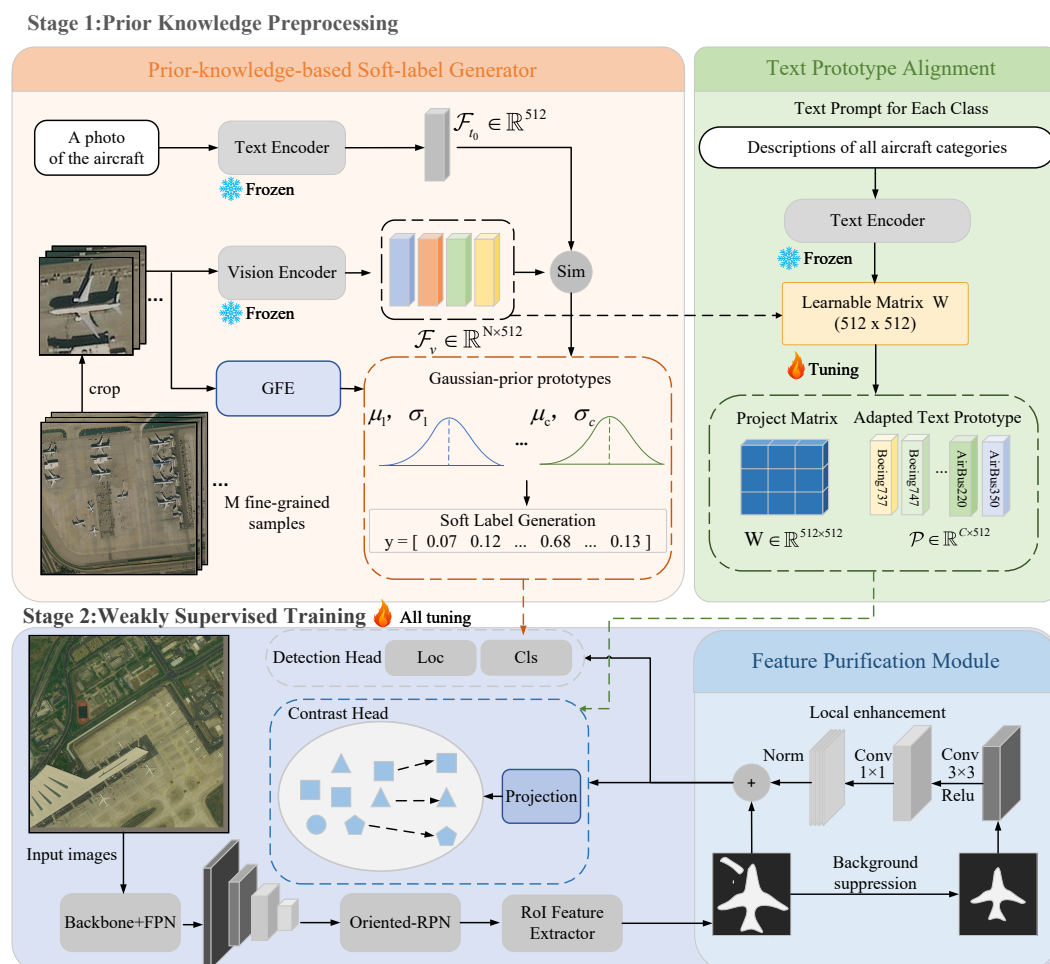


Figure 1. Overview of the proposed weakly supervised fine-grained aircraft detection framework. It consists of two stages: prior knowledge preprocessing and weakly supervised training.

4.2. Stage I: Prior Knowledge Preprocessing

4.2.1. Prior-Knowledge-Based Soft-Label Generator

Supervised-learning-based methods typically require a one-to-one correspondence between samples and annotations. In weakly supervised fine-grained detection tasks, the model can access only a few fine-grained annotations and cannot map them to a broader range of samples. To address the above challenges, we construct a set of prior knowledge prototypes by mining the provided fine-grained annotated geometric information and the aircraft's semantic knowledge. And based on this, we generate soft labels to map more coarse-grained annotated samples, guiding the model to achieve fine-grained category learning under WSL.

(1) Geometric Feature Extraction:

As illustrated in Figure 2, we design the GFE for estimating aircraft horizontal size and vertical height.

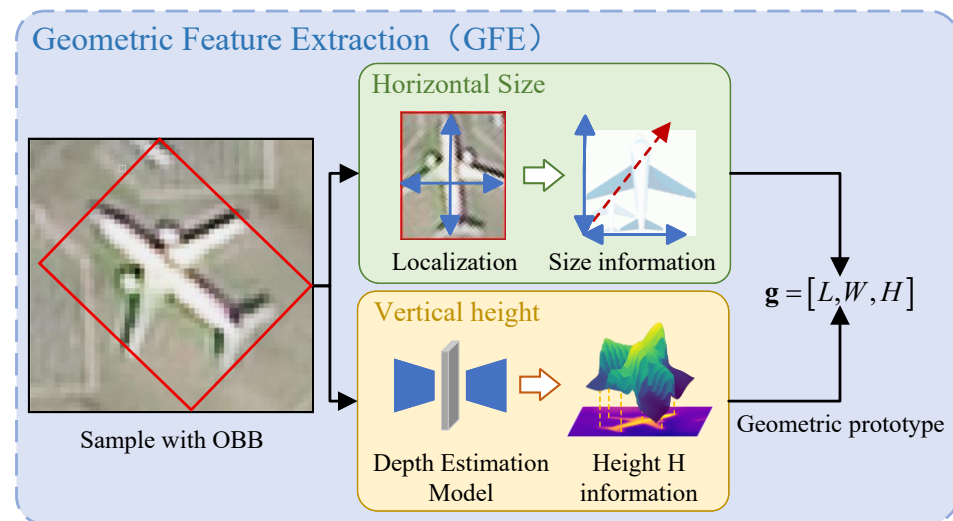


Figure 2. Geometric feature extraction, where the Depth Estimation Model is a monocular depth estimation model for obtaining height information.

Given a coarse-grained sample, its oriented bounding box is defined as $B = (x, y, w, h, \theta)$, where the length and width are defined as $L = \max(w, h)$ and $W = \min(w, h)$, respectively. They precisely represent the aircraft's length and width, which are specific to each model. Considering that aircraft altitude information is equally important for distinguishing aircraft models, but altitude information cannot be directly obtained from RS images.

We have introduced existing monocular depth estimation methods for estimating aircraft altitude. It is worth noting that monocular depth estimation cannot accurately recover the physical altitude of aircraft from RS images. Therefore, the derived height should not be interpreted as a real-world altitude. It characterizes the relative depth contrast between aircraft and surrounding regions. Specifically, the cropped target region is fed into the well-trained depth estimation model to obtain a depth map. Let the average depth value of the four corner points in the depth map be d_{corner} and the maximum depth value be d_{max} , then the estimated height is calculated as:

$$H = d_{\text{max}} - d_{\text{corner}} \quad (6)$$

Accordingly, we define a three-dimensional geometric feature vector represented as:

$$\mathbf{g} = [L, W, H] \in \mathbb{R}^3 \quad (7)$$

(2) Visual-Text Semantic Similarity:

Relying solely on geometric features to generate soft labels is far from enough, just as it does not consider the apparent structure of the aircraft. In practical scenarios, many aircraft models (e.g., Boeing 737 and Airbus A320, or Boeing 777 and Boeing 787) exhibit highly similar physical dimensions, including fuselage length, wingspan, and aspect ratio. Therefore, when encountering situations in which the size is the same but the type differs, prototypes constructed based on geometric features often exhibit significant ambiguity.

To overcome this limitation of purely geometric representations, we leverage the cross-modal representation capability of VLMs, specifically CLIP [28], and introduce semantic features as complementary supervisory information. Specifically, we construct a unified textual description of an aircraft (e.g., “a high-resolution RS image of an aircraft”) and

encode it into a general text embedding using the VLM's pretrained text encoder. For the visual feature $\mathcal{F}_v \in \mathbb{R}^{512}$ extracted from any fine-grained target, we explicitly calculate its cosine similarity with the general text embedding. This similarity quantifies the degree to which the visual representations of all fine-grained categories lie within the same cross-modal semantic feature space. It enables the model to effectively perceive subtle appearance perturbations between different fine-grained subclasses, thereby extracting more inherently discriminative relative semantic features.

(3) Prior Prototype Construction and Soft Label Generation:

After fusing geometric features and semantic similarities, we assume that features of each class in the embedding space follow a Gaussian distribution. We then construct a Gaussian distribution-based probabilistic prototype (p_c, σ_c) , where the mean p_c represents the class center prototype and the variance σ_c captures intra-class variation and uncertainty.

$$p_c = \frac{1}{|\mathcal{D}_{\text{fine}}^c|} \sum_{j \in \mathcal{D}_{\text{fine}}^c} f_j, \quad (8)$$

$$\sigma_c = \sqrt{\frac{1}{|\mathcal{D}_{\text{fine}}^c|} \sum_{j \in \mathcal{D}_{\text{fine}}^c} \|f_j - p_c\|_2^2}, \quad (9)$$

For any sample i without fine-grained annotations, its feature representation f_i is extracted, and its the normalized Euclidean distance to each category prototype $d_{i,c}$ is computed.

$$d_{i,c} = \frac{\|f_i - p_c\|_2}{\sigma_c}, \quad (10)$$

Finally, the distance values are converted into a soft label distribution using the Softmax function:

$$y_{i,c}^{\text{soft}} = \frac{\exp(-d_{i,c})}{\sum_{c'=1}^C \exp(-d_{i,c'})} \quad (11)$$

where C denotes the number of fine-grained categories. These prototypes remain fixed after construction. The generated soft labels serve as pseudo-supervision signals for subsequent weakly supervised training. The overall procedure of the proposed prior-knowledge-based soft-label generation is summarized in Algorithm 1.

Algorithm 1 Prior-Knowledge-based Soft-label Generator

Require: Fine-grained prior set $\mathcal{D}_{\text{fine}}$, coarse-grained set $\mathcal{D}_{\text{coarse}}$

Ensure: Soft labels $\{y_i^{\text{soft}}\}_{i \in \mathcal{D}_{\text{coarse}}}$ for coarse samples

```

1: Stage 1: Prior prototype construction
2: for each fine-grained prior sample  $j \in \mathcal{D}_{\text{fine}}$  do
3:   Extract geometric features  $\mathbf{g}_j = [L_j, W_j, H_j]$ , where  $H_j = d_{\text{max}} - d_{\text{corner}}$ 
4:   Compute semantic similarity vector  $\mathbf{s}_j$ 
5:   Form fused feature  $f_j = [\mathbf{g}_j; \mathbf{s}_j]$ 
6: end for
7: for each fine-grained category  $c = 1, \dots, C$  do
8:   Initialize prior prototype  $p_c = \frac{1}{|\mathcal{D}_{\text{fine}}^c|} \sum_{j \in \mathcal{D}_{\text{fine}}^c} f_j$  {Fixed, not updated}
9:   Estimate category scale  $\sigma_c$  according to Equation (9)
10: end for
11: Stage 2: Soft label generation for coarse samples
12: for each coarse-grained sample  $i \in \mathcal{D}_{\text{coarse}}$  do
13:   Extract fused feature  $f_i$ 
14:   Compute normalized Euclidean distance  $d_{i,c}$  according to Equation (10)
15:   Generate soft label  $y_i^{\text{soft}}$  according to Equation (11)
16: end for

```

4.2.2. Text Prototype Alignment Strategy

Introducing contrastive learning into fine-grained detection tasks and employing prototype learning to optimize the feature-space distribution are effective approaches to enhancing the discriminative capability of RS fine-grained object detection. Prototype learning, by forcing intra-class samples to be more compact and inter-class samples more separated, naturally aligns with the demand for highly discriminative features in fine-grained detection. However, existing prototype-based methods heavily rely on the extraction of discriminative visual features. In RS fine-grained scenarios, different categories often exhibit only subtle structural differences and highly similar visual appearances, making it difficult to fully characterize category prototypes using visual features alone.

To alleviate this issue, we incorporate semantic information to address the limitations of visual representations. Textual descriptions provide higher-level semantic priors that enhance semantic discriminability among fine-grained categories. Nevertheless, existing VLMs are primarily pretrained on generic natural-image datasets and lack adaptation to RS scenarios. Moreover, they still struggle to distinguish subtle inter-class differences in fine-grained tasks. Therefore, we propose a TPA strategy to construct a shared semantic space tailored for RS fine-grained object detection.

Specifically, we first construct a text description for each fine-grained category and use it as a text prompt for the VLM text encoder, as illustrated in Table 1, thereby generating a text embedding with dimension N_t . Meanwhile, a well-trained object detector is employed to extract visual RoI features from M annotated fine-grained samples. These visual features are further pooled and interpolated into the same N_t -dimensional space to obtain corresponding visual representations.

Table 1. Examples of category-specific textual prompts used for text embedding generation.

Category	Text Prompt
Boeing737	a narrow-body twinjet airliner with characteristic winglets and underwing engines
Boeing747	a wide-body four-engine jumbo jet with a distinctive hump upper deck
Boeing777	a wide-body twinjet airliner with six-wheel main landing gear and large underwing engines
Boeing787	a wide-body twinjet airliner with a smooth nose contour and large cockpit windows
C919	a narrow-body twinjet airliner with winglets
A220	a narrow-body twinjet aircraft with a distinctive pointed nose cone and rear-mounted engines
A321	the longest variant of the A320 family with an elongated fuselage
A330	a wide-body twinjet airliner with distinctive cockpit windows and wingtip fences
A350	a wide-body twinjet airliner with black cockpit windows and curved carbon-fiber wingtips
ARJ21	a regional jet airliner with a T-tail configuration and aft-mounted engines
other-airplane	a generic commercial airplane with standard aircraft configuration

After obtaining the text embeddings and visual representations, we further introduce a learnable alignment matrix:

$$\mathbf{W} \in \mathbb{R}^{N_t \times N_t}$$

to jointly project text and visual features into a shared semantic space. The transformation is expected to satisfy two objectives simultaneously: (1) enforcing cross-modal consistency between visual representations and text embeddings belonging to the same category, and (2) enhancing the discriminability among text embeddings from different categories. Based on these objectives, we design a cross-modal alignment loss and a discriminative loss to jointly optimize the alignment matrix $\mathbf{W}$.

$$\mathcal{F}'_t = \mathbf{W}\mathcal{F}_t, \quad \mathcal{F}'_v = \mathbf{W}\mathcal{F}_v \quad (12)$$

First, to enhance semantic consistency between visual features and their corresponding text embeddings, we define the cross-modal alignment loss as:

$$\mathcal{L}_{\text{align}} = -\frac{1}{|\mathcal{D}_{\text{fine}}|} \sum_{j \in \mathcal{D}_{\text{fine}}} \log \frac{\exp(\mathcal{F}'_{v,j} \mathcal{F}'_{t,c_j} / \tau)}{\sum_{k=1}^C \exp(\mathcal{F}'_{v,j} \mathcal{F}'_{t,k} / \tau)} \quad (13)$$

where $\mathcal{F}'_{v,j}$ denotes the j -th transformed visual feature, and $\mathcal{F}'_{t,k}$ denotes the transformed text embedding of the k -th category. C is the number of categories and τ is the temperature parameter. τ is empirically set to 0.5. $\mathcal{D}_{\text{fine}}$ denotes the set of M annotated fine-grained training samples. This loss follows the InfoNCE-style, encouraging visual features to be close to their corresponding text embeddings while pushing them away from non-matching categories in the shared semantic space.

Furthermore, to improve the discriminability among different text embeddings, we introduce a margin-based discriminative loss:

$$\mathcal{L}_{\text{dis}} = \frac{1}{C} \sum_{i \neq j} \max\left(0, \frac{\mathcal{F}'_{t,i} \mathcal{F}'_{t,j}}{\|\mathcal{F}'_{t,i}\|_2 \|\mathcal{F}'_{t,j}\|_2} - m\right) \quad (14)$$

where $\mathcal{F}'_{t,i}$ and $\mathcal{F}'_{t,j}$ denote the transformed text embeddings of different categories. m is a margin hyperparameter used to constrain the cosine similarity between heterogeneous category embeddings. Given the high semantic similarity among fine-grained categories, we empirically set $m = 0.5$, a value commonly used in fine-grained object detection tasks.

This loss enforces a margin constraint on the cosine similarity between different text embeddings, penalizing only those pairs whose similarity exceeds the margin threshold. As a result, it explicitly reduces semantic overlap among categories and improves the structural separability of the text prototypes.

Finally, the overall optimization objective for the alignment matrix $\mathbf{W}$ is formulated as:

$$\mathcal{L}_W = \lambda_{\text{align}} \mathcal{L}_{\text{align}} + \lambda_{\text{dis}} \mathcal{L}_{\text{dis}} \quad (15)$$

where λ_{align} and λ_{dis} are weighting factors that balance the cross-modal alignment loss and the prototype discriminative loss, respectively. By minimizing the above objective, an effective shared semantic space suitable for RS fine-grained scenarios can be obtained. Subsequently, in the Stage II weakly supervised training, the alignment matrix ($\mathbf{W}$) and the adapted text prototypes are fixed. A contrastive learning strategy is then employed to guide visual features toward their corresponding text prototypes $\mathcal{P}_c$, thus improving RS fine-grained object detection performance.

4.3. Stage II: Weakly Supervised Training

After obtaining the soft labels, the next step is to train the model in a WSL manner. We will introduce two key modules: the FPM and the Loss Function.

4.3.1. Feature Purification Module (FPM)

Due to the unique geometry of aircraft, the proposed regions often contain substantial background content. Feature maps are prone to isolated “spike” noise due to overactivation within local receptive fields and interference from complex airport backgrounds during feature extraction. To obtain more robust contextual representations, inspired by morphological opening operations, we propose a background-suppressed Feature Purification Module.

Given an input RoI feature map $\mathcal{F}_{\text{RoI}} \in \mathbb{R}^{H_{\text{RoI}} \times W_{\text{RoI}} \times N_{\text{RoI}}}$, we approximate morphological erosion and dilation using 3×3 Max Pooling operations:

$$\mathcal{F}_{\text{BS}} = \text{MaxPool}_3(-\text{MaxPool}_3(-\mathcal{F}_{\text{RoI}})) \quad (16)$$

After background suppression, we enhance the purified feature $\mathcal{F}_{\text{BS}}$ by employing a 3×3 convolution to capture local detailed information, followed by a 1×1 convolution for channel fusion.

$$\mathcal{F}_{\text{LE}} = \text{Conv}_{1 \times 1}(\text{ReLU}(\text{Conv}_{3 \times 3}(\mathcal{F}_{\text{BS}}))) \quad (17)$$

Combined with residual connections, the forward propagation of FPM is expressed as:

$$\mathcal{F}_{\text{P}} = \mathcal{F}_{\text{RoI}} + \text{Norm}(\mathcal{F}_{\text{LE}}) \quad (18)$$

With this design, the FPM effectively suppresses background noise and enhances object features.

4.3.2. Overall Loss Function

The overall end-to-end training objective of the detector is defined as

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_{\text{reg}} \mathcal{L}_{\text{reg}} + \lambda_{\text{con}} \mathcal{L}_{\text{con}} \quad (19)$$

Regarding classification and localization losses, the implementation can be based on Formulas (4) and (5). λ_{con} is the weight of the contrastive loss.

For contrastive learning, we first extract the RoI feature after purifying $\mathcal{F}_{\text{P}} \in \mathbb{R}^{H_{\text{RoI}} \times W_{\text{RoI}} \times N_{\text{RoI}}}$ from FPM.

The RoI feature is flattened into a vector representation and passed through the head consisting of shared fully connected layers to obtain a compact embedding:

$$\mathcal{F}'_{\text{P}} = \text{FC}(\text{Flatten}(\mathcal{F}_{\text{P}})), \mathcal{F}'_{\text{P}} \in \mathbb{R}^{N_v}. \quad (20)$$

After that, the MLP projection head is applied to further map the feature into a lower-dimensional embedding space:

$$\mathcal{F}_{\text{mlp}} = \text{MLP}(\mathcal{F}'_{\text{P}}), \mathcal{F}_{\text{mlp}} \in \mathbb{R}^{N_t}. \quad (21)$$

Finally, a pretrained and fixed alignment matrix $\mathbf{W} \in \mathbb{R}^{N_t \times N_t}$ (learned in Stage I) is used to map visual features into the shared semantic space:

$$\mathcal{F}_{\text{con}} = \mathbf{W} \mathcal{F}_{\text{mlp}}. \quad (22)$$

For the contrastive loss $\mathcal{L}_{\text{con}}$, the cosine similarity between the feature $\mathcal{F}_{\text{con},i}$ and all adapted text prototypes $\mathcal{P}$ is computed, and the most similar prototype is selected as the positive sample c_i :

$$c_i = \arg \max_c \frac{\mathcal{F}_{\text{con},i} \cdot \mathcal{P}_c}{\|\mathcal{F}_{\text{con},i}\| \|\mathcal{P}_c\|} \quad (23)$$

The contrastive loss is formulated as:

$$\mathcal{L}_{\text{con}} = -\log \frac{\exp(\mathcal{F}_{\text{con},i} \cdot \mathcal{P}_{c_i} / \tau)}{\sum_{c=1}^C \exp(\mathcal{F}_{\text{con},i} \cdot \mathcal{P}_c / \tau)} \quad (24)$$

5. Experiments

5.1. Dataset

We conduct experiments on the FAIR1M-Aircraft dataset [1], an RS image collection designed for fine-grained aircraft detection. The dataset comprises 7895 images with sizes ranging from 600 to 10,000 pixels. It contains approximately 46,000 instances across 11 fine-grained aircraft categories. As Figure 3 shows, they are Boeing 737, Boeing 747, Boeing 777, Boeing 787, ARJ21, C919, A220, A321, A330, A350, and other airplanes. As Figure 4 shows, there exists a severe long-tail distribution. The majority class (“other-airplane”) contains over 13,000 samples, while the rarest classes (C919, ARJ21) have fewer than 200 samples each. This imbalance presents a significant challenge for few-shot and weakly supervised fine-grained detection.



Figure 3. Visualization of all categories of aircraft in the FAIR1M-Aircraft dataset.

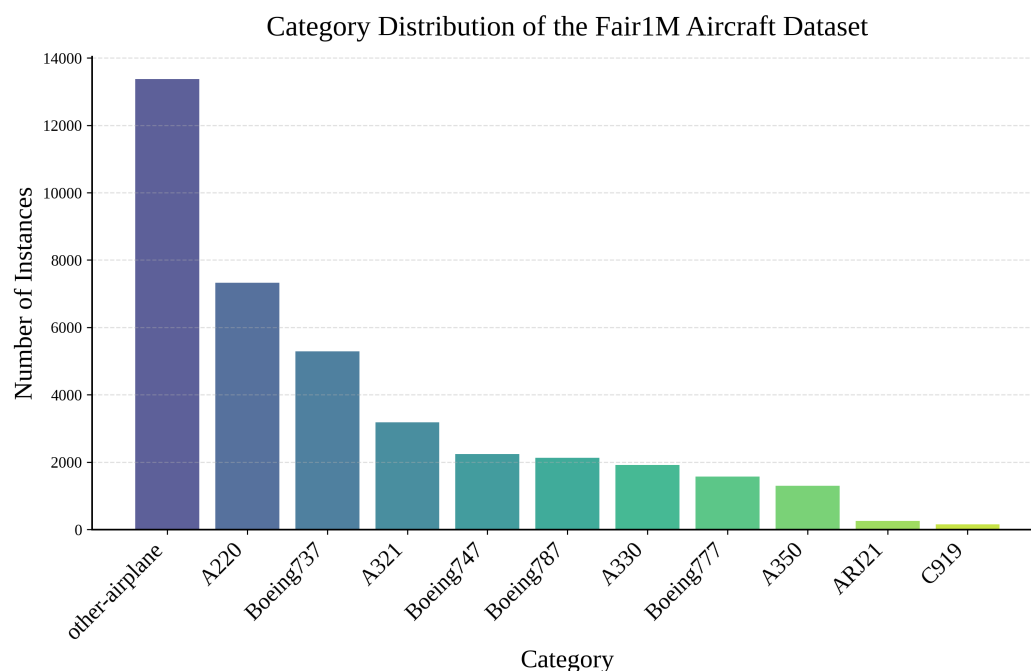


Figure 4. Category distribution of the FAIR1M-Aircraft dataset. The dataset comprises 11 fine-grained aircraft categories, totaling nearly 46,000 instances.

5.2. Experimental Setup

All experiments are conducted on a Linux platform using PyTorch 1.13. Training is performed on a single NVIDIA RTX 4090 GPU with 24 GB of memory. We adopt ORCNN [15] with a ResNet-50 backbone [53] pretrained on ImageNet [54] as the base detector for all experiments. The model parameters are optimized using the SGD optimizer. The initial learning rate, weight decay, temperature parameter τ , and momentum are set

to 0.005, 0.0001, 0.5, and 0.9, respectively. The total number of training epochs is set to 12, and the learning rate is decayed at the 8th and 11th epochs. The batch size is set to 2. The non-maximum suppression (NMS) threshold and IoU threshold are set to 0.1 and 0.5, respectively. Mean Average Precision (mAP), which is widely used in object detection tasks, is adopted as the evaluation metric.

We employ CLIP [28] as the vision-language model. Specifically, its text encoder extracts semantic text embeddings from category-specific textual descriptions. For visualization, we used the RoI feature extractor from the pretrained ORCNN detector to obtain region-level visual features, which serve as the visual branch for cross-modal alignment. To incorporate height priors, we employ Depth Anything V2 [55] as the monocular depth estimation model. The generated depth maps are further used to estimate object height, providing complementary geometric information for fine-grained aircraft recognition. For weakly supervised fine-grained detection, we randomly sample M fine-grained samples from the training set. The sampled fine-grained subset is constructed by first ensuring that each aircraft category is represented by at least one annotated instance; the remaining samples are then randomly selected to match the original class distribution. Multiple selections of M have been evaluated to assess the method's effectiveness in providing a few samples across various proportions. The remaining $N - M$ training images are treated as coarsely annotated samples. For each fine-grained category, the corresponding category description listed in Table 1 is adopted as the text prompt to generate fine-grained text embeddings.

For evaluation, we employ the official FAIR1M-Aircraft [1] test benchmark. During training, three data augmentation strategies, including horizontal, vertical, and diagonal flipping, are applied. For a fair comparison, all competing methods adopt the same data augmentation policy.

5.3. Ablation Study

To investigate the contribution of individual components, we have conducted extensive ablation studies. To ensure fairness, all experiments follow the same hyperparameter configuration.

(1) Effects of inter-modular ablation on model performance:

As shown in Table 2, we systematically evaluate the marginal contributions of each core component to the overall model performance. To establish clear performance boundaries, we train the model using supervised learning on $M = 50$ labeled samples. Its experimental results are presented as a baseline in Table 2. The experimental results show that its mAP is only 20.4%, confirming that the feature generalization capability of general detectors is severely limited when fine-grained annotations are extremely scarce, making it inadequate for challenging fine-grained object detection tasks.

Table 2. Ablation study of key components.

Setting	PSG	TPA	FPM	mAP (%)
Baseline	×	×	×	20.4
Ours	✓	×	×	28.4
	✓	✓	×	29.4
	✓	×	✓	29.2
	✓	✓	✓	30.5

In contrast, our proposed method adopts a WSL paradigm. We first construct a soft label distribution from the provided few annotations and then generalize it to a larger dataset. Thanks to this, the model can be trained in a supervised manner on datasets with

a large number of unlabeled samples. The experimental results showed that after adopting the PSG strategy, the model's mAP increased to 28.4%, a significant 8.0% increase. This convincingly demonstrates that mining multi-source priors and weakly supervised training on large amounts of unlabeled data can significantly mitigate the performance loss caused by the lack of fine-grained annotations.

In addition, we validated the sample prototype alignment and Feature Purification Modules separately. They steadily improved the model performance to 29.4% and 29.2%, respectively, indicating that each module contributed significantly to overall performance. In the end, the complete framework integrating all components achieved 30.5% of the optimal mAP. This result not only significantly outperforms the baseline (+10.1%), but also fully validates the synergistic effect of the three parts in weakly supervised fine-grained learning tasks.

(2) Ablation study of the PSG Strategy:

The PSG strategy aims to generate soft-label information by leveraging multimodal prior information to extend labeling to a large number of unlabeled samples. As shown in Figure 5, different types of prior information contribute proportionally to the accuracy of the generated soft labels. We use the target's original geometric features as a baseline to evaluate the incremental benefits of the additional priors we introduce. As shown in Table 3, combining altitude information increases mAP by 0.8%, underscoring its key role in distinguishing fine-grained aircraft models with similar sizes but different heights. In addition, integrating semantic priors based on VLMs can yield an additional 0.4% benefit. Ultimately, the synergistic effect of multi-source priors (geometry and semantics) enables the model to achieve an optimal mAP of 28.4%. This indicates that our proposed PSG strategy effectively bridges the “supervision gap” in weakly supervised fine-grained object detection.

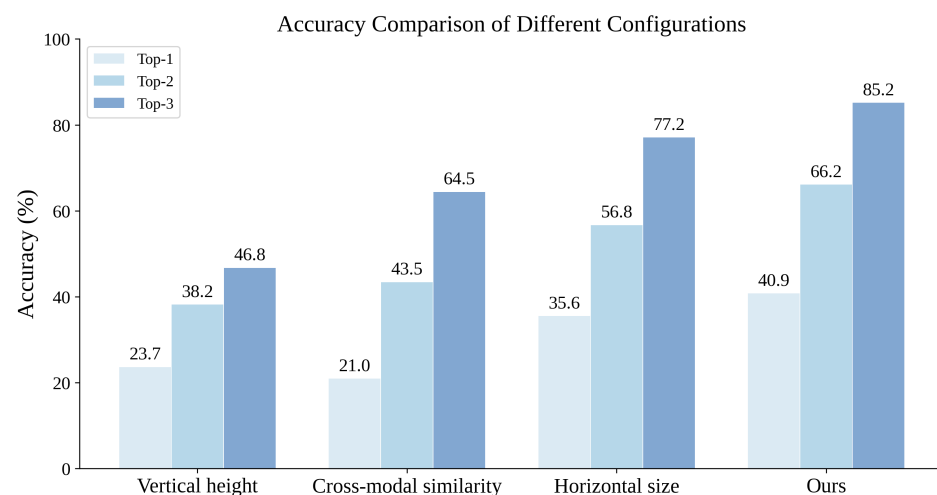


Figure 5. Accuracy comparison of different configurations. The x-axis denotes different configurations, including vertical height, cross-modal similarity, horizontal size, and our method.

Table 3. Ablation study of soft-label generation strategies. The bold results represent the best performance.

	Height	Cross-Modal Similarity	mAP (%)
Baseline	×	×	27.4
PSG	✓	×	28.2
	×	✓	27.8
	✓	✓	28.4

To evaluate the reliability of the derived height feature obtained from monocular depth estimation, we provide qualitative and quantitative analyses. Figure 6 visualizes the depth maps generated by Depth Anything V2. Although the estimated depth values do not correspond to physical aircraft altitudes, aircraft regions consistently exhibit distinct relative depth responses compared with the surrounding background, indicating that the model captures meaningful structural differences. Figure 7 further presents the distribution of the derived height feature across different aircraft categories. The results show category-dependent variations, where larger aircraft generally exhibit higher derived height values than smaller aircraft. These observations suggest that the proposed feature contains useful geometric information related to aircraft scale and structure.

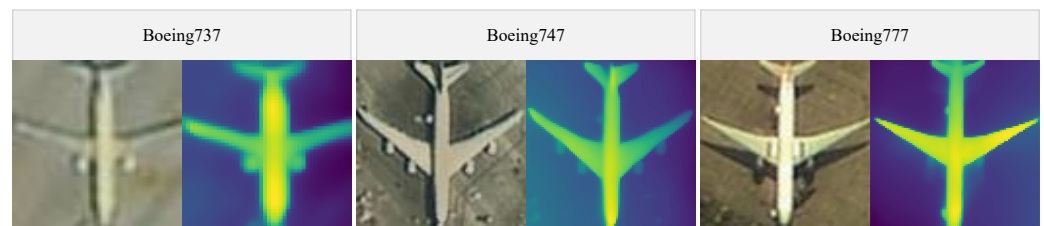


Figure 6. Visualization of monocular depth estimation results generated by Depth Anything V2. Columns 2, 4, and 6 show the predicted depth, with darker colors indicating greater distances.

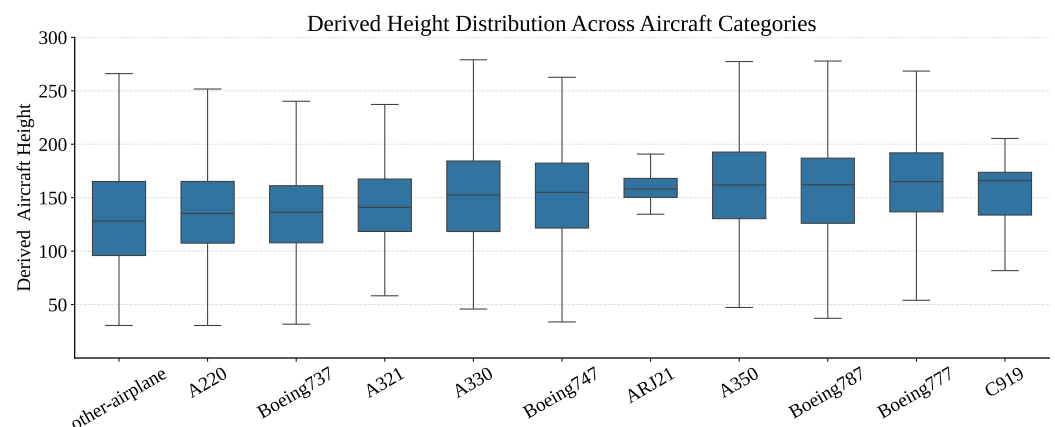


Figure 7. Distribution of derived height features among different aircraft categories. The derived feature shows category-dependent geometric characteristics. The vertical axis denotes relative depth values without physical units.

Furthermore, we investigate the impact of the number of labeled samples M used for soft-label generation, as reported in Table 4. In the extremely one-shot labeling with only $M = 11$ target instances (one per category), the baseline model cannot be trained or evaluated because it relies on fully supervised learning and is therefore denoted as “—”. In contrast, our method achieves 23.4% mAP without requiring additional annotations, demonstrating strong adaptability in scarce-data scenarios. As the number of M increases, performance consistently improves, reaching its optimum at $M = 250$ images (5% of the whole dataset). When the sample size is further increased to $M = 500$ (10% of the whole dataset), a slight performance degradation is observed. Nevertheless, it still outperforms supervised learning methods. We analyze that too many samples may introduce instances with low representativeness or high noise. This will affect the quality of the generated soft labels and the discriminative power of the prior-knowledge prototypes.

Table 4. Effect of the number of labeled samples M used. The two "-" in the first row indicate that the sampling rate was not calculated and that the baseline model could not be trained. The bold results represent the best performance.

Sample Size	Sampling Ratio	mAP (%)	
		Baseline	Ours
11 target instances	—	—	23.4
50 images	1%	20.4	30.5
250 images	5%	24.9	31.4
500 images	10%	28.0	30.2

(3) Ablation study of the Text Prototype Alignment (TPA):

As shown in Table 5, we conduct an ablation study on the proposed Text Prototype Alignment (TPA) strategy. First, we evaluate a baseline in which the original text embeddings are used directly as category prototypes without applying TPA. This baseline achieves a mAP of 28.9%, which already improves over training with only soft labels by 0.5 mAP. This result indicates that text embeddings generated from category-specific textual descriptions using VLMs exhibit some discriminative power.

Table 5. Ablation study of loss weight coefficients in the proposed TPA strategy.

	λ_{align}	λ_{dis}	mAP (%)
Baseline	—	—	28.9
TPA	0.1	0.9	28.5
	0.3	0.7	28.3
	0.5	0.5	29.1
	0.7	0.3	29.4
	0.9	0.1	28.0

We perform a systematic ablation on the two loss weight coefficients in TPA, λ_{align} and λ_{dis} , while keeping their sum fixed at 1. The results show that the best performance is achieved with a 7:3 ratio, yielding a mAP of 29.4%, corresponding to an additional 0.5 mAP gain over the baseline. This demonstrates that the proposed TPA strategy effectively enhances the expressiveness of the text prototypes, thereby improving the discriminability of fine-grained aircraft categories in remote sensing object detection.

Furthermore, we analyzed the contrastive loss weight of the Contrast Head. As shown in Figure 8, the model's performance increases with the contrastive loss weight λ_{con} , exhibiting a convex trend. Specifically, its mAP fluctuates slightly between 29 and 30, reflecting the stability of the contrastive learning process. The model achieves the best performance when $\lambda_{\text{con}} = 0.8$. Therefore, we set it as the default value for all experiments.

Finally, we investigate the sensitivity of the margin parameter m in the discriminative loss. As shown in Figure 9, the detection performance first increases and then decreases as m becomes larger. When the margin is too small, the discriminative constraint between categories is insufficient, resulting in limited inter-class separation. Conversely, an excessively large margin imposes overly strict constraints and may hinder optimization. The best performance is achieved at $m = 0.5$, yielding 29.4% mAP.

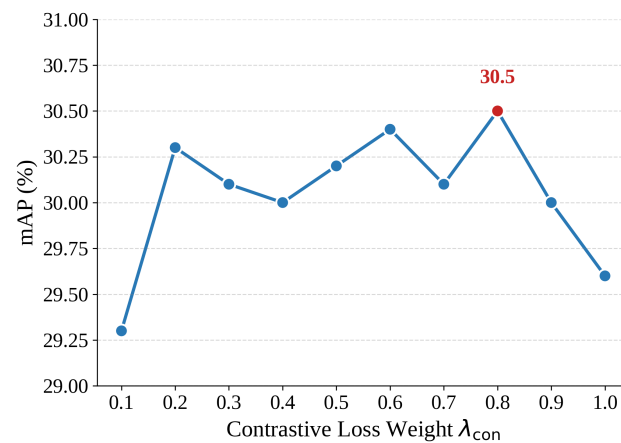


Figure 8. The impact of the contrastive loss weight λ_{con} on the detection performance (mAP). The red dot represents the highest value.

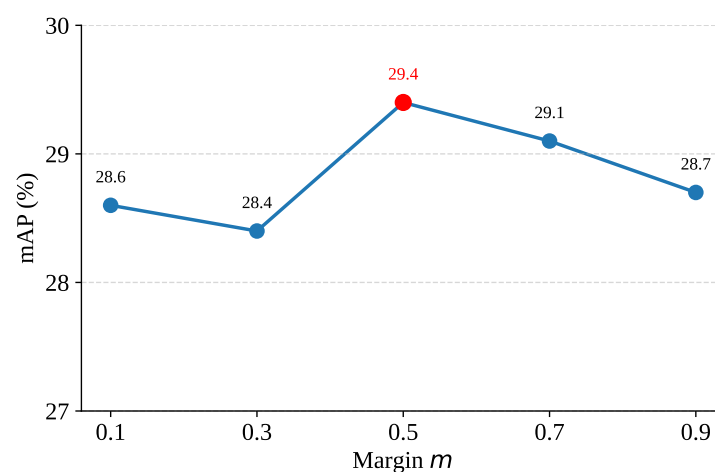


Figure 9. Sensitivity analysis of the margin parameter m in the discriminative loss. The red dot represents the highest value.

(4) Ablation study of the Feature Purification Module (FPM):

During feature extraction, to suppress background noise and enhance the target's contextual features, we designed FPM. Specifically, we compared three configurations: only background suppression (BS), only local enhancement convolution (LE), and the full FPM. As shown in Table 6, compared with the original baseline, the introduction of the BS alone can increase mAP by 0.8%, confirming the necessity of effective inhibition in complex backgrounds. Meanwhile, the Refine module achieved a 1.1% performance gain over the baseline. It helps the model learn discriminative features of the target by enhancing its perception of local details. When the two are combined to form a complete FPM, the model's mAP reaches 30.5%. This indicates a significant complementary effect between noise suppression and feature enhancement, which helps to learn more discriminative feature representations.

To investigate the influence of the pooling kernel size in the morphology-guided background suppression operation, we conduct a sensitivity analysis with different kernel sizes, including 1×1 , 3×3 , and 5×5 . As shown in Table 7, the proposed FPM achieves the best performance with a 3×3 kernel, obtaining an mAP of 30.5%. The 1×1 setting, which omits the background suppression operation, reduces performance to 30.2%, demonstrating the effectiveness of morphological filtering. Meanwhile, the 5×5 kernel slightly reduces the mAP to 30.3%, suggesting that excessive spatial aggregation may weaken the discriminative

power of target details. Therefore, the 3×3 kernel is adopted as it provides a better trade-off between background suppression and feature preservation.

Table 6. Ablation study of the Feature Purification Module (FPM) components. The bold results represent the best performance.

	Background Suppressed	Local Enhancement	mAP (%)
Baseline	×	×	29.4
FPM	✓	×	29.9
	×	✓	30.2
	✓	✓	30.5

Table 7. Sensitivity analysis of pooling kernel size in FPM. The bold results represent the best performance.

Module	Kernel Size	mAP (%)
FPM	1×1 (<i>w/o</i> BS)	30.2
	3×3	30.5
	5×5	30.3

In addition, feature map visualizations are further provided to demonstrate the effectiveness of the proposed FPM. As illustrated in Figure 10, the baseline model without FPM exhibits dispersed activations with considerable background noise and irrelevant responses distributed around the target region. In contrast, after introducing the FPM, the feature responses become significantly more concentrated on the aircraft body while the surrounding background activations are effectively suppressed. This phenomenon indicates that the proposed FPM can purify noisy RoI features and enhance target-aware representations, thereby enabling the network to focus more accurately on discriminative aircraft structures.

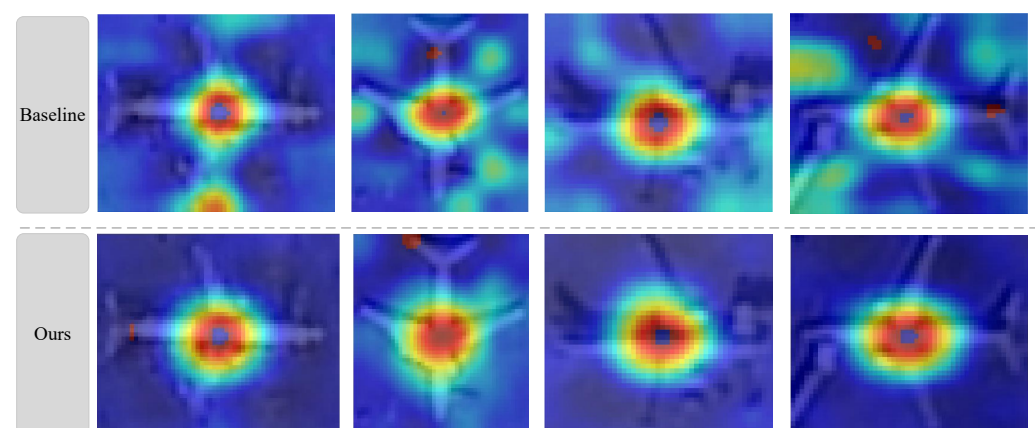


Figure 10. Visualization results of the RoI feature map of the aircraft target. The first row is the baseline model, and the second is equipped with FPM.

5.4. Comparisons Study

To comprehensively evaluate the superiority of our proposed method, we have conducted systematic comparative experiments across a range of mainstream detector architectures. Specifically, we instantiated the proposed weakly supervised method using the two-level detector ORCNN. Meanwhile, we introduce other representative two-stage detectors, namely ReDet [16] and Faster R-CNN [19], as well as the single-stage detector Rotating FCOS [56] (R-FCOS), for comparison. The experiment covered two different

supervised paradigms: supervised learning and WSL (classification branch). Table 8 lists the quantitative performance of the proposed method and the comparison method.

Table 8. Performance comparison under different supervision settings on fine-grained airplane dataset (mAP, %)

Class	Supervised Learning				Weakly Supervised Learning (M Samples Provided)				
	ORCNN [15]	ReDet [16]	Faster R-CNN [19]	R-FCOS [56]	ORCNN [15]	ReDet [16]	Faster R-CNN [19]	R-FCOS [56]	Ours
Boeing737	38.9	41.4	38.6	38.3	26.7	24.9	18.7	22.3	38.0
Boeing747	79.6	82.7	80.9	80.3	31.1	34.3	15.1	23.8	49.5
Boeing777	21.5	27.2	15.4	20.3	10.3	9.6	1.9	12.3	8.8
Boeing787	44.6	43.3	42.6	49.6	19.1	21.4	12.7	14.3	31.8
C919	10.7	1.8	10.6	0.5	0.0	0.0	0.0	0.0	0.0
A220	41.7	46.9	38.8	39.9	29.9	23.9	22.8	24.9	39.1
A321	53.6	55.9	52.2	56.3	22.2	14.1	4.2	11.6	39.8
A330	48.5	44.0	41.0	46.5	16.1	16.1	10.4	10.7	31.5
A350	36.9	35.3	39.8	48.8	18.8	9.3	5.0	4.5	27.3
ARj21	18.1	18.5	18.2	15.4	0.0	0.0	0.0	3.5	3.6
other-air	70.7	73.2	70.9	72.4	50.5	49.6	33.4	31.3	66.7
mAP(%)	42.2	42.8	40.8	42.6	20.4	18.5	11.3	14.5	30.5

Under the fully supervised setting, all detectors achieve relatively satisfactory detection performance on fine-grained aircraft categories. Among them, ReDet ranks first with a mAP of 42.8%, while ORCNN, FCOS, and Faster R-CNN achieve mAP values of 42.2%, 42.6%, and 40.8%, respectively. Notably, for common categories such as Boeing 737 and Boeing 747, all detectors perform well. For categories with extremely limited samples (e.g., C919 and ARJ21), the available visual variations and category-specific patterns are insufficient for reliable fine-grained learning; therefore, the low AP under such extreme conditions is expected.

Under WSL, the performance of all compared detectors drops substantially. The mAP values of ORCNN, ReDet, Faster R-CNN, and FCOS are only 20.4%, 18.5%, 11.3%, and 14.5%, respectively. Particularly in the C919 and ARJ21 categories, the AP of most detectors drops to 0%, indicating that few annotated samples alone are insufficient to enable existing rotated detectors to learn effective fine-grained feature representations. This result further underscores the challenges of insufficient generalization and the risk of overfitting that the WSL paradigm faces in fine-grained object detection tasks on RS imagery.

As the table lists, our method achieves an mAP of 30.5%, significantly better than the baseline under WSL (ORCNN, 20.4%), representing an improvement of 10.1%. Compared with the fully supervised ORCNN (42.2%), our method achieves approximately 72% of its performance while requiring only 1% of the fine-grained annotation cost of full supervision. It demonstrates the great potential of the weakly supervised paradigm in scenarios with limited annotation budgets. From a per-category perspective, our method substantially outperforms all compared detectors under WSL across multiple categories, including Boeing737, Boeing747, A220, A321, A330, and other-airplane. Notably, in the A321 category, our method achieves an AP of 39.8%, a significant improvement over the baseline (22.2%).

In addition, we compare the confusion matrices between our method and the baseline model (see Figure 11). It shows that our method outperforms the baseline model in classification performance. First, as illustrated in Figure 11b, the diagonal values of our model are generally higher than those of the baseline in Figure 11a. For instance, the classification accuracies for Boeing 747, Boeing 777, and Boeing 787 increase to 23%, 22%, and 21%, indicating stronger discriminative ability. Second, high-probability predictions in our model are more concentrated along the diagonal, and misclassifications exhibit a clear “block-wise” aggregation pattern. Errors mainly occur between semantically or structurally similar categories (e.g., within the Boeing family), while random cross-category errors are significantly reduced. This reasonable error distribution effectively enhances

model interpretability, thereby demonstrating the superiority of the proposed method in discriminative feature learning.

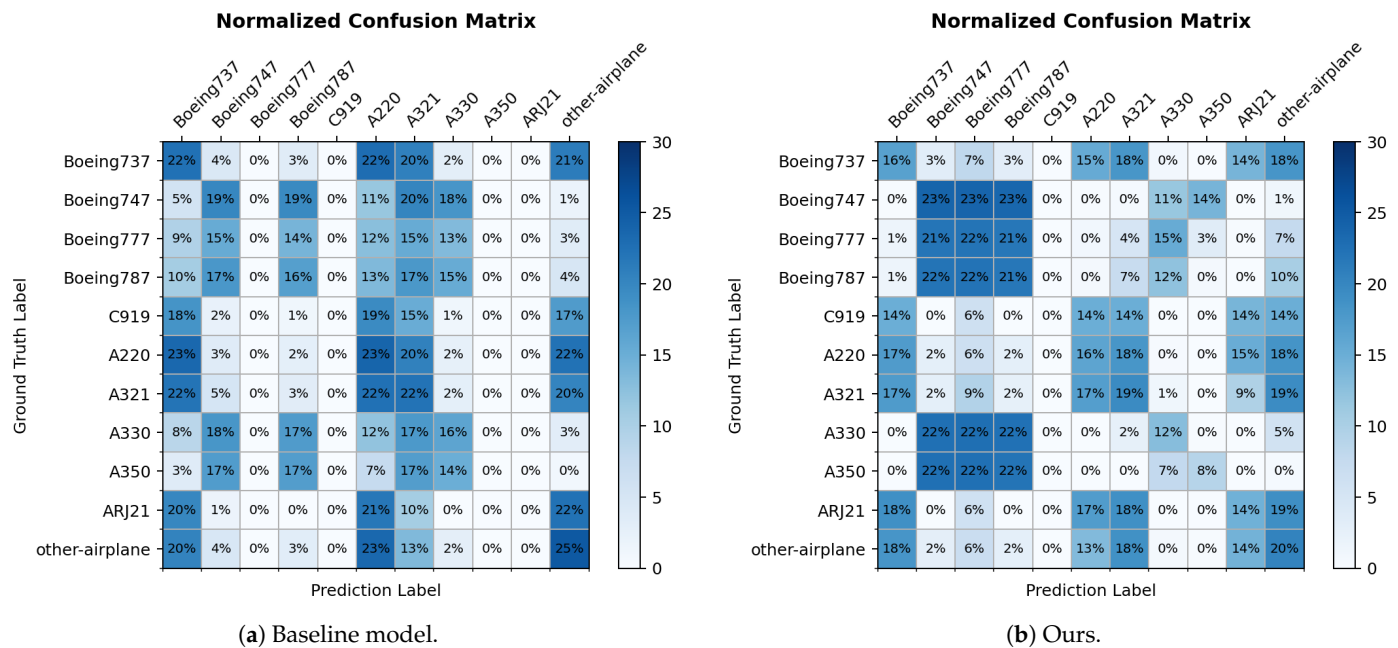


Figure 11. Comparison of fine-grained confusion matrices on the FAIR1M-Aircraft dataset.

Table 9 reports the computational cost of the proposed method. Compared with Oriented R-CNN, our method increases the number of parameters from 41.36 M to 43.59 M and GPU memory consumption from 3.63 GB to 4.11 GB. Although the training time per iteration increases from 0.28 s to 0.44 s, the inference speed remains largely unchanged (13.56 FPS vs. 13.16 FPS). This is because PSG is used only for offline soft-label generation, whereas TPA and FPM introduce only lightweight computations during inference. These results demonstrate that the proposed framework achieves significant performance gains with only a modest increase in computational cost.

Table 9. Comparison of computational costs.

Method	Params (M)	GPU Memory (GB)	Training Time (s/iter)	FPS (Images/s)
Oriented R-CNN	41.36	3.63	0.28	13.56
Ours	43.59	4.11	0.44	13.16

Figure 12 shows qualitative detection results on test images, compared with the baseline method. Compared to the baseline, our method significantly reduces misclassifications in both densely parked and structurally similar airplane scenarios.

In summary, the quantitative comparative experiments demonstrate the effectiveness and robustness of our proposed weakly supervised rotated object detection method for fine-grained aircraft recognition. Compared with supervised learning methods, our method significantly reduces the dependence on precisely rotated annotations. Compared with the WSL methods, our method achieves substantial improvements in detection accuracy, demonstrating promising prospects for practical application.



Figure 12. Qualitative detection results on test images. Green bounding boxes denote correct detections ($\text{IoU} \geq 0.5$ and accurate classification), while red boxes indicate incorrect detections (localization error or class mismatch). Compared with the baseline methods, the proposed method achieves higher detection accuracy.

5.5. Analysis of Error Cases

We analyze representative misclassification cases selected from highly confused category pairs in the confusion matrix. As shown in Figure 13, three major factors are identified: (1) Geometric ambiguity, where categories such as A330 and Boeing 777 exhibit highly similar geometric characteristics; (2) Semantic ambiguity, where visually similar categories (e.g., A220 and Boeing 737) contain limited discriminative details due to small object sizes; (3) Background interference caused by neighboring aircraft, shadows, and complex airport facilities.

The analysis indicates that most errors arise from highly similar aircraft categories and challenging background environments. These observations further validate the design motivations of PSG, TPA, and FPM, which respectively exploit more prior knowledge, vision-language semantic prototypes, and feature purification to enhance category discrimination and suppress background distractions. Nevertheless, categories with highly overlapping geometric and semantic characteristics remain difficult to distinguish, suggesting potential directions for future improvement.

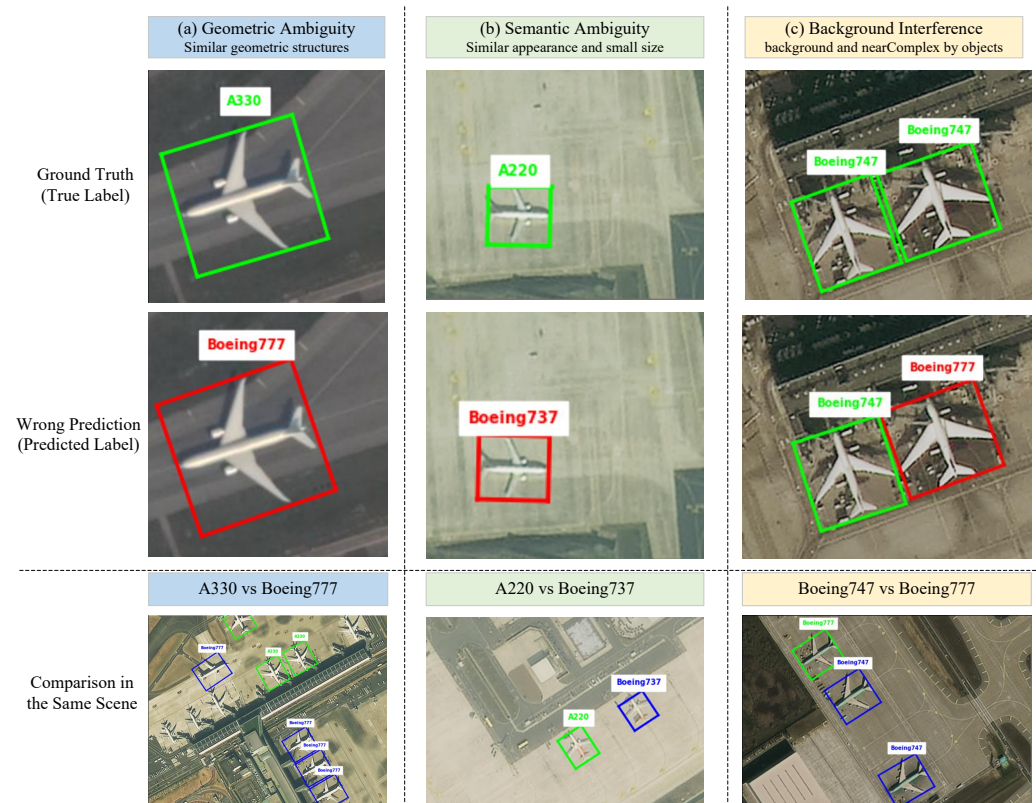


Figure 13. Examples of typical error cases in fine-grained aircraft detection, including geometric ambiguity, semantic ambiguity, and background interference.

6. Conclusions

In this paper, we have presented an innovative weakly supervised framework for fine-grained aircraft detection in remote sensing images. First, based on the exploration of multi-source prior knowledge about aircraft, we constructed a category-specific prototype library and extended it with a large number of unlabeled samples. Second, we introduced mainstream vision-language models and adjusted their text-output embeddings to enhance inter-class discrimination and to align visual features with text prototypes. Third, to mitigate background interference and inter-class similarity issues arising from geometric differences, we proposed FPM applied to aircraft RoI features. Extensive experiments on the FAIR1M-Aircraft dataset demonstrate that our method systematically outperforms supervised-learning-based detection frameworks, achieving a superior mAP of 30.5% with few fine-grained categorical labels. Our framework strikes a favorable trade-off by significantly reducing annotation costs while maintaining competitive performance.

Beyond aircraft, the proposed framework has potential for extension to other rigid RS targets. The PSG and TPA modules are generally applicable to categories with distinctive geometric and semantic characteristics, such as ships and vehicles. FPM is better tailored to aircraft due to their cross-like structures, which introduce substantial background interference; for more compact targets like ships, its benefits may be less pronounced.

Future research may focus on developing robust mechanisms for learning long-tailed distributions and enhancing the model's ability to extract locally discriminative features.

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

Funding: This work was supported in part by the National Natural Science Foundation of China under Grant 62401388 and in part by Jiangsu Provincial Department of Education under Grant 1020241530.

Data Availability Statement: All data that support the findings of this study are available from the corresponding author upon reasonable request.

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

References

1. 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. <https://doi.org/10.1016/j.isprsjprs.2021.12.004>.
2. Xia, G.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A Large-Scale Dataset for Object Detection in Aerial Images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2018; pp. 3974–3983.
3. Li, K.; Wan, G.; Cheng, G.; Meng, L.; Han, J. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS J. Photogramm. Remote Sens.* **2020**, *159*, 296–307. <https://doi.org/10.1016/j.isprsjprs.2019.11.023>.
4. Li, S.; Song, W.; Fang, L.; Chen, Y.; Ghamisi, P.; Benediktsson, J.A. Deep learning for hyperspectral image classification: An overview. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 6690–6709. <https://doi.org/10.1109/TGRS.2019.2907932>.
5. Chen, J.; Guo, P.; Liu, Y.; Cao, L.; Ran, D.; Wang, K.; Hu, W.; Wan, L. SPOD-YOLO: A Modular Approach for Small and Oriented Aircraft Detection in Satellite Remote Sensing Imagery. *Remote Sens.* **2025**, *17*, 3963. <https://doi.org/10.3390/rs17243963>.
6. Wang, J.; Bai, Z.; Zhang, X.; Qiu, Y. A Lightweight Remote Sensing Aircraft Object Detection Network Based on Improved YOLOv5n. *Remote Sens.* **2024**, *16*, 857. <https://doi.org/10.3390/rs16050857>.
7. El Ghazouali, S.; Gucciardi, A.; Venturini, F.; Venturi, N.; Rueegsegger, M.; Michelucci, U. FlightScope: An Experimental Comparative Review of Aircraft Detection Algorithms in Satellite Imagery. *Remote Sens.* **2024**, *16*, 4715. <https://doi.org/10.3390/rs16244715>.
8. Wang, Q.; Geng, X.; Li, P.; Zhang, L.; Niu, B.; Wang, F.; Zhou, G.; Hu, Y. Keypoint Detection-Based Aircraft Fine-Grained Recognition for High-Resolution Optical Remote Sensing Images. *Remote Sens.* **2025**, *17*, 3577. <https://doi.org/10.3390/rs17213577>.
9. Wan, H.; Nurmatamat, P.; Chen, J.; Cao, Y.; Wang, S.; Zhang, Y.; Huang, Z. Fine-Grained Aircraft Recognition Based on Dynamic Feature Synthesis and Contrastive Learning. *Remote Sens.* **2025**, *17*, 768. <https://doi.org/10.3390/rs17050768>.
10. Liang, W.; Li, J.; Diao, W.; Sun, X.; Fu, K.; Wu, Y. FGATR-Net: Automatic Network Architecture Design for Fine-Grained Aircraft Type Recognition in Remote Sensing Images. *Remote Sens.* **2020**, *12*, 4187. <https://doi.org/10.3390/rs12244187>.
11. Zhu, Z.; Kang, J.; Diao, W.; Wang, B.; Ni, J. Align and Complete Samples in Remote Sensing Fine-Grained Rigid Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5655617. <https://doi.org/10.1109/TGRS.2025.3635340>.
12. Gao, Y.; Deng, C.; Chen, L.; Zhu, Z. A Scale-Aware and Discriminative Feature Learning Network for Fine-Grained Rigid Object Recognition. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2025**, *18*, 1695–1705. <https://doi.org/10.1109/JSTARS.2024.3484411>.
13. 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>.
14. Huang, X.; Xu, K.; Huang, C.; Wang, C.; Qin, K. Multiple Instance Learning Convolutional Neural Networks for Fine-Grained Aircraft Recognition. *Remote Sens.* **2021**, *13*, 5132. <https://doi.org/10.3390/rs13245132>.
15. Xie, X.; Cheng, G.; Wang, J.; Yao, X.; Han, J. Oriented R-CNN for Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2021; pp. 3520–3529.
16. Han, J.; Ding, J.; Xue, N.; Xia, G.S. ReDet: A Rotation-equivariant Detector for Aerial Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2021.
17. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: Piscataway, NJ, USA, 2017; pp. 2980–2988.
18. Zhu, Z.; Sun, X.; Diao, W.; Chen, K.; Xu, G.; Fu, K. Invariant Structure Representation for Remote Sensing Object Detection Based on Graph Modeling. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5625217. <https://doi.org/10.1109/TGRS.2022.3181686>.

19. 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. <https://doi.org/10.1109/TPAMI.2016.2577031>.
20. Yang, X.; Zhang, G.; Li, W.; Wang, X.; Zhou, Y.; Yan, J. H2RBox: Horizontal Box Annotation is All You Need for Oriented Object Detection. In Proceedings of the International Conference on Learning Representations (ICLR), Kigali, Rwanda, 1–5 May 2023.
21. Yu, Y.; Yang, X.; Li, Q.; Zhou, Y.; Zhang, G.; Da, F.; Yan, J. H2RBox-v2: Incorporating Symmetry for Boosting Horizontal Box Supervised Oriented Object Detection. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), New Orleans, LA, USA, 10–16 December 2023; Volume 36, pp. 18218–18231.
22. Cao, G.; Yu, X.; Yu, W.; Han, X.; Yang, X.; Li, G.; Jiao, J.; Han, Z. P2RBox: Point Prompt Oriented Object Detection with SAM. *arXiv* **2023**, arXiv:2311.13128.
23. Yu, Y.; Yang, X.; Li, Q.; Da, F.; Dai, J.; Qiao, Y.; Yan, J. Point2RBox: Combine Knowledge from Synthetic Visual Patterns for End-to-end Oriented Object Detection with Single Point Supervision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2024.
24. Yu, Y.; Ren, B.; Zhang, P.; Liu, M.; Luo, J.; Zhang, S.; Da, F.; Yan, J.; Yang, X. Point2RBox-v2: Rethinking Point-supervised Oriented Object Detection with Spatial Layout Among Instances. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2025; pp. 19283–19293.
25. Zhang, T.; Fan, Z.; Liu, M.; Zhang, X.; Lu, X.; Li, W.; Zhou, Y.; Yu, Y.; Li, X.; Yan, J.; et al. Point2RBox-v3: Self-Bootstrapping from Point Annotations via Integrated Pseudo-Label Refinement and Utilization. *arXiv* **2025**, arXiv:2509.26281. <https://doi.org/10.48550/ARXIV.2509.26281>.
26. 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. <https://doi.org/10.1109/TGRS.2023.3290091>.
27. Li, Y.; Chen, L.; Li, W.; Bian, C.; Wang, N. Fine-Grained Ship Recognition With Spatial-Aligned Feature Pyramid Network and Adaptive Prototypical Contrastive Learning. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5403914. <https://doi.org/10.1109/TGRS.2025.3574821>.
28. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Brooks, T.; Dhariwal, P.; Nichol, A.; Child, R.; et al. Learning Transferable Visual Models From Natural Language Supervision. In *38th International Conference on Machine Learning (ICML)*; PMLR: Vienna, Austria, 2021; pp. 8748–8763.
29. Li, J.; Li, D.; Xiong, C.; Hoi, S.C. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *39th International Conference on Machine Learning (ICML)*; PMLR: Baltimore, MD, USA, 2022; pp. 12888–12900.
30. Li, J.; Li, D.; Savarese, S.; Hoi, S. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *40th International Conference on Machine Learning (ICML)*; PMLR: Honolulu, HI, USA, 2023; pp. 19730–19742.
31. Liu, H.; Li, C.; Li, Y.; Shen, Y.; Lee, Y.J. LLaVA: Large Language and Vision Assistant. *arXiv* **2023**, arXiv:2304.08485.
32. Serra, J. *Image Analysis and Mathematical Morphology*; Academic Press: Cambridge, MA, USA, 1982.
33. Wu, H.; Zhang, H.; Zhang, J.; Xu, F. Fast aircraft detection in satellite images based on convolutional neural networks. In *IEEE International Conference on Image Processing (ICIP)*; IEEE: Piscataway, NJ, USA, 2015; pp. 4210–4214.
34. Zhang, F.; Du, B.; Zhang, L.; Xu, M. Weakly supervised learning based on coupled convolutional neural networks for aircraft detection. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 5553–5563. <https://doi.org/10.1109/TGRS.2016.2569141>.
35. Li, Y.; Fu, K.; Sun, X.; Sun, H. An aircraft detection framework based on reinforcement learning and convolutional neural networks in remote sensing images. *Remote Sens.* **2018**, *10*, 243. <https://doi.org/10.3390/rs10020243>.
36. Zuo, J.; Xu, G.S.; Fu, K.; Sun, X.; Sun, H. Aircraft type recognition based on segmentation with deep convolutional neural networks. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 282–286. <https://doi.org/10.1109/LGRS.2017.2786232>.
37. Xiong, Y.; Niu, X.; Dou, Y.; Qie, H.; Wang, K. Non-locally Enhanced Feature Fusion Network for Aircraft Recognition in Remote Sensing Images. *Remote Sens.* **2020**, *12*, 681. <https://doi.org/10.3390/rs12040681>.
38. Jia, H.; Guo, Q.; Chen, J.; Wang, F.; Wang, H.; Xu, F. Adaptive component discrimination network for airplane detection in remote sensing images. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2021**, *14*, 7699–7713. <https://doi.org/10.1109/JSTARS.2021.3098296>.
39. Cai, B.; Jiang, Z.N.; Zhang, H.; Yao, Y.; Nie, S. Online exemplar-based fully convolutional network for aircraft detection in remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 1095–1099. <https://doi.org/10.1109/LGRS.2018.2829147>.
40. Wei, H.; Zhang, Y.; Wang, B.; Yang, Y.; Li, H.; Wang, H. X-LineNet: Detecting aircraft in remote sensing images by a pair of intersecting line segments. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 1645–1659. <https://doi.org/10.1109/TGRS.2020.2999082>.
41. Zeng, B.; Ming, D.; Ji, F.; Yu, J.; Xu, L.; Zhang, L.; Lian, X. Top-down aircraft detection in large-scale scenes based on multi-source data and FEF-R-CNN. *Int. J. Remote Sens.* **2022**, *43*, 1108–1130. <https://doi.org/10.1080/01431161.2022.2034194>.
42. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
43. Liu, C.; Yu, H.; Wei, H.; Sun, X.; Fu, K. S2CGNet: A robust aircraft detector based on the sword-shaped component geometry. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5602715. <https://doi.org/10.1109/TGRS.2023.3242147>.

44. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.Y.; et al. Segment anything. In *IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: Piscataway, NJ, USA, 2023; pp. 4015–4026.
45. Qian, Y.; Pu, X.; Jia, H.; Wang, H.; Xu, F. ARNet: Prior knowledge reasoning network for aircraft detection in remote-sensing images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5205214. <https://doi.org/10.1109/TGRS.2024.3359764>.
46. Tang, P.; Wang, X.; Bai, X.; Liu, W. Multiple Instance Detection Network with Online Instance Classifier Refinement. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2017; pp. 3051–3059.
47. Li, J.; Zhou, P.; Xiong, C.; Socher, R.; Hoi, S.C.H. Prototypical Contrastive Learning of Unsupervised Representations. In *Proceedings of the International Conference on Learning Representations (ICLR)*, Virtual, 3–7 May 2021.
48. Guo, M.; You, Y.; Liu, F. Discriminative Prototype Learning for Few-Shot Object Detection in Remote-Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5624313. <https://doi.org/10.1109/TGRS.2023.3326992>.
49. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum Contrast for Unsupervised Visual Representation Learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2020; pp. 9726–9735.
50. Wu, A.; Han, Y.; Zhu, L.; Yang, Y. Universal-Prototype Enhancing for Few-Shot Object Detection. In *IEEE International Conference on Computer Vision (ICCV)*; IEEE: Piscataway, NJ, USA, 2021; pp. 9745–9754.
51. Cheng, G.; Yan, B.; Shi, P.; Li, K.; Yao, X.; Guo, L.; Han, J. Prototype-CNN for Few-Shot Object Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5604610. <https://doi.org/10.1109/TGRS.2021.3078507>.
52. Ding, J.; Xue, N.; Long, Y.; Xia, G.S.; Lu, Q. Learning RoI Transformer for Oriented Object Detection in Aerial Images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2019.
53. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2016; pp. 770–778.
54. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60*, 84–90. <https://doi.org/10.1145/3065386>.
55. Yang, L.; Kang, B.; Huang, Z.; Zhao, Z.; Xu, X.; Feng, J.; Zhao, H. Depth Anything V2. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, Vancouver, BC, Canada, 10–15 December 2024.
56. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: Fully Convolutional One-Stage Object Detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: Piscataway, NJ, USA, 2019; pp. 9627–9636.

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.