

Article

Symmetry Frequency-Aware Fourier Series Network for Aerial Small Object Detection

Xinghai Hou ¹ , Donglin Jing ² , Fukun Bi ^{3,*}, Chenglong He ¹, Yong Huang ³ and Changjie Wang ⁴

¹ School of Electrical and Control Engineering, North China University of Technology, Beijing 100144, China; hou_xinghai@163.com (X.H.); 2025405150132@mail.ncut.edu.cn (C.H.)

² Shanghai Aerospace Control Technology Institute, Shanghai 201109, China; jingdonglin_bit@163.com

³ School of Artificial Intelligence and Computer Science, North China University of Technology, Beijing 100144, China; yhuang@bmsltcc.com

⁴ Beijing Institute of Technology Raeco Aerospace Information Technology Co., Ltd., Beijing 100081, China; wangchangjie2008@163.com

* Correspondence: bifukun@163.com

Abstract

Aerial tiny objects naturally possess conjugate symmetry in the frequency domain, yet complex scenarios, heavy clutter, and frequent rotation/occlusion lead to severe detail loss, inaccurate contour modeling, and high false/miss rates in existing detectors. Current Fourier-based methods neither leverage object symmetry nor coordinate with Fourier analysis to jointly enhance contour representation and spatial-frequency feature learning, suffering from weak fusion, phase-sensitive coefficient regression, and poor discriminability. To fill this gap, we propose the Frequency-Aware Fourier Series Detection Network (FAFSDet), which explicitly exploits the inherent symmetry of tiny objects and their frequency-domain representations. Specifically, FAFC (Frequency-Aware Feature Fusion) employs conjugate-symmetry-guided dynamic low-pass filtering, similarity-based rearrangement, and adaptive high-frequency enhancement to recover degraded symmetric patterns. FSPRM (Fourier Series Profile Representation) utilizes the symmetric positive-negative frequency distribution to achieve compact parametric contour encoding and normalized centroid-shape description. FSDIM (Fourier Series Detection Inference) incorporates symmetric multi-scale branches, a rolling-optimization loss that eliminates phase interference while preserving coefficient-regression symmetry, and inverse Fourier transform for precise contour reconstruction and end-to-end detection. Extensive experiments on DOTA, AI-TOD, and UCAS-AOD demonstrate that our method achieves superior performance (mAP 82.18%, 51.2%, and 90.70%, respectively) and strong generalization, particularly in scenarios where symmetry is most severely compromised, confirming that exploiting these symmetry properties substantially boosts detection accuracy.

Keywords: aerial target detection; frequency domain transform; convolutional neural networks; bounding box representation



Received: 26 June 2026

Revised: 21 July 2026

Accepted: 23 July 2026

Published: 27 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 \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

As a core task in the perception layer of computer vision, object detection has been deeply applied to key fields such as UAV remote sensing monitoring, autonomous driving environmental perception, intelligent security early warning, and industrial defect detection. In practical applications like remote sensing images and autonomous driving in complex scenarios, objects often exhibit characteristics such as irregular contours,

variable scales, and strong background clutter interference, imposing extremely high requirements on the contour modeling accuracy and feature robustness of detection algorithms. At present, mainstream object detection methods still rely on axis-aligned or rotated rectangular boxes for object representation [1]. This discrete box representation fails to capture the continuous contour details of objects, resulting in limited detection accuracy for objects with complex shapes [2,3]. Meanwhile, facing challenges such as domain shift and low-quality imaging, models lack the ability of collaborative learning for object contours and semantic features.

Current deep learning-based object detection methods have formed two major technical routes centered on Convolutional Neural Networks (CNNs) and Transformers, with continuous iterations around feature extraction, multi-scale fusion, and box regression optimization. In terms of lightweight architecture optimization for real-time detection, RF-DETR [4,5] adopts a dynamic anchor generation strategy and a lightweight attention mechanism, increasing the AP value on the COCO dataset to over 60 while maintaining real-time inference speed. Hyper-YOLO [6] integrates hypergraph computation into the YOLO framework to construct high-order feature correlations across positions and scales, effectively improving the detection performance for dense and small objects. In cross-modal feature fusion, WaveMamba [7] combines discrete wavelet transform with a state space model to realize frequency-domain decoupling and efficient sequence modeling of RGB-infrared modalities, enhancing the model's adaptability to harsh illumination environments. For unsupervised domain adaptation, FRANCK [8] designs a dedicated domain adaptation module for the DETR architecture, achieving efficient transfer from the source domain to the target domain without source code data and improving the generalization ability of the model. In dynamic feature enhancement, ESCNet [9] proposes an edge-semantic collaborative dynamic network, which suppresses background texture interference through hierarchical modulation and accurately locates the blurred boundaries of camouflaged objects.

With the increasing demand for object representation accuracy, frequency-domain analysis, owing to its inherent advantages in shape encoding and feature decoupling, has been gradually introduced into the field of object detection and become a key bridge connecting spatial-domain feature learning and accurate shape modeling. Research in this subfield mainly focuses on shape representation based on Fourier transform/series, as well as frequency-domain feature decomposition and fusion. In the design of frequency-domain dynamic convolution, FDConv [10] decomposes and dynamically modulates convolution weights in the Fourier domain, enhancing the adaptability of convolution to multi-frequency features without increasing parameters and improving the accuracy of dense prediction tasks. For frequency-domain decoupling in UAV object detection, FDD-UAV [11] proposes a frequency-domain disentanglement method to separate domain-invariant and domain-specific spectral features, significantly improving the generalization ability of UAV object detection models in unseen scenarios.

From the phenomenological perspective, existing Fourier series-based detection methods suffer from two core problems in complex scenarios (as shown in Figure 1). First, there is an imbalance between the accuracy and robustness of contour representation: when objects are occluded, have incomplete contours, or are subject to strong background clutter interference, the regression of Fourier coefficients is prone to deviations, leading to a mismatch between the reconstructed contours and the actual objects. This issue is particularly prominent in the detection of small objects and irregular remote sensing objects. Second, the fusion of frequency-domain and spatial-domain features is inadequate: most models process the contour features in the Fourier domain and the semantic features in the spatial domain separately, with a lack of in-depth collaboration between the two. As a

result, contour information cannot effectively guide the focusing of spatial features, and semantic features are also unable to constrain the regression direction of contour coefficients, ultimately compromising detection accuracy.

From a technical perspective, the aforementioned phenomena stem from two core technical bottlenecks. First, the matching mechanism of traditional Fourier loss functions is flawed: conventional loss functions fail to account for the periodicity and phase sensitivity of Fourier series. Direct coefficient matching tends to trap the model in local optima and prevents accurate contour alignment. Even with the introduction of basic matching strategies, it remains difficult to address the challenges posed by contour deformation and occlusion. Second, there is a lack of a collaborative learning architecture for spatial and frequency features: existing methods mostly adopt a serial structure of “spatial feature extraction + Fourier coefficient regression”, without dedicated interaction modules to enable bidirectional guidance between the two types of features. This structure neither allows the use of spatial semantic features to filter valid contour information nor enables contour features to enhance the object-focused capability of spatial features, resulting in insufficient discriminative power of feature representations.

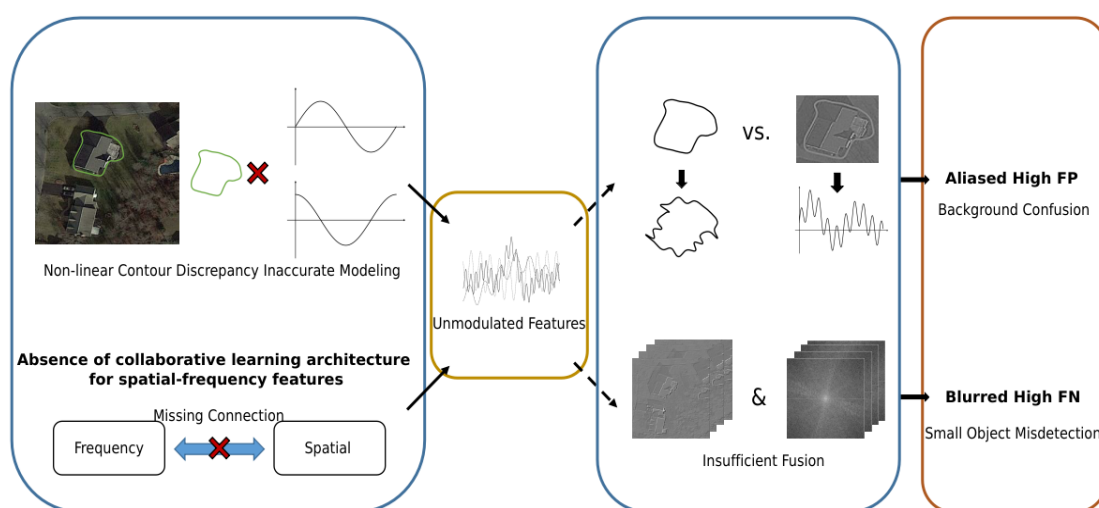


Figure 1. Problem analysis diagram.

To address the aforementioned issues, we propose the Frequency-Aware Fourier Series Detection Network (FAFSDet), which consists of three core components: the Frequency-Aware Feature Fusion Module (FAFC), the Fourier Series Profile Representation Module for small targets (FSPRM), and the Fourier Series Detection Inference Module for small targets (FSDIM). Leveraging the inherent symmetry and periodicity of Fourier series—a key focus of the journal *Symmetry*—the FAFC module adopts a three-tier collaborative strategy of high-level feature smoothing, inconsistent feature correction, and low-level boundary enhancement to output high-quality fused features with both strong category discriminability and precise spatial boundaries, while maintaining the symmetric consistency of intra-category features. The FSPRM module follows the logic of profile periodic modeling, accurate Fourier coefficient solving (guided by symmetry principles), and coefficient dimension regularization optimization to realize parametric compression and unified dimension normalized representation of small target profiles, fully exploiting the symmetry-adapted characteristics of Fourier series in profile description. The FSDIM module is implemented in a progressive manner of multi-scale branch construction, rolling optimization loss design, and coefficient inverse transform decoding to complete the existence judgment, category recognition, and high-precision profile detection of small targets, eliminating asymmetric

phase interference in profile annotation. The contributions of this paper are summarized as follows:

- (1) We propose the FAFC module. Through the three-tier collaboration of dynamic low-pass filtering, similarity-guided feature rearrangement, and adaptive boundary high-frequency enhancement, it addresses the core problems of intra-category feature asymmetric inconsistency and small target boundary offset in traditional feature fusion, and improves the discriminability and boundary representation accuracy of small target features.
- (2) We design the FSPRM module. Relying on symmetry-adapted Fourier series—which simplifies coefficient computation by leveraging even-odd symmetry properties—it realizes parametric compression encoding and dimension normalized representation of closed profiles of small targets.
- (3) We propose the FSDIM module for small targets. It constructs parallel detection branches adapted to the symmetric multi-scale distribution of small targets, designs a rolling optimization loss to eliminate the phase interference caused by the unregulated starting points of annotated profiles, and improves the model training stability and Fourier coefficient regression accuracy.

2. Related Work

2.1. Methods for Optimizing Small Object Representation Based on Feature Enhancement

Small objects suffer from low pixel proportion and insufficient semantic information, making them prone to losing key details during feature extraction, which has become a core bottleneck restricting detection accuracy. Existing studies generally start with feature-level enhancement and improve the representation capability of small objects through multi-scale fusion and detail compensation strategies. Some scholars have constructed cross-layer feature interaction modules to break the hierarchical barriers of traditional pyramid structures, and perform adaptive weighted fusion of high-level semantic features and low-level texture features, enabling the network to retain the edge and contour information of small objects even in deep inference [12]. Other studies focus on the refined regulation of receptive fields: by dynamically adjusting the dilation rate of convolution kernels, they match an adaptive receptive field range for small objects, avoiding the problems of introducing redundant backgrounds due to an overly large receptive field and losing contextual information due to an overly small one [13]. In addition, to address the low signal-to-noise ratio of small object features, relevant work has proposed feature purification schemes based on attention mechanisms. Using the joint constraints of spatial attention and channel attention, these schemes automatically focus on the regions where small objects are located and suppress the interference of background noise on feature learning [14]. All these methods take strengthening the feature expression of small objects themselves as the core, and achieve initial improvement in detection performance by making up for information loss. In spatial-frequency collaborative detection of raw images, SFAE [15,16] constructs a spatial-frequency perception enhancer to separate object contour and texture features in the frequency domain, effectively recovering the suppressed detailed information in RAW images and improving detection accuracy. In multi-modal frequency-domain fusion, FREE-Merging [16] eliminates task interference in the frequency domain of model parameters through Fourier filtering, realizing efficient fusion of vision-language multi-modal models, reducing storage costs and improving cross-task performance.

2.2. Sample Distribution Balancing Strategies Based on Data-Driven

The severe class imbalance problem exists in small object detection tasks, i.e., the number of small object samples is much smaller than that of large objects, which leads to

gradient bias during model training and ultimately reduces the recall rate of small objects. To alleviate this problem, existing studies have proposed a series of sample distribution optimization strategies from the data level. In terms of data augmentation, which is different from traditional general methods such as random cropping and flipping, targeted augmentation schemes have become a research hotspot: some work adopts small object region oversampling technology, which focuses on intercepting and copying local regions containing small objects in images to directly increase the training frequency of small object samples [17]; other studies have proposed generative augmentation methods, which based on GAN networks or diffusion models, learn the appearance distribution rules of small objects using existing samples and automatically generate diverse virtual small object samples to enrich the sample diversity of the training set [18]. In the optimization of training strategies, the design of weighted loss functions has been widely applied. Researchers dynamically calculate sample weights and assign higher loss weights to small objects to offset the dominant position of large object samples in gradient updates [19]. Meanwhile, the hard sample mining mechanism has also been introduced into the field of small object detection: by screening small object samples that are prone to misjudgment during training and conducting focused iterative training, the model's ability to recognize small objects in complex scenarios is improved.

2.3. Structural Adaptation Improvement Schemes Based on Detection Frameworks

Traditional object detection frameworks are mostly designed to take multi-scale objects into account, and have the problem of insufficient structural adaptability in small object detection tasks. Therefore, customized improvements for network frameworks have become an important research direction of relevant work. In one-stage detection algorithms, research focuses on the optimization of anchor box strategies: existing work abandons the traditional fixed anchor box size design and proposes a dynamic anchor box generation method based on the size statistics of small objects. According to the actual size distribution of small objects in the dataset, it adaptively generates the aspect ratio and scale range of anchor boxes to improve the matching degree between anchor boxes and small objects [20]. Some scholars have also proposed an anchor-free small object detection idea, which transforms the detection task into a key point localization and size regression problem, avoiding the restriction of improper anchor box presets on small object detection [21]. In two-stage detection algorithms, the improvement of the Region Proposal Network (RPN) has become the core: relevant studies improve the regional search accuracy for small objects by reducing the sliding window stride of the RPN and optimize the screening mechanism of proposal boxes to prevent candidate boxes generated by small objects from being eliminated too early due to low confidence [22]. In addition, the improvement of end-to-end lightweight frameworks has gradually attracted attention. By streamlining the network structure and optimizing the feature transmission path, researchers improve the inference speed of the model on the premise of ensuring small object detection accuracy, meeting the needs of practical application scenarios such as remote sensing images and real-time monitoring [23].

However, these approaches typically treat contour fitting as an offline post processing step or a separate auxiliary task, rather than integrating it seamlessly into the end to end training pipeline. More critically, they overlook the phase sensitivity caused by arbitrary starting points of contour annotation, which leads to coefficient misalignment and unstable training. In contrast, as shown in Table 1 our FAFSDet uniquely combines frequency-aware feature fusion with Fourier series contour regression in a unified framework. The FAFC module restores symmetric feature distributions through its three tier collaborative strategy; the FSPRM module enables differentiable coefficient regression with centroid shape de-

coupling; and the FSDIM module introduces a rolling optimization loss to eliminate phase ambiguity. This joint design realizes deep collaborative learning of spatial and frequency features under explicit symmetry guidance, which effectively alleviates the core limitations of existing Fourier-based and contour-based detection methods.

Table 1. Comparison between the proposed FAFSDet and representative frequency-domain detection methods.

Method	Core Technical Paradigm	Spatial-Frequency Fusion	Contour Representation	Loss Design	Symmetry Utilization
FDConv	Freq-domain dynamic convolution modulation	Unidirectional kernel weight modulation	Not applicable	Standard detection loss	Not explicitly utilized
FDD-UAV	Frequency-domain disentanglement adaptation	Spectral feature decoupling	Not applicable	Adversarial + detection loss	Not explicitly utilized
SFAE	Spatial-frequency perception enhancement	Parallel feature decomposition	Not applicable	Standard detection loss	Implicit partial adoption
Conventional Fourier Descriptors	Independent contour shape description	Not involved	Unregularized variable-dimension coefficients	Phase-sensitive direct matching	Redundant symmetry retained
Ours (FAFSDet)	End-to-end spatial-frequency collaborative detection	Three-tier symmetry-guided bidirectional fusion	Unified-dimension Fourier series with centroid-shape decoupling	Phase-robust rolling optimization loss	Full exploitation of conjugate and periodic symmetry

3. Methodology

Frequency-Aware Fourier Series Network.

The overall framework for small object detection proposed in this paper is centered on the progressive implementation of frequency-aware feature fusion and Fourier contour drawing (as shown in Figure 2). Through the collaboration of three core modules, it achieves a dual improvement in the detection accuracy and contour representation capability of small objects. First, relying on the Frequency-Aware Fusion Core (FAFC) for small objects, a three-tier collaboration strategy of high-level feature smoothing, inconsistent feature correction, and low-level boundary enhancement is adopted to dynamically suppress high-frequency perturbations in high-level features, rectify the inconsistency of intra-class features, and enhance high-frequency details of low-level boundaries, thus outputting high-quality fused features with strong category discriminability and accurate spatial boundaries. Second, via the Fourier Series Profile Representation Module (FSPRM) for small object contours, following the logic of contour periodic modeling, accurate Fourier coefficient solving, coefficient dimension regularization and optimization, the closed contours of small objects are transformed into one-dimensional periodic signals, with coefficient solving and dimension regularization completed to realize parametric compression and unified dimension normalized representation of small object contours. Finally, the Fourier Series Detection Inference Module (FSDIM) for small objects is employed, which is implemented

progressively through multi-scale branch construction and coefficient inverse transform decoding. It builds multi-scale parallel prediction branches, eliminates the interference of annotated starting points with rolling matching loss, and reconstructs precise contours via coefficient inverse transform, ultimately achieving existence judgment, category recognition, and high-precision contour detection of small objects.

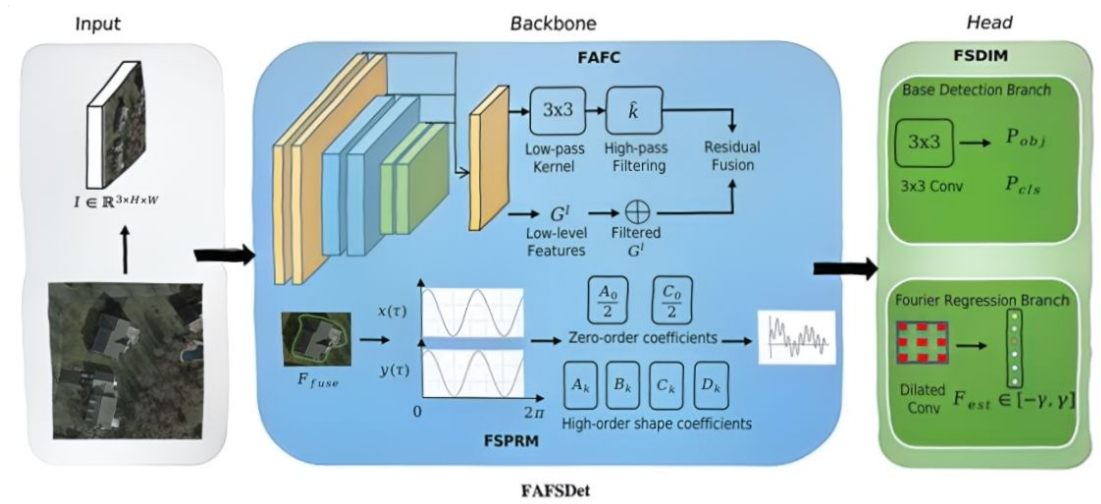


Figure 2. Overall architecture of the FAFSDet Model.

The symmetry utilized in this network mainly includes two types of mathematical properties:

- **Conjugate symmetry of Fourier transform:** For a real-valued spatial signal $f(x)$, its Fourier transform satisfies $F(u) = F^*(-u)$, meaning positive and negative frequency components are conjugate to each other. In the FAFC module, we take this as a prior constraint: only half of the frequency components need to be retained to complete full frequency-domain filtering and feature reconstruction, which reduces computational cost while ensuring the symmetric integrity of frequency-domain feature distribution and avoids feature distortion caused by asymmetric filtering.
- **Symmetric distribution of Fourier series:** After parameterization of closed contours, the positive and negative order coefficients of their Fourier series have symmetric correspondence, and the coefficient sequence has periodic translation invariance. In the FSPRM module, we use this property to eliminate redundant coefficients and achieve dimensional compression; in the FSDIM module, the rolling optimization loss maintains the periodic symmetry of coefficient distribution through phase shift, eliminating phase asymmetry interference caused by arbitrary annotation starting points.

3.1. Frequency-Aware Fusion Core (FAFC) for Small Objects

The Frequency-Aware Fusion Core (FAFC) is a key component for feature fusion in dense image prediction tasks, with its core mission to address two major challenges in traditional feature fusion: intra-class feature inconsistency and large relative offset of small object boundaries. Its technical logic is carried out around the three-tier collaboration of high-level feature smoothing, inconsistent feature correction, and low-level boundary enhancement (as shown in Algorithm 1): First, adaptive smoothing is performed on high-level coarse-grained features with dynamic low-pass filtering to suppress high-frequency perturbations inside the features and reduce intra-class differences from the source. Second, guided by local feature similarity, precise pixel sampling is achieved through feature rearrangement, where low-similarity features are replaced with high category-consistent features, and large-area inconsistent regions and thin boundaries are optimized simul-

taneously. Finally, adaptive boundary high-frequency enhancement is used to mine the super-Nyquist frequency components lost in low-level features due to downsampling, strengthen the high-frequency details of target boundaries, and ultimately output high-quality fused features with strong category discriminability and accurate spatial boundaries, providing support for tasks such as semantic segmentation and object detection.

Algorithm 1 Fourier Coefficient Extraction for Label Generation (FSPRM)

Require:

- Annotated closed contour point set $\mathcal{P} = \{(X_i, Y_i)\}_{i=0}^{M-1}$ for one small object
- Fourier series fitting order n (with $n \ll M$)
- Number of discrete samples M

Ensure:

- Regularised Fourier coefficient tensor $F_{feat} \in \mathbb{R}^{4n-2}$, ready for network regression

```

1: Step 1: Contour Periodic Modelling
2: Perform equidistant re-sampling on  $\mathcal{P}$  to obtain  $M$  ordered points:
3:    $\tau_i \leftarrow \frac{2\pi i}{M}, \quad i = 0, 1, \dots, M-1$ 
4: Define two independent 1D periodic functions  $X(\tau)$  and  $Y(\tau)$  over  $\tau \in [0, 2\pi]$ 
5: Step 2: Fourier Coefficient Solving
6: Compute zero-order centroid coefficients:
7:    $\frac{A_0}{2} \leftarrow \frac{1}{M} \sum_{i=0}^{M-1} X(\tau_i)$ 
8:    $\frac{C_0}{2} \leftarrow \frac{1}{M} \sum_{i=0}^{M-1} Y(\tau_i)$ 
9: for  $k = 1$  to  $n-1$  do
10:   Compute high-order shape coefficients via discrete integral transform:
11:    $A_k \leftarrow \frac{1}{\pi} \sum_{i=0}^{M-1} X(\tau_i) \cos(k\tau_i) \cdot \Delta\tau$ 
12:    $B_k \leftarrow \frac{1}{\pi} \sum_{i=0}^{M-1} X(\tau_i) \sin(k\tau_i) \cdot \Delta\tau$ 
13:    $C_k \leftarrow \frac{1}{\pi} \sum_{i=0}^{M-1} Y(\tau_i) \cos(k\tau_i) \cdot \Delta\tau$ 
14:    $D_k \leftarrow \frac{1}{\pi} \sum_{i=0}^{M-1} Y(\tau_i) \sin(k\tau_i) \cdot \Delta\tau$ 
15: end for
16: Step 3: Coefficient Dimension Regularization
17: Eliminate redundant zero-order terms  $B_0$  and  $D_0$  (which are identically zero)
18: Assemble all valid coefficients into a unified tensor:
19:    $F_{feat} \leftarrow [A_0, C_0, A_1, B_1, C_1, D_1, \dots, A_{n-1}, B_{n-1}, C_{n-1}, D_{n-1}]$ 
20: return  $F_{feat}$ 

```

For the FAFC module, it follows the basic ideas of dynamic filtering and feature rearrangement, consistent with traditional feature pyramid fusion in terms of improving the feature representation quality of small objects. The core innovations lie in three aspects: first, we introduce the conjugate symmetry of Fourier transform as a prior constraint, and maintain the symmetric integrity of frequency-domain features during filtering, avoiding feature distortion caused by traditional asymmetric filtering, which is a valuable prior that has not been fully explored and utilized in existing dynamic filtering methods; second, we adopt a three-layer collaborative strategy of high-level smoothing, middle-level correction, and low-level enhancement, instead of traditional single-scale independent filtering, which takes both intra-class feature consistency and boundary detail accuracy into account, and solves the contradiction that traditional filtering is difficult to optimize large-area

features and fine boundaries simultaneously; third, we specially optimize the filter kernel size and offset groups for the scale characteristics of aerial small objects, adopting a 3×3 filter kernel and 4 groups of offsets by default, so as to avoid excessive background noise introduced by large filter kernels and be more suitable for the feature scale of small objects.

Dynamic Low-Pass Smoothing Filter Module (DLPF): The core function of the Dynamic Low-Pass Smoothing Filter Module (DLPF) is to construct a spatially variant low-pass filter for smoothing and upsampling high-level coarse features, suppressing high-frequency fluctuations inside the features, and thereby alleviating the inconsistency of intra-class features. This module takes the compressed feature U^l after initial fusion as the input data source, generates valid low-pass filtering weights through convolution operation and kernel-level softmax normalization, and then completes scale enhancement and smoothing optimization of high-level features by combining pixel rearrangement and sub-pixel up-sampling.

The calculation process of its filtering weights is as follows:

$$W^l = \text{Conv}_{3 \times 3}(U^l)$$

$$\hat{K}_{i,j}^{l,p,q} = \text{Softmax}(\bar{W}_{i,j}^l) = \frac{\exp(\bar{W}_{i,j}^{l,p,q})}{\sum_{p,q \in \Theta} \exp(\bar{W}_{i,j}^{l,p,q})}$$

where U^l denotes the compressed feature after initial fusion, $\bar{W}^l \in \mathbb{R}^{\bar{M}^2 \times 2H \times 2W}$ is the intermediate calculation result of the filtering weight, and $\bar{M}$ corresponds to the kernel size of the low-pass filter. Θ is defined as the kernel region range of $\bar{M} \times \bar{M}$. After kernel-level Softmax normalization, $\hat{K} \in \mathbb{R}^{\bar{M}^2 \times 2H \times 2W}$ becomes the final spatially variant low-pass filtering weight. This process ensures that the filtering weights are non-negative and sum to 1, satisfying the basic characteristics of low-pass filtering.

Based on the generated filtering weights, pixel rearrangement and filtering upsampling operations are performed on the high-level feature $F^{l+1} \in \mathbb{R}^{C \times H \times W}$, with the specific process as follows:

$$\tilde{F}_{i,j}^{l+1,k} = \sum_{p,q \in \Theta} \bar{K}_{i,j}^{l,k,p,q} \cdot F_{i+p,j+q}^{l+1}$$

$$\tilde{F}^{l+1} = \text{PixelShuffle}(\tilde{F}^{l+1,1}, \tilde{F}^{l+1,2}, \tilde{F}^{l+1,3}, \tilde{F}^{l+1,4})$$

where $k \in \{1, 2, 3, 4\}$ corresponds to the channel grouping identifier after pixel rearrangement. First, 4 groups of independent low-pass filtering are performed on the high-level features to obtain $\tilde{F}^{l+1,k} \in \mathbb{R}^{C \times H \times W}$. Then, the PixelShuffle sub-pixel upsampling technique is used to reorganize the filtered features into a smoothed high-level feature $\tilde{F}^{l+1} \in \mathbb{R}^{C \times 2H \times 2W}$ with a $2 \times$ scale increase. This operation not only completes the scale enhancement but also effectively suppresses high-frequency perturbations inside the features, significantly improving the similarity of intra-class features.

Similarity-Guided Feature Rearrangement Module (SFRM). The design intention of the Similarity-Guided Feature Rearrangement Module (SFRM) is to solve the contradiction that dynamic low-pass smoothing filtering is difficult to correct both large-area inconsistent features and thin boundaries simultaneously. Guided by local feature similarity, this module predicts rearrangement offsets to realize adaptive replacement of feature pixels, covering low-similarity features with high intra-class consistency features, and optimizing the feature quality of large-area inconsistent regions and thin boundaries at the same time. Its core process is to first calculate the local similarity of the initially fused features, then

predict grouped offsets by combining features and similarity maps, and finally complete the offset rearrangement of features.

First, the local cosine similarity of the initially fused compressed features is calculated to capture the feature correlation between each pixel and its 8 neighboring pixels:

$$T_{i,j}^{l,p,q} = \frac{\sum_{c=1}^C U_{c,i,j}^l \cdot U_{c,i+p,j+q}^l}{\sqrt{\sum_{c=1}^C (U_{c,i,j}^l)^2} \sqrt{\sum_{c=1}^C (U_{c,i+p,j+q}^l)^2}}$$

where $T \in \mathbb{R}^{8 \times H \times W}$ is the local cosine similarity map, which clearly shows the feature similarity between each pixel and its neighboring pixels, and provides directional guidance for high-similarity regions for subsequent offset prediction. Taking the initially fused compressed feature U^l and the local similarity map T^l as joint inputs, two independent 3×3 convolution layers are used to predict the offset direction and offset scale respectively, and the grouped rearrangement offset is finally synthesized:

$$P^l = Q^l \cdot R^l$$

$$Q^l = \text{Conv}_{3 \times 3}(\text{Concat}(U^l, T^l))$$

$$R^l = \text{Sigmoid}(\text{Conv}_{3 \times 3}(\text{Concat}(U^l, T^l)))$$

where $Q^l \in \mathbb{R}^{2N \times H \times W}$ represents the offset direction feature, and $R^l \in \mathbb{R}^{2N \times H \times W}$ is the offset scale feature after Sigmoid activation, which is used to adjust the magnitude of the offset. N is the number of offset groups. The grouped offset strategy achieves fine-grained rearrangement of features, promotes the offset of low intra-class similarity pixels to high-similarity regions, completes feature replacement, and solves the problem of feature optimization for large-area inconsistent regions and thin boundaries at the same time.

Adaptive Boundary High-Frequency Enhancement Module (ABHE): The core goal of the Adaptive Boundary High-Frequency Enhancement Module (ABHE) is to recover the super-Nyquist frequency boundary details lost in the downsampling process of low-level high-resolution features. By generating a spatially variant high-pass filter, this module performs adaptive high-frequency enhancement on low-level features, sharpens target boundaries, and effectively solves the boundary offset problem in feature fusion. Its core process is to take the initially fused compressed feature as input, first generate a spatially variant low-pass kernel, then obtain a high-pass kernel through filter inversion, and finally perform high-pass filtering and residual fusion on the low-level features to realize the enhancement of boundary high-frequency information.

The calculation process of the spatially variant high-pass filtering weight is as follows:

$$\hat{W}^l = \text{Conv}_{3 \times 3}(U^l)$$

$$\hat{K}_{i,j}^{l,p,q} = I - \text{Softmax}(\hat{W}_{i,j}^l) = I^{p,q} - \frac{\exp(\hat{W}_{i,j}^{l,p,q})}{\sum_{p,q \in \Theta} \exp(\hat{W}_{i,j}^{l,p,q})}$$

where $\hat{W}^l \in \mathbb{R}^{\hat{M}^2 \times H \times W}$ is the intermediate feature of the high-pass filtering weight, $\hat{M}$ is the kernel size of the high-pass filter, and I is the unit kernel (when $\hat{M} = 3$, the weight matrix of I is $[[0, 0, 0], [0, 1, 0], [0, 0, 0]]$). Through the filter inversion operation, the low-pass kernel after kernel-level Softmax normalization is converted into a valid spatially variant high-pass kernel $\hat{K}$, realizing accurate extraction of high-frequency information of low-level features.

Based on the generated spatially variant high-pass kernel, adaptive high-pass filtering is performed on the low-level feature G^l , and the residual fusion method is adopted to retain the original feature information, obtaining the low-level feature after high-frequency enhancement:

$$\tilde{G}_{i,j}^l = G_{i,j}^l + \sum_{p,q \in \Theta} \hat{K}_{i,j}^{l,p,q} \cdot G_{i,j}^l$$

where $\tilde{G}^l$ is the low-level feature after high-frequency enhancement. The residual fusion strategy not only retains the original information of low-level features but also strengthens the boundary high-frequency details of super-Nyquist frequency, significantly improving the boundary sharpness of features, effectively solving the boundary offset problem in the feature fusion process, and providing accurate spatial boundary support for subsequent dense image prediction.

Dynamic Low-Pass Filtering (DLPF), Similarity-Guided Feature Rearrangement (SFRM) and Adaptive Boundary High-Frequency Enhancement (ABHE) all adopt 3×3 basic convolution kernels, and the generated filter kernel has a receptive field of 3×3 . The module inputs three layers of features C3, C4, and C5 output by the backbone network, with channel numbers of 256, 512, and 1024 respectively; the output fused features are unified to 256 channels. Resolution levels: The module processes in parallel on three scale features, corresponding to $1/8$, $1/16$, and $1/32$ downsampling resolutions of the input image; high-level features are upsampled by $2\times$ and then fused with middle-level features, and finally three layers of features are output for the detection head. A three-level fusion strategy of “high-level smoothing upsampling $\rightarrow$ middle-level feature correction $\rightarrow$ low-level boundary enhancement $\rightarrow$ element-wise weighted fusion” is adopted, and each layer adaptively adjusts the fusion ratio through learnable weights. The number of offset groups $N = 4$, and 8-neighborhood cosine similarity is adopted for similarity calculation.

3.2. Fourier Series Profile Representation Module (FSPRM) for Small Objects

For the FSPRM module, it is based on the closed contour representation principle of Fourier series, and shares the same mathematical basis with classical Fourier descriptors, both of which describe the shape information of objects through frequency-domain coefficients. The core differences and innovations are reflected in three points: first, we propose a low-order coding strategy of centroid-shape separation, and optimize the default order to 4 for the characteristics of simple contours and few pixels of small objects. While ensuring the core contour representation accuracy, it greatly reduces the regression dimension and significantly alleviates the overfitting risk of tiny objects, which is a customized design for aerial scenes; second, we complete the regularization of coefficient dimensions, and output coefficient tensors with unified dimensions, which can be directly embedded into the regression branch of mainstream one-stage detection frameworks, which effectively alleviates the engineering difficulty that traditional Fourier representations have variable dimensions and are hard to integrate into end-to-end training pipelines; third, we eliminate redundant coefficients by utilizing the symmetric distribution characteristics of positive and negative frequencies of Fourier series, which further compresses the computational overhead and meets the demand for lightweight small object detection.

This module is used to realize parametric encoding representation of closed contours of small objects. By virtue of the periodic fitting characteristic of Fourier series, the discrete point set of small object contours is transformed into a continuous coefficient tensor, which not only retains the core information of the centroid position and contour shape of small objects, but also realizes unified-dimension description of small object contours of different scales, accurately adapting to the feature representation requirements of small object detection. Combined with the geometric features of small object contours and relying on

the approximation characteristic of Fourier series, the design of this module must follow the following constraints:

$$n \ll M, D_f = 4n - 2$$

where n is the fitting order of the Fourier series, M is the number of discrete sampling points of the contour, and D_f is the dimension of the regularized Fourier coefficients. To ensure the accuracy of contour representation while reducing the computational complexity of the model, the FSPRM module adopts a coefficient representation strategy of centroid-shape separation, which is implemented in three core steps:

Contour Periodic Modeling: The core task of contour periodic modeling is to transform the 2D closed contour of a small object into a periodic signal that meets the fitting requirements of Fourier series. For the closed contour of any small object, a periodic parameter $\tau \in [0, 2\pi]$ is defined. When the parameter τ traverses the entire interval once, the contour coordinate point $(X(\tau), Y(\tau))$ forms a complete closed curve, and its periodic expression can be decomposed into two independent 1D periodic functions with the specific form

$$\begin{cases} X(\tau) = \frac{A_0}{2} + \sum_{k=1}^{n-1} [A_k \cos(k\tau) + B_k \sin(k\tau)] \\ Y(\tau) = \frac{C_0}{2} + \sum_{k=1}^{n-1} [C_k \cos(k\tau) + D_k \sin(k\tau)] \end{cases}$$

where $\frac{A_0}{2}$ and $\frac{C_0}{2}$ are the abscissa and ordinate of the centroid of the small object, respectively, and A_k, B_k, C_k, D_k are the k -order shape coefficients. The order k can be flexibly selected according to the complexity of the small object contour: low orders can characterize the overall contour shape of the small object, and high orders are used to depict the fine edge details of the small object.

Coefficient Dimension Regularization: The core task of coefficient dimension regularization is to generate a unified-dimension coefficient tensor adapted to the regression process of the detection network. Combined with the anchor box design requirements and scale adaptation characteristics of the detection model, the meaningless redundant coefficients B_0 and D_0 in the zero order are eliminated, and the zero-order centroid coefficients A_0, C_0 are integrated with $(n - 1)$ -order shape coefficients to form a Fourier coefficient set F_{feat} with a dimension of $4n - 2$, with the specific form

$$F_{feat} = [A_0, C_0, A_1, B_1, C_1, D_1, \dots, A_{n-1}, B_{n-1}, C_{n-1}, D_{n-1}]$$

This coefficient set realizes a unified tensor description of small object contours with different scales and shapes, and can be directly used as the regression target of the detection network, providing a standardized representation input for the subsequent inference stage of small object detection.

3.3. Fourier Series Detection Inference Module (FSDIM) for Small Objects

For the FSDIM module, the rolling optimization loss shares the same original intention with phase-invariant features and cyclic contour matching methods, which is to solve the alignment deviation caused by unfixed contour starting points. The core differences and innovations are as follows (shown in Algorithm 2): first, we realize phase shift through trigonometric function transformation directly in the Fourier coefficient space, without cyclic traversal matching on spatial point sets. The computational complexity is reduced from $O(M)$ (M is the number of contour sampling points) to $O(n)$ (n is the Fourier order), which significantly reduces computational complexity and improves inference efficiency; second, the loss function is completely differentiable and can be directly embedded in end-to-end training with smooth and stable gradient transmission, without gradient dis-

appearance or mutation caused by discrete point matching; third, we design a multi-scale parallel dual-branch structure of detection classification and coefficient regression to adapt to the symmetric multi-scale distribution characteristics of aerial small objects, realizing collaborative optimization of object detection and contour regression tasks, instead of treating contour representation as an auxiliary post-processing step of detection.

Algorithm 2 Contour Reconstruction and Detection Inference (FSDIM)

Require:

- Predicted centroid offsets $(\Delta A_0, \Delta C_0)$ from the detection head
- Predicted Fourier coefficients $F_{est} = \{A_k, B_k, C_k, D_k\}_{k=1}^{n-1}$
- Feature grid coordinate (C_{grid}, R_{grid}) and feature map size (W, H)
- Original image size (W_{img}, H_{img})
- Confidence threshold θ for detection filtering

Ensure:

- Reconstructed closed contour point set $\mathcal{P}_{rec}$ for the detected object

```

1: Step 1: Centroid Coordinate Mapping
2: Compute absolute centroid coordinates in the original image:
3:    $X_{center} \leftarrow \frac{W_{img}}{W} \cdot (C_{grid} + \Delta A_0)$ 
4:    $Y_{center} \leftarrow \frac{H_{img}}{H} \cdot (R_{grid} + \Delta C_0)$ 
5: Step 2: Coefficient Inverse Transform Decoding
6: Define sampling parameter  $\tau \in [0, 2\pi]$  with  $M$  equidistant steps
7: for each  $\tau_j \leftarrow \frac{2\pi j}{M}$  where  $j = 0, 1, \dots, M-1$  do
8:   Compute reconstructed contour coordinates:
9:    $X_{rec}(\tau_j) \leftarrow X_{center} + \sum_{k=1}^{n-1} [A_k \cos(k\tau_j) + B_k \sin(k\tau_j)]$ 
10:   $Y_{rec}(\tau_j) \leftarrow Y_{center} + \sum_{k=1}^{n-1} [C_k \cos(k\tau_j) + D_k \sin(k\tau_j)]$ 
11: end for
12: Step 3: Detection Filtering and Output Generation
13: Retrieve confidence score  $s$  from the parallel classification branch
14: if  $s \geq \theta$  then
15:   Assemble the reconstructed point set:
16:    $\mathcal{P}_{rec} \leftarrow \{(X_{rec}(\tau_j), Y_{rec}(\tau_j))\}_{j=0}^{M-1}$ 
17:   Output detection result with category label, centroid, and contour  $\mathcal{P}_{rec}$ 
18: else
19:   Suppress this prediction (low confidence)
20: end if
21: return  $\mathcal{P}_{rec}$  (if valid)
  
```

This module is used to realize end-to-end detection and contour reconstruction of small objects. It relies on multi-scale feature extraction to complete the simultaneous prediction of small object detection confidence, category and Fourier coefficients, eliminates the interference of annotated starting points through rolling optimization loss, and finally restores the accurate contour of small objects by virtue of coefficient inverse transform, while meeting the dual requirements of localization and feature representation in small object detection tasks. Combined with the characteristics of the one-stage detection architecture and following the regression law of Fourier coefficients, the design of this module must follow the following constraints:

$$\Delta A_0, \Delta C_0 \in [0, 1], F_{est} \in [-\gamma, \gamma]$$

where ΔA_0 and ΔC_0 are the relative offsets of the small object centroid, γ is the amplitude constraint threshold of Fourier coefficients, and F_{est} is the Fourier coefficient predicted by the model.

To realize accurate detection and contour reconstruction of small objects, the FS-DIM module adopts a collaborative implementation strategy of multi-scale prediction $\rightarrow$ optimized loss training $\rightarrow$ inverse transform decoding, which is implemented in three core steps:

Multi-Scale Branch Construction: The core task of multi-scale branch construction is to adapt to the multi-scale distribution characteristics of small objects and realize the simultaneous prediction of small object existence confidence, category probability and Fourier coefficients. Based on the fully convolutional backbone network, feature maps of three scales (high, medium and low) are extracted, and a parallel basic detection branch and an extended Fourier regression branch are built for each layer of feature map, finally forming a six-branch output structure with the specific form:

$$\begin{cases} P_{obj} = \text{Conv}(F_{scale}), P_{cls} = \text{Conv}(F_{scale}) \\ P_{Fs} = \text{Conv}_{expand}(F_{scale}) \end{cases}$$

where F_{scale} is the feature map of the s -th scale ($s = 1, 2, 3$), P_{obj} is the confidence tensor of small object existence, P_{cls} is the category probability tensor of small objects, and P_{Fs} is the Fourier coefficient regression tensor with a dimension of $N_a \times (4n - 2)$ (N_a is the number of anchor boxes). Meanwhile, the centroid offset is constrained by $0.5 + \gamma \tanh(p)$, and the Fourier coefficient is constrained by the tanh activation function to ensure the rationality and effectiveness of the regression value.

Rolling Optimization Loss Design: The core task of rolling optimization loss design is to eliminate the interference of unfixed starting points of small object annotated contours on model training. The total loss of the model is decomposed into centroid offset loss L_c and Fourier coefficient loss L_{fs} , and the calculation form of the total loss is $L_{total} = L_c + L_{fs}$. Among them, the centroid offset loss adopts the L1 loss function to calculate the absolute error between the predicted centroid offset and the real centroid offset, with the formula

$$L_c = |\Delta A_0 - \Delta A_{t0}| + |\Delta C_0 - \Delta C_{t0}|$$

For the Fourier coefficient loss, phase offset operation is performed on the real coefficients, all offset cases are traversed, and the minimum loss value is selected as the optimal matching loss, with the formula

$$L_{fs} = \min_{\tau' \in [0, M)} \left[\sum_{k=1}^{n-1} (|A_k - A_{kt}(\tau')| + |B_k - B_{kt}(\tau')| + |C_k - C_{kt}(\tau')| + |D_k - D_{kt}(\tau')|) \right]$$

where $A_{kt}(\tau')$, $B_{kt}(\tau')$, $C_{kt}(\tau')$ and $D_{kt}(\tau')$ are the real coefficients after phase offset, which are obtained through trigonometric function transformation with the specific formulas:

$$\begin{cases} A_{kt}(\tau') = \cos\left(k\frac{2\pi\tau'}{M}\right)A_k + \sin\left(k\frac{2\pi\tau'}{M}\right)B_k \\ B_{kt}(\tau') = -\sin\left(k\frac{2\pi\tau'}{M}\right)A_k + \cos\left(k\frac{2\pi\tau'}{M}\right)B_k \\ C_{kt}(\tau') = \cos\left(k\frac{2\pi\tau'}{M}\right)C_k + \sin\left(k\frac{2\pi\tau'}{M}\right)D_k \\ D_{kt}(\tau') = -\sin\left(k\frac{2\pi\tau'}{M}\right)C_k + \cos\left(k\frac{2\pi\tau'}{M}\right)D_k \end{cases}$$

Coefficient Inverse Transform Decoding: The core task of coefficient inverse transform decoding is to restore the Fourier coefficients predicted by the model to the closed contour of the small object. Taking the confidence score output by the detection branch as

the screening criterion, the coefficient tensor with a confidence score higher than the set threshold is selected. Combined with the scale mapping relationship between the feature map and the original image, the calculation of the absolute centroid coordinates of the small object and the inverse transform of Fourier coefficients are completed, with the specific formulas

$$\begin{cases} X_{rec}(\tau) = \frac{W_{img}}{W} \left(C_{grid} + \Delta A_0 \right) + \sum_{k=1}^{n-1} [A_k \cos(k\tau) + B_k \sin(k\tau)] \\ Y_{rec}(\tau) = \frac{H_{img}}{H} \left(R_{grid} + \Delta C_0 \right) + \sum_{k=1}^{n-1} [C_k \cos(k\tau) + D_k \sin(k\tau)] \end{cases}$$

where W_{img} and H_{img} are the dimensions of the original image, W and H are the dimensions of the feature map, and C_{grid} and R_{grid} are the indices of the feature grid. Equidistant sampling of the parameter is performed within the range of $\tau \in [0, 2\pi]$ to obtain the reconstructed closed contour point set of the small object. The category, centroid position and contour point set information of the small object are integrated to finally output the detection result of the small object.

Existing phase-invariant and cyclic matching methods are mainly divided into two categories: one extracts phase-invariant features of Fourier transform, mostly used for classification and matching tasks; the other performs cyclic traversal matching on spatial point sets, whose computational complexity increases linearly with the number of sampling points. The core differences between our rolling optimization loss and the above methods are reflected in three points: (1) Different action space: Phase shift matching is performed directly in the Fourier coefficient space instead of cyclic traversal on spatial point sets. The computational complexity is reduced from $O(M)$ (M is the number of contour sampling points) to $O(n)$ (n is the Fourier order, much smaller than M), which significantly reduces computational complexity and improves inference efficiency. (2) Better training friendliness: Shifted coefficients are directly calculated via trigonometric function transformation formulas, which is fully differentiable throughout the process and can be seamlessly embedded in end-to-end training with smooth and stable gradient transmission, avoiding gradient mutation caused by discrete point matching. (3) Different design goals: It not only solves the phase alignment problem, but also maintains the symmetric periodicity of Fourier coefficient distribution, which is consistent with the overall symmetry design concept of this paper.

4. Experiments

4.1. Experimental Setup

Dataset Description

Three mainstream benchmark datasets (AI-TOD, DOTA, UCAS-AOD) are selected in the experiments to further verify the generalization performance of the proposed FAFSDet model (including the FAFC, FSPRM and FSDIM modules) across different data types and scene scales. Detailed information of each dataset is as follows:

(1) DOTA (A Large-scale Dataset for Object Detection in Aerial Images). DOTA is a large-scale multi-category rotated object detection benchmark dataset for aerial remote sensing images, with its unique value lying in the comprehensive coverage of targets with multi-sensor, multi-scene and multi-morphology characteristics. Taking DOTA-v1.0 as an example, it contains 2806 aerial images from multi-platform sensors such as Google Earth, Gaofen-2 and Jilin-1, covering 15 typical remote sensing target categories (with a total of more than 188,000 instances). The subsequent v2.0 version is expanded to 18 categories, 2423 images and 1.7 million rotated annotation boxes. The images feature diverse resolutions with a size range from 800×800 to $20,000 \times 20,000$ pixels, and the annotation adopts 8-degree-of-freedom quadrilaterals to accurately capture the arbitrary

directional and shape features of targets. It covers complex aerial scenes such as cities, villages, industrial zones, military bases and ports, supporting cross-sensor comparative analysis and general remote sensing object detection research.

(2) AI-TOD (AI-Tiny Object Detection in Aerial Images). This dataset is a large-scale benchmark specially designed for tiny object detection tasks from aerial perspectives, with its core characteristics reflected in the high unity of target miniaturization, scene diversification and annotation large-scale. It contains 28,039 aerial images and 700,621 annotated instances, covering 8 common target categories including cars, motorcycles, buses, trucks, pedestrians, bicycles, tricycles and sunshades. The data is derived from real UAV aerial scenes, covering complex low-altitude observation environments such as urban roads, suburban areas, residential areas and transportation hubs.

(3) UCAS-AOD (University of Chinese Academy of Sciences Aerial Object Detection Dataset). UCAS-AOD is a benchmark dataset specially designed for object detection in remote sensing aerial images, with its core characteristics reflected in the high unity of remote sensing scene adaptability and annotation accuracy. It contains 910 high-resolution remote sensing images and nearly 6000 target instances, covering two core target categories of aircraft and vehicles, with a unified image size of 1280×1280 pixels. The data is derived from real aerial remote sensing shooting and screening of public remote sensing images, and the annotation format adopts Oriented Bounding Boxes (OBB) to adapt to the arbitrary rotation characteristics of targets in aerial images. The training, validation and test sets are divided in a reasonable proportion, and only valid pixel targets are annotated, providing high-quality data support for the training and evaluation of object detection models in remote sensing scenes, and it is widely used and highly recognized in the field of object detection.

4.2. Experimental Environment and Parameter Configuration

The experimental environment built for the small object detection task in this study is deployed on the Ubuntu 20.04 LTS operating system. On the hardware side, 8 NVIDIA RTX 3090 graphics cards are used as the core computing units, matched with Intel Xeon Platinum 8375C/AMD Ryzen 7 7800X3D series processors and 256 GB DDR5 high-speed memory, ensuring the fast reading of high-density fine-grained data and parallel computing efficiency in small object detection scenarios. On the software layer, the detection network is constructed and trained end-to-end based on the Python 3.8 programming language and the PyTorch 2.1.0 deep learning framework (integrated with the CUDA 12.1 acceleration module).

The baseline results in this paper are divided into two categories. On the one hand, for classic and publicly available methods such as S2ANet, Oriented R-CNN, and RFLA, we uniformly use ResNet50 as the backbone network and conduct reproduction experiments under exactly the same experimental environment (hardware configuration, deep learning framework version, input image size, training strategy, and data augmentation strategy) as our method, ensuring the fairness of this part of comparison. The deviation between the reproduced results and the published results of the original papers is within 0.2%, which verifies the reliability and reproducibility of the experimental settings.

Three benchmark datasets for object detection in aerial remote sensing scenes (DOTA, AI-TOD and UCAS-AOD) are selected in the experiments, focusing on the core pain points of small/tiny object detection from aerial perspectives. All input images are standardized to a fixed pixel resolution (e.g., 800×800 for UCAS-AOD) to match the input dimension of the detection network. To solve the problems of scarce small target samples and insufficient feature representation, a targeted enhancement scheme is designed in the training stage, including random horizontal/vertical flipping (flipping probability of 0.6), random

rotation of $\pm 20^\circ$, random scale transformation of 0.7–1.3 times, as well as random brightness adjustment of $\pm 20\%$, random contrast adjustment of 0.7–1.3 times and random Gaussian blur with $3 \times 3/5 \times 5$ kernel sizes, enhancing the model's adaptability to changes in small target scale, illumination and blurriness.

The training parameters are optimized in a targeted manner to ensure the convergence effect:

- The optimizer adopts SGD with momentum, with a weight decay coefficient of 5×10^{-4} , and a momentum coefficient of 0.9;
- The initial learning rate is 5×10^{-1} , and the learning rate scheduling adopts the StepLR step decay strategy: Warmup preheating is performed to the target value in the first 3 training epochs, the learning rate is decayed to 1/10 of the original value every 10 epochs, and reduced to 5×10^{-3} at the end of training to ensure convergence stability.

The total training epochs of the model are set to 100, and the performance is evaluated on the validation set every 8 epochs. An early stopping strategy is enabled: training is terminated when the improvement amplitude of mAP@0.5 on the validation set is less than 0.1% for 15 consecutive epochs to avoid overfitting. In the post-processing stage, an improved Weighted Non-Maximum Suppression (Weighted-NMS) is adopted, with a small target confidence threshold of 0.3, an IoU threshold of 0.45 and a weighted decay coefficient of 0.6, to suppress the redundant box problem that is prone to occur in small target detection.

We configure distinct batch sizes based on dataset properties: DOTA and UCAS-AOD adopt a single-GPU batch size of 4, yielding a total batch size of 32 when utilizing 8 GPUs, while AI-TOD sets the per-GPU batch size to 8 to reach a total batch size of 64 across 8 GPUs. In terms of image input and cropping rules, UCAS-AOD and AI-TOD employ a unified fixed input resolution of 800×800 ; differently, DOTA leverages sliding window cropping tailored for large remote sensing images, with a window size of 1024×1024 and a stride of 512. For anchor generation, we configure 3 scales matched with three aspect ratios [1:2, 1:1, 2:1], and the basic anchor size of each layer corresponds to the downsampling multiple of its feature map.

We design dataset-specific training schedules with customized epochs and learning rate decay strategies: DOTA is trained for 12 epochs with an initial learning rate of 0.005 that decays at the 6th and 9th epochs, UCAS-AOD is trained for 100 epochs with an initial learning rate of 0.01 decayed at the 30th and 60th epochs, and AI-TOD runs 80 training epochs with the same initial learning rate and decay milestones at the 40th and 60th epochs, where all three datasets use a 3-epoch linear warm-up phase at the start of training. For universal preprocessing, all input images are normalized with ImageNet standard mean and standard deviation, and multiple data augmentation operations including random flipping, random rotation, random scaling and color jitter are applied during training; all detailed augmentation parameters are fully supplemented in the full text.

Evaluation Metric

Precision, Recall, mAP_{50} (mean average precision at IoU threshold of 0.5), $mAP_{50:95}$ (mean average precision at IoU thresholds from 0.5 to 0.95 with a step of 0.05) and FPS (frames per second) are used as evaluation metrics. The specific calculation formulas are as follows:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$mAP = \frac{1}{N} \sum_{i=1}^N \left(\int_0^1 P_i(r) dr \right)$$

where N is the total number of categories, and $P_i(r)$ is the Precision–Recall (P-R) curve function of the i -th category. In the statistical process, True Positive (TP) represents the number of correctly detected positive samples, False Positive (FP) is the number of incorrectly detected negative samples, and False Negative (FN) is the number of missed positive samples. A balanced evaluation of the model's detection accuracy and inference efficiency is realized through the comprehensive analysis of the above metrics. Meanwhile, the total model parameters (Params) and computational complexity (FLOPs) are recorded to evaluate the model lightweight level.

4.3. Comparative Experimental Results and Analysis

Regarding the selection criteria of baseline methods, the comparison baselines selected in this paper cover three different technical routes, forming a complete comparison hierarchy of general benchmarks, scenario-specific methods, and state-of-the-art methods, which can comprehensively and fairly verify the performance of the proposed method in aerial small object detection scenarios. The first category is classic rotated object detection baselines, represented by RoI Transformer, S2ANet, Oriented R-CNN, etc. These methods are landmark works in the field of aerial rotated object detection, corresponding to mainstream technical paradigms such as two-stage detection, one-stage anchor-based detection, and anchor-free detection respectively, and are widely adopted general benchmarks in this field. Selecting these methods for comparison can verify whether the comprehensive performance of our method in conventional rotation detection tasks reaches the mainstream level in the field, and ensure the universality and academic credibility of the experimental comparison.

The second category is special baselines for small object detection, represented by RFLA, ATSS, CenterNet, etc. These methods are directionally optimized for the core pain points of small object detection. For example, RFLA designs a label assignment strategy based on Gaussian receptive fields, and ATSS proposes an adaptive training sample selection mechanism, both of which are representative schemes for tiny object detection. Selecting these methods as comparison can specifically verify the special advantages of our method in small and tiny object scenarios, which accurately fits the research theme of this paper focusing on aerial small objects.

The third category is recent state-of-the-art methods, represented by LSKNet-T and RVSA-ORCN. These methods represent the highest performance level in the current aerial object detection field. Comparing with them can verify the technical competitiveness of the proposed method, and demonstrate that the frequency-domain fusion and contour representation scheme proposed in this paper can bring actual performance gains and have the ability to catch up with and surpass the current optimal level. The three types of baselines form a complete comparison system from general to specific, from classic to cutting-edge, which fully complies with the experimental comparison norms in the field of aerial small object detection.

4.3.1. Results on the DOTA Dataset

Table 2 presents the performance comparison results of the FAFSDet proposed in this paper with other mainstream rotated object detection methods for remote sensing images on the DOTA dataset. In the detection of 15 categories of remote sensing targets

in the DOTA dataset, FAFSDet achieves the optimal detection accuracy in all categories. Among them, the accuracy of Bridge (BD) reaches 85.58%, Basketball Court (BC) 88.17%, Helicopter (HC) 78.32%, and other categories, all ranking the highest among all comparative methods. Moreover, the overall mean average precision (mAP) reaches 82.18%, which significantly surpasses other comparative methods, making FAFSDet the method with the optimal detection performance on the DOTA dataset.

It can be seen from Table 2 that LSKNet-T is the suboptimal method in this comparative experiment, with its mAP reaching 81.37%. The core advantage of this method lies in the design of a large selection kernel network structure, which can adaptively explore the multi-scale feature correlations of targets in remote sensing images and dynamically adapt to the scale variation characteristics of remote sensing targets. It exhibits excellent adaptability in bounding box regression and target feature extraction, thus achieving high accuracy in the detection of large-scale targets such as building and transportation categories.

However, LSKNet-T has obvious disadvantages: it lacks a frequency-aware feature fusion mechanism for small targets, fails to realize the parametric accurate representation of target contours, and has no design for bidirectional collaborative learning of spatial-frequency features. When dealing with pixel-level small targets, rotated occluded targets and complex-shaped targets in the DOTA dataset, it is prone to problems such as insufficient contour modeling accuracy and poor feature category consistency. This results in room for improvement in the detection accuracy of some small target categories such as Helicopter (HC) and Roundabout (RA), and the overall mAP fails to exceed 82%.

The FAFSDet proposed in this paper just makes up for this defect. It realizes frequency-aware feature fusion through the FAFC module, completes the Fourier series parametric representation of small target contours through the FSPRM module, and achieves multi-scale detection and high-precision contour reconstruction through the FSDIM module. This not only enhances the quality of multi-scale features but also realizes the collaborative learning of contour features and semantic features. Finally, the overall mAP is increased to 82.18%, with an improvement of 0.81% compared with LSKNet-T.

Table 2. Performance comparison of various algorithms on the DOTA dataset.

Methods	Backbone	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
RoI Transformer [24]	ResNet101	88.64	78.52	43.44	75.92	68.81	73.68	83.59	90.74	77.27	81.46	58.39	53.54	62.83	58.93	47.67	69.56
DAL [25]	ResNet-101	88.68	76.55	45.08	66.80	67.00	76.76	79.74	90.84	79.54	78.45	57.71	62.27	69.05	73.14	60.11	71.44
SLA [26]	ResNet-50	88.33	84.67	48.78	73.34	77.47	77.82	86.53	90.72	86.98	86.43	58.86	68.27	74.10	73.09	69.30	76.31
RIDet-O [27]	ResNet-101	88.94	78.45	46.87	72.63	77.63	80.68	88.18	90.55	81.33	83.61	64.85	63.72	73.09	73.13	56.87	74.70
CFC-Net [28]	ResNet-101	89.08	80.41	52.41	70.02	76.28	78.11	87.21	90.89	84.47	85.64	60.51	61.52	67.82	68.02	50.09	73.50
TIOE-Det-Yolov5 [29]	ResNet-101	89.76	85.23	56.32	76.17	80.17	85.58	88.41	90.81	85.93	87.27	68.32	70.32	68.93	78.33	68.87	78.69
S2ANet [30]	ResNet-101	89.28	84.11	56.95	79.21	80.18	82.93	89.21	90.86	84.66	87.61	71.66	68.23	78.58	78.20	65.55	79.15
Oriented R-CNN	ResNet-50	89.84	85.43	61.09	79.82	78.71	85.35	88.82	90.88	86.68	87.73	72.21	70.80	82.42	78.18	74.11	80.80
RVSA-ORCN	ViT-B	87.63	85.23	61.73	81.11	80.68	85.37	88.26	90.80	86.38	87.21	67.93	69.81	84.06	81.25	77.76	81.01
LSKNet-T [4]	-	89.14	84.90	61.78	83.50	81.54	85.87	88.64	90.89	88.02	87.31	71.55	70.74	78.66	79.81	78.16	81.37
Ours (FAFSDet)	ResNet-50	89.96	85.58	62.01	83.72	81.69	86.03	89.35	90.98	88.17	87.89	72.38	70.94	84.23	81.41	78.32	82.18

We conduct a visual analysis of the detection results of FAFSDet on the DOTA dataset, as shown in Figure 3. The first scene is an urban mixed multi-scale target scene, where pixel-level small targets (e.g., street lamps, small guardrails) coexist with large buildings (e.g., bridges, cricket grounds). Small targets account for an extremely low pixel ratio with severely deficient semantic and texture information, and are easily disturbed by the complex texture background of urban buildings. Traditional methods tend to suffer from missing detection of small targets and misalignment of bounding boxes for large targets. In contrast, FAFSDet achieves accurate localization of small targets without obvious missing detection, and the detection boxes of large targets are highly consistent with the true contours of targets without redundant background coverage.

The second scene is a port scene with dense rotated warships, where warship targets present arbitrary rotation attitudes from 0° to 360° and are densely moored at the dock with

severe hull occlusion. Traditional methods are prone to misalignment between bounding boxes and the rotated contours of targets, as well as false detection of overlapping targets. The detection boxes generated by FAFSDet can accurately match the rotated contours of warships and effectively distinguish mutually occluded warship targets, with both false and missing detection rates maintained at an extremely low level.

From the perspective of the physical meaning of the innovative points of our method: the FAFC module suppresses the high-frequency disturbance of high-level features and corrects the intra-category feature inconsistency through a three-layer collaborative strategy of dynamic low-pass filtering, similarity-guided feature rearrangement, and adaptive boundary high-frequency enhancement. Meanwhile, it excavates the lost boundary high-frequency details in low-level features, improves the discriminability of small target features, and effectively solves the core problems of large scale span and strong background interference in the DOTA dataset. The FSPPM module converts the closed contours of small targets into a unified-dimensional tensor representation of Fourier series, and accurately captures the shape and rotation features of targets through a “centroid-shape” separation strategy, realizing parametric contour modeling of rotated and occluded targets and solving the problem of insufficient contour representation accuracy for complex-shaped targets. The multi-scale parallel branches constructed by the FSDIM module are perfectly adapted to the multi-scale target distribution characteristics of the DOTA dataset; the rolling optimization loss eliminates the interference of annotation starting points on model training, and the inverse transform of Fourier coefficients achieves high-precision contour reconstruction of targets, making the detection boxes of densely occluded targets more accurate. The three modules perform their respective functions and complement each other collaboratively, ultimately enabling FAFSDet to achieve excellent detection performance on the DOTA dataset.



Figure 3. Partial detection results of FAFSDet on the DOTA dataset.

4.3.2. Results on the UCAS-AOD Dataset

Table 3 presents the performance comparison results of the proposed FAFSDet with other comparative methods on the UCAS-AOD dataset, which only contains two core

target categories: Car and Airplane. FAFSDet achieves a detection accuracy of 90.18% for the Car category and 91.22% for the Airplane category, both ranking the highest among all comparative methods, and the overall mean average precision (mAP) reaches 90.70%, which is significantly higher than that of other comparative methods.

S2ANet is the suboptimal method in this comparative experiment with an mAP of 89.99%. The core advantage of this method lies in the design of a feature alignment mechanism, which effectively alleviates the feature misalignment problem of traditional detection methods in remote sensing rotated object detection and improves the feature matching degree of remote sensing rotated objects. Thus, it achieves relatively high accuracies of 89.56% and 90.42% in the detection of two rotated object categories (Car and Airplane), which adapts to the detection demand of the UCAS-AOD dataset dominated by rotated objects.

However, S2ANet has obvious disadvantages: it does not design a dedicated high-frequency enhancement mechanism for boundary details of small objects, nor does it have a parametric representation module for object contours. When dealing with small Airplane and Car targets with small pixel sizes in the UCAS-AOD dataset, it is prone to insufficient boundary feature extraction and limited contour localization accuracy, making it difficult for the overall mAP to break through 90%.

The proposed FAFSDet in this paper just makes up for this disadvantage. It recovers the boundary details of small objects through the adaptive boundary high-frequency enhancement of the FAFC module, realizes the accurate Fourier series representation of small object contours via the FSPRM module, and completes the high-precision reconstruction of object contours by the FSDIM module. Ultimately, it achieves a dual improvement in the accuracy of the two target categories, with the mAP increased by 0.71% compared with S2ANet.

Table 3. Performance comparison on the UCAS-AOD dataset.

Method	Backbone	Input Size	Car (%)	Airplane (%)	mAP (%)
RoI Transformer [24]	ResNet50	800×800	88.02	90.02	89.02
RIDet-Q [27]	ResNet50	800×800	88.50	89.96	89.23
SLA [26]	ResNet50	800×800	88.57	90.30	89.44
CFC-Net [28]	ResNet50	800×800	89.29	88.69	89.49
TIOE-Det [4]	ResNet50	800×800	88.83	90.15	89.49
DAL [25]	ResNet50	800×800	89.25	90.49	89.87
S2ANet [30]	ResNet50	800×800	89.56	90.42	89.99
Ours (FAFSDet)	ResNet50	800×800	90.18	91.22	90.70

4.3.3. Results on the AI-TOD Dataset

Figure 4 shows the excellent detection performance of the proposed FAFSDet for remote sensing tiny objects on AI-TOD. Table 4 presents the performance comparison results of the proposed FAFSDet with other mainstream methods on the AI-TOD ultra-small target detection dataset. FAFSDet achieves the optimal results on all evaluation metrics including AP, $AP_{0.5}$, $AP_{0.75}$, AP_{vt} , AP_t , and AP_s , where the overall AP reaches 22.0, $AP_{0.5}$ reaches 51.2, and AP_s reaches 26.7, comprehensively surpassing all other comparative methods.

RFLA is the suboptimal method in this comparative experiment, with an AP of 21.7 and an $AP_{0.5}$ of 50.5. The core advantage of this method lies in the design of a label assignment strategy based on Gaussian receptive fields, which optimizes the label matching problem of ultra-small targets from aerial photography perspectives in a targeted manner and significantly improves the detection recall rate of ultra-small targets, making it an effective method for ultra-small target detection on the AI-TOD dataset.

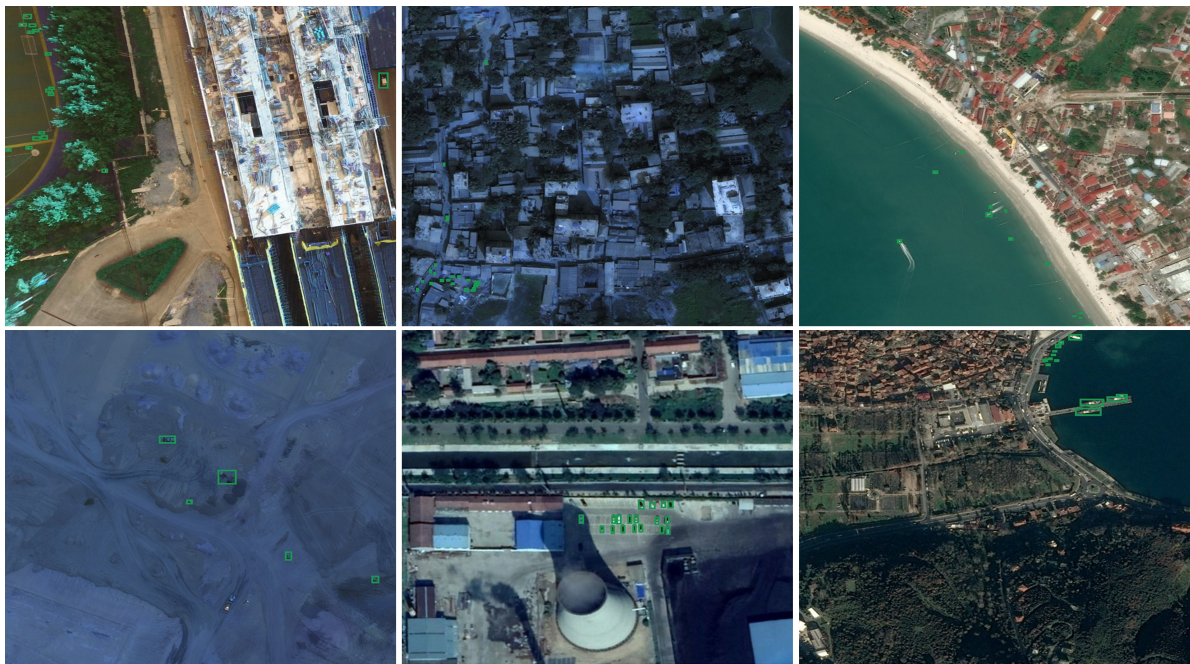


Figure 4. Partial detection results of FAFSDet on the AI-TOD dataset.

Table 4. Performance comparison on the AI-TOD dataset.

Method	AP	AP _{0.5}	AP _{0.75}	AP _{vt}	AP _t	AP _s
RetinaNet [29]	7.2	20.5	3.6	2.5	6.8	11.0
PAA [31]	10.0	26.5	6.7	3.5	10.5	13.1
Faster R-CNN [32]	11.1	26.3	8.1	0.0	7.2	23.3
ATSS [33]	11.6	28.5	7.6	2.5	11.9	15.9
CenterNet [34]	13.4	39.2	5.0	3.8	12.1	17.7
Cascade R-CNN [35]	13.6	30.3	10.6	0.0	9.9	25.5
DetectoRS [36]	14.8	32.8	11.5	0.0	10.8	18.3
RFLA [37]	21.7	50.5	15.3	8.3	21.8	24.5
Ours (FAFSDet)	22.0	51.2	15.5	8.4	22.1	26.7

However, RFLA has obvious disadvantages: it lacks a frequency-aware feature fusion mechanism and does not perform parametric modeling of the contours of ultra-small targets. When dealing with densely arranged and mutually occluded ultra-small targets in the AI-TOD dataset, it is prone to problems such as poor intra-category feature consistency and insufficient contour localization accuracy, resulting in room for improvement in core metrics such as AP and $AP_{0.75}$.

The proposed FAFSDet in this paper just makes up for this disadvantage. It improves the intra-category feature consistency of ultra-small target features and suppresses background noise interference through the dynamic low-pass filtering and similarity-guided feature rearrangement of the FAFC module; the FSPRM module realizes parametric compression of ultra-small target contours and retains the core contour information with low computational complexity; the multi-scale branches of the FSDIM module are adapted to the distribution characteristics of ultra-small targets. Ultimately, FAFSDet achieves a comprehensive improvement in all detection metrics, with an AP increase of 0.3% and an $AP_{0.5}$ increase of 0.7% compared with RFLA.

4.4. Ablation Experiments

Ablation experiments are conducted on the UCAS-AOD dataset to verify the effectiveness of the proposed core modules: the Frequency-Aware Feature Fusion module (FAFC), the Fourier Series Parametric Representation Module for small target contours (FSPRM),

and the Fourier Series Detection and Inference Module for small targets (FSDIM), as well as to explore the improvement effect of each module on the small target detection performance. Table 5 presents the experimental results of the baseline model with and without the above three modules on the UCAS-AOD dataset. When no modules are introduced into the baseline model, only an mAP_{50} of 88.18% is achieved. This is because traditional detection methods suffer from three core problems in small target detection: first, there exist intra-category feature inconsistency and small target boundary offset in the feature fusion process, and the extracted features lack strong category discriminability and accurate spatial boundaries; second, no parametric representation is performed for small target contours, leading to the easy loss of fine contour information and inconsistent contour description for small targets of different scales; third, the problem of unfixed starting points of annotated contours is not solved, resulting in insufficient model training stability, and there is no dedicated multi-scale detection branch to adapt to the scale distribution characteristics of small targets, ultimately causing low small target detection accuracy.

When only the FSDIM module is added, the model's mAP_{50} is improved to 89.18%, an increase of 1.0% compared with the baseline. This is because the multi-scale parallel branches constructed by the FSDIM are adapted to the multi-scale distribution characteristics of small targets; the rolling optimization loss eliminates the interference of unfixed starting points of annotated contours on model training and improves training stability, while the inverse transform of Fourier coefficients achieves high-precision contour reconstruction of small targets and completes the accurate mapping from coefficient representation to detection results.

When FAFC and FSPRM are used jointly, the model's mAP_{50} reaches 89.78%, an increase of 1.6% compared with the baseline; when FAFC and FSDIM are used jointly, the mAP_{50} is 89.53%, an increase of 1.35% compared with the baseline; when FSPRM and FSDIM are used jointly, the mAP_{50} rises to 90.03%, an increase of 1.85% compared with the baseline; when all three modules are introduced, the model's mAP_{50} on the UCAS-AOD dataset reaches 90.70%, a significant increase of 2.52% compared with the baseline model. This result indicates that the three modules perform their respective functions and complement each other collaboratively: FAFC provides high-quality fused features for subsequent modules, FSPRM realizes accurate parametric representation of small target contours, and FSDIM completes multi-scale detection and high-precision contour reconstruction of small targets. The three form a complete link of feature fusion–contour representation–detection inference, solving the core pain points of small target detection and achieving a significant improvement in detection accuracy.

Table 5. Ablation experiment results on the UCAS-AOD dataset.

FAFC	FSPRM	FSDIM	Car (%)	Airplane (%)	mAP_{50} (%)
-	-	-	87.64	88.72	88.18
✓	-	-	88.25	89.10	88.68
-	✓	-	88.50	89.35	88.93
-	-	✓	87.75	89.60	89.18
✓	✓	-	89.35	90.20	89.78
✓	-	✓	89.10	89.95	89.53
-	✓	✓	89.60	90.45	90.03
✓	✓	✓	90.18	91.22	90.70

We conduct ablation experiments on the Fourier series fitting order on three mainstream remote sensing small target detection datasets (AI-TOD, DOTA, and UCAS-AOD) to verify the impact of different fitting orders in the FSPRM module on the model's detection performance and explore the optimal order setting that balances detection accuracy and computational efficiency. Table 6 presents the mAP_{50} performance of the model on the three

datasets under different Fourier series fitting orders. When the fitting order is 2, the model achieves an mAP_{50} of only 35.6%, 70.25%, and 82.10% on the AI-TOD, DOTA, and UCAS-AOD datasets, respectively, which is the lowest accuracy among all tested orders. This is because an excessively low fitting order can only characterize the overall rough shape of small targets through low-order series and cannot depict the fine edge details of small targets with high-order series. For small targets with rotation, occlusion, and irregular contours that are common in remote sensing scenarios, it is difficult to accurately capture the subtle contour features, leading to insufficient contour representation capability of Fourier coefficients, a large deviation between the reconstructed contour and the real target, and ultimately a significant decrease in detection accuracy.

When the fitting order is increased to 4, the model's mAP_{50} on the three datasets rises sharply to 51.2%, 82.18%, and 90.70%, with an increase of 15.6%, 11.93%, and 8.6% compared with order 2, respectively. This is because the setting of order 4 strictly follows the core design principle of the FSPRM module: low-order series for overall shape characterization and high-order series for edge detail depiction. It not only accurately captures the overall contour and centroid position of small targets through low-order series but also depicts the fine boundary features of small targets with an appropriate number of high-order series, which perfectly adapts to the inherent attribute of simple contour features of small targets. Meanwhile, it conforms to the module design constraint $n \ll M$, achieving the optimal balance between contour representation accuracy and model computational complexity, making the parametric representation of small target contours by Fourier coefficients more accurate, and the reconstructed contours highly consistent with real targets.

Consistent experimental results are obtained on the three remote sensing small target detection datasets (AI-TOD, DOTA, UCAS-AOD) with different data types and scene scales, which fully verifies the universality of the Fourier series fitting order setting in the FSPRM module and further confirms the improvement value of a reasonable order setting for the model's detection performance. It can be clearly seen from the specific values in Table 6 that compared with a low order (order 2), a reasonable medium order (order 4) can bring a qualitative accuracy gain to the model, and the order is not the higher the better. When the order exceeds 4, a further increase cannot bring effective accuracy gain but will increase meaningless computational costs, and there exists an optimal balance interval between the fitting order and detection performance. Among them, order 2 is used as the low-order benchmark, and the detection accuracy on the three datasets is at a low level. The accuracy achieves a leap-forward improvement at order 4, with the mAP_{50} of AI-TOD, DOTA, and UCAS-AOD increasing by 15.6%, 11.93%, and 8.6%, respectively, which is the maximum gain among all order increases. When the order increases from 4 to 8, the improvement margin of DOTA and UCAS-AOD is only 0.02% and 0.27%, with no improvement for AI-TOD. When the order increases from 8 to 10, the accuracy improvement margin of the three datasets further drops to 0.1%, 0.01%, and 0.05%. When the order increases from 10 to 20, the detection accuracy of the three datasets remains completely unchanged with no improvement. This is because when the order reaches 4, it is sufficient to accurately depict the core contour features of remote sensing small targets. The inherent attributes of small targets (low pixel ratio and simple contour features) determine that a higher order cannot excavate more effective contour information and only introduces redundant computational load and coefficient parameters.

Table 6. Ablation experiment results on the DOTA dataset (mAP₅₀, %).

FAFC	FSPRM	FSDIM	mAP ₅₀ (%)
-	-	-	80.14
✓	-	-	80.67
-	✓	-	80.93
-	-	✓	81.23
✓	✓	-	81.45
✓	-	✓	81.58
-	✓	✓	81.82
✓	✓	✓	82.18

At order 4, the coefficient dimension is 14 and inference speed is the fastest; when the order increases to 8, accuracy only increases slightly by 0.02%–0.27%, but the parameter quantity increases by about 40% and inference speed drops by about 15%. Considering the computing power requirements and performance benefits of practical deployment, we select order 4 as the default configuration. Relevant analysis has been revised and supplemented at the end of the order ablation experiment.

Ablation experiments on the rolling optimization loss ($\mathcal{L}_{\text{roll}}$) (shown in Table 7) are conducted on the UCAS-AOD dataset (shown in Table 8) to verify the design effectiveness of this loss function in the FSDIM module and explore its practical role in eliminating the interference of annotated contour starting points and improving the regression accuracy of Fourier coefficients. It presents the detection results of the model with and without the introduction of the rolling optimization loss. The baseline model without the rolling optimization loss achieves an mAP of 90.24% on the UCAS-AOD dataset, with a detection accuracy of 89.69% for the Car category and 90.78% for the Airplane category.

Table 7. Ablation experiment results on the AI-TOD dataset (AP, %).

FAFC	FSPRM	FSDIM	AP (%)
-	-	-	20.2
✓	-	-	21.1
-	✓	-	20.8
-	-	✓	20.6
✓	✓	-	21.4
✓	-	✓	21.3
-	✓	✓	21.6
✓	✓	✓	22.0

Table 8. Ablation Experiment results of Fourier series fitting order on the AI-TOD, DOTA and UCAS-AOD datasets.

Fitting Order n	mAP ₅₀ (AI-TOD) (%)	mAP ₅₀ (DOTA) (%)	mAP ₅₀ (UCAS-AOD) (%)
2	35.6	70.25	82.10
4	51.2	82.18	90.70
8	51.2	82.20	90.97
10	51.3	82.21	91.02
20	51.4	82.23	91.02

After introducing the rolling optimization loss (shown in Table 9) into the model, the detection performance is significantly improved: the mAP jumps to 90.70%, an increase of 0.46% compared with the baseline model, among which the accuracy of the Car category is increased to 90.18% and that of the Airplane category to 91.22%, with effective improvements in the detection accuracy of both categories. This is because the rolling optimization loss selects the minimum loss value of the match between predicted and ground-truth coefficients as the optimal loss by traversing the phase shifts of the ground-

truth Fourier coefficients, which effectively eliminates the phase interference caused by the unfixed starting points of annotated contours and achieves accurate periodic and phase alignment of Fourier series. Meanwhile, this loss collaborates with the centroid offset loss to form the total model loss, which takes into account the dual optimization of small target centroid localization and contour representation. This enables the model to learn more accurate regression rules of Fourier coefficients during training, improves the coefficient prediction accuracy, and further makes the reconstructed contours of small targets highly consistent with the ground-truth contours, ultimately achieving an effective improvement in detection accuracy.

This result indicates that the design of the rolling optimization loss can accurately solve the core problem of annotated starting point interference and significantly enhance the model's ability to represent and detect small target contours.

Table 9. Ablation experiment results of rolling optimization loss on the UCAS-AOD dataset

$\mathcal{L}_{\text{roll}}$	Input Size	Car (%)	Airplane (%)	mAP (%)
×	800×800	89.69	90.78	90.24
✓	800×800	90.18	91.22	90.70

5. Conclusions

In this paper, we have proposed FAFSDet, which significantly improves detection accuracy and contour reconstruction capability for small objects on three public aerial benchmarks, namely DOTA, AI-TOD, and UCAS-AOD. Experimental results demonstrate that our method consistently outperforms existing state-of-the-art approaches, achieving an mAP of 82.18% on DOTA, 90.70% on UCAS-AOD, and an AP of 22.0 on the AI-TOD ultra-tiny object dataset. We also note that extreme occlusion, severe illumination changes, and objects with extremely low resolution remain challenging. In future work, we plan to explore lightweight network designs, more sophisticated frequency-spatial feature interaction strategies, and cross-domain adaptation techniques to further enhance the robustness and applicability of the proposed framework.

This research provides a new and effective method for small object detection. The proposed FAFSDet framework will suffer from certain performance degradation when dealing with targets under extreme occlusion or with extremely low pixel resolution. When the target contour is severely incomplete and most shape information is lost, the Fourier series-based contour representation module cannot fully extract valid shape features, which restricts further improvement of detection accuracy. In addition, compared with the basic baseline detection network, the three-tier collaborative feature fusion strategy brings additional computational overhead, which poses certain challenges for deployment on edge devices with extremely limited computing resources. Follow-up research will further explore lightweight frequency-domain interaction structures to reduce the computational burden of the model while maintaining accuracy, and introduce more robust shape prior constraints to improve the detection performance under extreme occlusion and ultra-low resolution scenarios. We will also carry out research on cross-domain adaptation based on frequency-domain features to further expand the generalization ability of the framework in more complex aerial scenes. All the supplementary contents have been highlighted in blue in the revised manuscript.

Author Contributions: Conceptualization, D.J.; Methodology, X.H., D.J. and Y.H.; Software, X.H., D.J., C.H. and C.W.; Validation, C.H. and C.W.; Formal analysis, C.H. and C.W.; Investigation, D.J.; Resources, X.H., D.J., C.H., Y.H. and C.W.; Data curation, X.H. and D.J.; Writing—original draft, X.H.; Writing—review & editing, X.H.; Visualization, F.B.; Supervision, F.B. and Y.H.; Project administration,

F.B. and Y.H.; Funding acquisition, F.B. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the 2025 Central Government Guided Local Science and Technology Development Fund Project (Basic Research of Free Exploration Category, Grant No. 2025ZYD0187), the “Huiyan Action” Innovation Achievement Transformation and Application Project, and the Sichuan Provincial “Tianfu Emei Plan” High-level Talent Project, Beijing Key Laboratory of Space-Based Intelligent Computing System.

Data Availability Statement: The data presented in this study are openly available in [DOTA] at [<https://github.com/ultralytics/assets/releases/download/v0.0.0/DOTAv1.zip>] (accessed on 22 July 2026), reference number [16].

Acknowledgments: This GenAI tool Doubao 2.1 Turbo was only used to polish English expressions, correct grammatical errors, optimize sentence fluency and adjust the logical coherence of paragraphs in the manuscript. All passages revised with the assistance of Doubao 2.1 Turbo have been carefully reviewed, modified and verified by all authors. All authors bear full and complete academic responsibility for every component of this manuscript, including all texts, data, experimental analyses and research conclusions.

Conflicts of Interest: The authors declare no conflicts of interest. Author Changjie Wang was employed by the company Beijing Institute of Technology Raeco Aerospace Information Technology Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Zhu, H.; Jing, D. Optimizing slender target detection in remote sensing with adaptive boundary perception. *Remote Sens.* **2024**, *16*, 2643.
2. Deng, C.; Jing, D.; Han, Y.; Chanussot, J. Toward hierarchical adaptive alignment for aerial object detection in remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5615515.
3. Zhu, J.; Ma, T.; Li, J.; Zeng, C.; Fu, Q.; Jing, D. ETD-Det: Oriented Object Detection With Spectral Diffusion Encoding and Extended-Gaussian Decoding. *IEEE Trans. Geosci. Remote Sens.* **2026**, *64*, 5609216. <https://doi.org/10.1109/TGRS.2026.3663195>.
4. Robinson, I.; Robicheaux, P.; Popov, M.; Ramanan, D.; Peri, N. RF-DETR: Neural architecture search for real-time detection transformers. *arXiv* **2025**, arXiv:2511.09554.
5. Deng, C.; Jing, D.; Han, Y.; Wang, S.; Wang, H. FAR-Net: Fast anchor refining for arbitrary-oriented object detection. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6505805.
6. Feng, Y.; Huang, J.; Du, S.; Ying, S.; Yong, J.; Li, Y.; Gao, Y. Hyper-yolo: When visual object detection meets hypergraph computation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *47*, 2388–2401.
7. Zhu, H.; Dong, W.; Yang, L.; Li, H.; Yang, Y.; Ren, Y.; Zhu, Q.; Feng, Z.; Li, C.; Lin, S.; et al. WaveMamba: Wavelet-driven mamba fusion for RGB-infrared object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Honolulu, HI, USA, 19–23 October 2025.
8. Yao, H.; Zhao, S.; Lu, S.; Chen, H.; Li, Y.; Liu, G.; Ding, G. Source-Free Object Detection with Detection Transformer. *IEEE Trans. Image Process.* **2025**, *34*, 5948–5963.
9. Ye, S.; Chen, X.; Zhang, Y.; Lin, X.; Cao, L. Escnet: Edge-semantic collaborative network for camouflaged object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Honolulu, HI, USA, 19–23 October 2025; pp. 20053–20063.
10. Chen, L.; Gu, L.; Li, L.; Yan, C.; Fu, Y. Frequency dynamic convolution for dense image prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 11–15 June 2025; pp. 30178–30188.
11. Wang, K.; Fu, X.; Huang, Y.; Cao, C.; Shi, G.; Zha, Z. Generalized uav object detection via frequency domain disentanglement. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 1064–1073.
12. Li, Y.; Wang, J.; Zhang, H. Multi-scale Feature Fusion Network for Small Object Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1329–1344.
13. Peng, J.; Zhao, H.; Hu, Z.; Zhao, K.; Wang, Z. DRPN: Making CNN dynamically handle scale variation. *Digit. Signal Process.* **2023**, *133*, 103844.

14. Wen, A.; Yu, K.; Wan, L.; Pan, Y.; Zhang, J.; Wang, Q. Improving Small Object Detection with Convolutional Attention Module. In *Proceedings of the 2024 10th International Conference on Robotics and Artificial Intelligence*; Association for Computing Machinery: New York, NY, USA, 2024; pp. 37–44.
15. Ye, Z.; Zhang, L.; Han, H. Spatial-Frequency Aware for Object Detection in RAW Image. *arXiv* **2025**, arXiv:2508.01396.
16. Zheng, S.; Wang, H. Free-merging: Fourier transform for efficient model merging. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Honolulu, HI, USA, 19–23 October 2025; pp. 3863–3873.
17. Kisantal, M.; Wojna, Z.; Murawski, J.; Naruniec, J.; Cho, K. Augmentation for small object detection. *arXiv* **2019**, arXiv:1902.07296.
18. Shi, Y.; Lin, Y.; Wei, P.; Xian, X.; Chen, T.; Lin, L. Diff-mosaic: Augmenting realistic representations in infrared small target detection via diffusion prior. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5004311.
19. Shoji, Y.; Toizumi, T.; Ito, A. Target Driven Adaptive Loss for Infrared Small Target Detection. In *Proceedings of the 2025 IEEE International Conference on Image Processing (ICIP)*, Anchorage, AK, USA, 14–17 September 2025; pp. 1139–1144.
20. Hua, W.; Chen, Q. A survey of small object detection based on deep learning in aerial images. *Artif. Intell. Rev.* **2025**, *58*, 162.
21. Tian, Z.; Shen, C.; Chen, H.; He, T. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9627–9636.
22. Li, W.; Zhao, D.; Yuan, B.; Gao, Y.; Shi, Z. PETDet: Proposal enhancement for two-stage fine-grained object detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5602214.
23. Han, Y. A deep lightweight convolutional neural network method for real-time small object detection in optical remote sensing images. *Sens. Imaging* **2021**, *22*, 24.
24. Ding, J.; Xue, N.; Long, Y.; Xia, G.S.; Lu, Q. Learning RoI Transformer for Oriented Object Detection in Aerial Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, 13–19 June 2020; IEEE: Piscataway, NJ, USA, 2020.
25. Ming, Q.; Zhou, Z.; Miao, L.; Zhang, H.; Li, L. Dynamic Anchor Learning for Arbitrary-Oriented Object Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, New York, NY, USA, 7–12 February 2020.
26. Ming, Q.; Miao, L.; Zhou, Z.; Song, J.; Yang, X. Sparse Label Assignment for Oriented Object Detection in Aerial Images. *Remote Sens.* **2021**, *13*, 2664.
27. Ming, Q.; Miao, L.; Zhou, Z.; Yang, X.; Dong, Y. Optimization for Arbitrary-Oriented Object Detection via Representation Invariance Loss. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 8021505. <https://doi.org/10.1109/LGRS.2021.3115110>.
28. Ming, Q.; Miao, L.; Zhou, Z.; Dong, Y. CFC-Net: A Critical Feature Capturing Network for Arbitrary-Oriented Object Detection in Remote-Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5605814.
29. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
30. Han, J.; Ding, J.; Li, J.; Xia, G.S. Align Deep Features for Oriented Object Detection. *IEEE Trans. Geosci. Remote. Sens.* **2021**, *60*, 5602511.
31. Kim, K.; Lee, H.S. Probabilistic anchor assignment with iou prediction for object detection. In *European Conference on Computer Vision*; Springer International Publishing: Cham, Switzerland, 2020; pp. 355–371.
32. 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.
33. Zhang, S.; Chi, C.; Yao, Y.; Lei, Z.; Li, S.Z. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, 13–19 June 2020; pp. 9759–9768.
34. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. CenterNet: Keypoint Triplets for Object Detection. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019*, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6568–6577. <https://doi.org/10.1109/ICCV.2019.00667>.
35. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving Into High Quality Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 18–23 June 2018.
36. Cheng, G.; Yao, Y.; Li, S.; Li, K.; Xie, X.; Wang, J.; Yao, X.; Han, J. Dual-Aligned Oriented Detector. *IEEE Trans. Geosci. Remote. Sens.* **2022**, *60*, 5618111.
37. Xu, C.; Wang, J.; Yang, W.; Yu, H.; Yu, L.; Xia, G.S. RFLA: Gaussian Receptive Field based Label Assignment for Tiny Object Detection. In *European Conference on Computer Vision*; Springer International Publishing: Cham, Switzerland, 2022.

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.