sensors

MDPI

Article

SPFDet: CLIP Text-Prior-Guided Structure-Enhanced Detector for Fine-Grained Ship Detection in Remote Sensing Images

Junbo Zhang [1] and Yuheng Li [2,*]

[1] School of Data Science and Big Data Technology, Central South University, Changsha 410083, China; 0106230118@csu.edu.cn
[2] School of Cyberspace Security (School of Cryptology), Hainan University, Haikou 570228, China
* Correspondence: liyuheng@hainanu.edu.cn

Abstract

Ship detection in remote sensing images is crucial for maritime surveillance, port management, and national security. However, existing detectors struggle with complex harbor backgrounds, large-scale variations, and fine-grained inter-class similarities among diverse ship categories. In this paper, we propose SPFDet, a CLIP text-prior-guided structure-enhanced detector that systematically addresses these challenges by integrating vision-language semantic priors with channel-selective feature enhancement. First, a Semantic Prior Component (SPC) employs a frozen CLIP text encoder to generate category-level semantic embeddings from textual descriptions of ship types, which are combined with visual scene priors and detail-aware priors to provide hierarchical domain-specific guidance. Second, a Structure-Enhanced Transformer Module (SETM), equipped with a Cross-level Channel Selection Module (CCSM) and Multi-Head Cross-Attention (MHCA), selectively enhances discriminative channel responses associated with hull contours, deck textures, and fine structural patterns across encoder layers. Third, a Fused Category-Dominated Feature (FCDF) generation mechanism integrates CLIP-based semantic priors with multi-scale visual features through category-guided attention, producing category-aware representations for precise classification and localization. We further introduce a Category-Semantic Consistency Loss to enforce alignment between predicted features and CLIP text priors. Extensive experiments on HRSC2016 and ShipRSImageNet benchmarks demonstrate that SPFDet achieves 97.2% and 72.8% mAP respectively, providing consistent improvements over strong detection baselines while maintaining competitive inference speed.

Keywords: ship detection; remote sensing; CLIP text prior; semantic prior; structure-enhanced transformer; vision-language model; object detection

Academic Editor: Zhe-Ming Lu

Received: 3 May 2026
Revised: 11 June 2026
Accepted: 9 July 2026
Published: 14 July 2026

1. Introduction

Ship detection in remote sensing images has become a fundamental task in earth observation, serving critical applications including maritime traffic monitoring, illegal fishing surveillance, port logistics management, and national defense [1,2]. With the rapid advancement of satellite imaging technology and the proliferation of high-resolution optical sensors, an unprecedented volume of remote sensing data is now available, demanding automated and accurate ship detection algorithms [3].

Deep learning-based object detection methods have achieved remarkable progress in natural image domains [4–6]. However, directly applying these general-purpose detectors to ship detection in remote sensing images introduces several unique challenges. First,

remote sensing images exhibit extreme scale variations—ships range from small fishing boats spanning only a few pixels to large aircraft carriers occupying hundreds of pixels, making multi-scale feature representation essential. Second, complex harbor backgrounds containing docks, buildings, and water reflections create substantial visual clutter that degrades detection precision [7]. Third, fine-grained ship categories (e.g., destroyers, cruisers, and frigates) share highly similar visual appearances, requiring the detector to capture subtle discriminative features for accurate classification [8].

Recent advances in remote sensing object detection have explored various directions to address these issues. Two-stage detectors such as Faster R-CNN [4] and Cascade R-CNN [9] provide high accuracy but suffer from slow inference. One-stage detectors including RetinaNet [5] and FCOS [10] achieve better speed-accuracy trade-offs but lack fine-grained discriminative capability. The YOLO family [11–13] delivers real-time performance yet often struggles with densely packed ships in harbors. Transformer-based detectors such as DETR [6], Deformable DETR [14], and RT-DETR [15] leverage global attention mechanisms but treat all object categories uniformly, ignoring the domain-specific semantic structure inherent in ship detection scenarios. Moreover, recent works on multi-modal remote sensing detection [16,17] and intelligent visual recognition [18] have demonstrated the importance of incorporating prior knowledge and cross-modal feature fusion for improving detection performance in complex scenes.

A critical observation motivating our work is that ships in remote sensing images possess strong categorical semantic priors. Different ship types exhibit distinctive structural patterns: warships feature elongated hulls with weapon systems, cargo ships have large container areas, and submarines present unique low-profile silhouettes. These categorical patterns provide valuable guidance that existing detectors fail to exploit. Furthermore, ship structures contain regular spatial patterns (e.g., deck layouts, container arrangements) that are reflected in multi-channel feature responses related to contours, textures, and local structural repetitions [19,20], suggesting that channel-selective feature enhancement can complement spatial-domain representations.

Recent advances in Vision-Language Models (VLMs) have demonstrated remarkable capabilities in encoding rich semantic knowledge about visual concepts [21,22]. CLIP [21], trained on massive image-text pairs, learns textual representations that capture fine-grained categorical attributes and structural characteristics while remaining aligned with visual features. Leveraging CLIP text embeddings as category-level priors offers a promising yet largely unexplored avenue for remote sensing detection, as textual descriptions of ship categories can encode discriminative structural knowledge (e.g., "an aircraft carrier with a flat flight deck and island superstructure") that complements purely visual features. Meanwhile, recent works on advanced target recognition in remote sensing [23–25] have further highlighted the importance of incorporating external knowledge and robust feature learning for accurate recognition in complex scenes.

Based on these insights, we propose SPFDet (CLIP text-prior-guided Structure-enhanced Detector), a detection framework that integrates CLIP-based semantic knowledge with structure-aware channel enhancement. Our key contributions are summarized as follows:

- We propose a Semantic Prior Component (SPC) that leverages a frozen CLIP text encoder to generate category-level semantic embeddings from textual descriptions of ship types, combined with visual scene priors and detail-aware priors, to provide hierarchical domain-specific guidance. Unlike conventional detectors that rely solely on data-driven feature extraction, SPC introduces vision-language semantic priors that improve fine-grained ship classification.
- We design a Structure-Enhanced Transformer Module (SETM) that integrates a Cross-level Channel Selection Module (CCSM) and Multi-Head Cross-Attention (MHCA)

to selectively enhance discriminative channel responses across multi-scale encoder features. SETM strengthens structure-sensitive representations, enabling the detector to capture subtle differences among ship categories.

- We introduce a Fused Category-Dominated Feature (FCDF) generation mechanism that fuses semantic priors with multi-scale visual features through category-guided attention, along with a Category-Semantic Consistency Loss that enforces alignment between predicted features and semantic priors during training.
- Extensive experiments on HRSC2016 [26] and ShipRSImageNet [8] demonstrate that SPFDet achieves competitive state-of-the-art performance. Ablation studies and visualization analyses validate the effectiveness of each proposed component.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 details the proposed SPFDet framework. Section 4 presents experimental results. Section 5 provides discussion, and Section 6 concludes the paper.

2. Related Work

2.1. Generic Object Detection

Modern object detection methods can be broadly categorized into two-stage and one-stage paradigms. Two-stage detectors, pioneered by Faster R-CNN [4], first generate region proposals and then refine them for classification and localization. Cascade R-CNN [9] extends this pipeline with multi-stage refinement, while Sparse R-CNN [27] introduces learnable proposals to eliminate hand-crafted anchors. These methods generally achieve high accuracy but at the cost of increased computational overhead.

One-stage detectors directly predict bounding boxes and class labels from feature maps. SSD [28] employs multi-scale feature maps for detection at different resolutions. RetinaNet [5] addresses the class imbalance problem through focal loss, and FCOS [10] proposes a fully convolutional anchor-free framework. The YOLO series [11–13,29–32] has continuously pushed the boundary of real-time detection, with recent versions incorporating advanced training strategies, architectural innovations, and efficient feature aggregation mechanisms. Detection frameworks based on YOLO have also been successfully adapted for domain-specific applications such as agricultural pest detection [33] and building supervision [34].

2.2. Transformer-Based Detection

The introduction of DETR [6] marked a paradigm shift by formulating object detection as a set prediction problem using Transformers [35]. Deformable DETR [14] addresses the slow convergence of DETR through deformable attention mechanisms. DINO [36] further improves performance with contrastive denoising training and mixed query selection. RT-DETR [15] achieves real-time performance by designing an efficient hybrid encoder and uncertainty-minimal query selection. Recent works have also explored cross-attention transformers for multi-modal detection tasks [17,37], demonstrating the versatility of attention mechanisms across different sensing modalities.

Vision Transformers [38] and their hierarchical variants such as Swin Transformer [39,40] have also been widely adopted as backbone networks, providing strong feature representations through self-attention mechanisms. However, most transformer-based detectors apply uniform attention across all object categories, missing the opportunity to leverage category-specific semantic structures that are particularly important in fine-grained detection scenarios.

2.3. Vision-Language Priors for Visual Recognition

Vision-Language Models (VLMs) have emerged as powerful tools for encoding semantic knowledge applicable to visual recognition tasks. CLIP [21] pioneered the alignment of visual and textual representations through contrastive learning on large-scale image-text pairs, enabling zero-shot and few-shot recognition capabilities. Subsequent works have explored leveraging text embeddings as semantic priors for various vision tasks, including open-vocabulary detection and language-driven semantic segmentation [22]. Grounding DINO [41] further demonstrated the effectiveness of grounding language representations in visual detection, achieving open-set object detection by marrying transformer-based detectors with grounded pre-training.

In the remote sensing domain, recent studies have highlighted the importance of advanced recognition paradigms and robust feature learning for accurate target identification. Hou et al. [24] proposed EfficientSARNet, a lightweight yet high-performance network for SAR image target recognition, demonstrating that efficient architectural design can achieve strong recognition accuracy. Privacy-preserving federated learning has also been explored for remote sensing recognition with adaptive resource management [23], while federated clustering approaches [25] address domain shift challenges in distributed SAR recognition scenarios. Despite these advances, the potential of CLIP text priors for guiding fine-grained ship detection in optical remote sensing images remains largely unexplored. Our work bridges this gap by employing a frozen CLIP text encoder to generate category-specific semantic embeddings that provide domain-aware guidance for the detection pipeline.

2.4. Ship Detection in Remote Sensing

Ship detection in remote sensing images has received significant attention due to its practical importance. Early deep learning approaches adapted generic detectors to aerial images with rotated bounding box predictions. R3Det [42] proposes a refined rotation detector with feature-level alignment. Oriented R-CNN [43] generates oriented proposals efficiently, while RoI Transformer [44] learns spatial transformations for oriented objects. ReDet [45] introduces rotation-equivariant networks to handle arbitrary orientations. GDet [46] models objects as Gaussian distributions for more accurate rotation representation.

Recent ship detection methods have explored frequency-aware features [20], multi-scale feature fusion [7,47], and semantic guidance [48] to improve detection accuracy. Multi-scale detection strategies have also been explored in UAV-based remote sensing scenarios [16], and slicing-aided inference has shown clear benefits for small object detection in large images [49]. In addition, complex-environment detection and attention-enhanced remote-sensing detectors have been studied in camouflage and aerial-target scenarios [50,51]. Despite these advances, existing methods rarely integrate categorical CLIP text priors with structure-aware channel enhancement in a unified framework. SPFDet addresses this gap by combining semantic prior guidance, channel-selective transformer enhancement, and category-dominated feature fusion.

3. Proposed Method

3.1. Overall Architecture

The overall architecture of SPFDet is illustrated in Figure 1. Given an input remote sensing image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$, SPFDet processes it through four main stages: (1) a backbone network extracts multi-scale visual features; (2) the Semantic Prior Component (SPC) generates three types of priors, where the category prior is derived from a frozen CLIP text encoder and the scene and detail priors are extracted from backbone features; (3) the Structure-Enhanced Transformer Module (SETM) enhances encoder features by selectively amplifying discriminative channel responses guided by semantic priors; and (4) the

Fused Category-Dominated Feature (FCDF) generation mechanism integrates CLIP-based semantic priors with visual features to produce category-aware representations for the detection head.

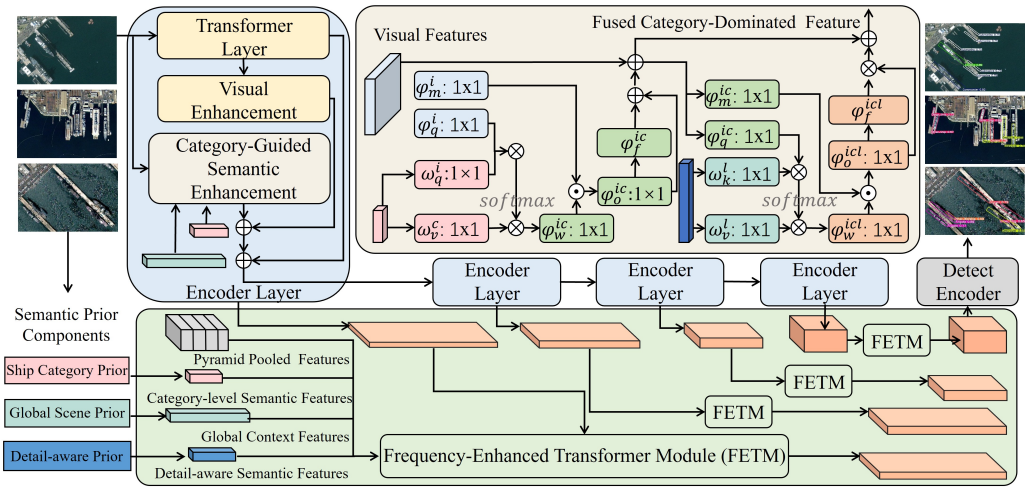

Figure 1. Overall architecture of SPFDet. The framework consists of three main components: Semantic Prior Component (SPC) that extracts ship category prior, global scene prior, and detail-aware prior; Structure-Enhanced Transformer Module (SETM) that enhances multi-scale encoder features through cross-level channel selection and cross-attention; and Fused Category-Dominated Feature (FCDF) generation that integrates semantic priors with visual features for category-aware detection.

Formally, the backbone network (e.g., ResNet-50 [52]) extracts a feature pyramid $\{C_2, C_3, C_4, C_5\}$ at strides $\{4, 8, 16, 32\}$. A Feature Pyramid Network [53] further refines these into multi-scale features $\{P_2, P_3, P_4, P_5\}$. The encoder consists of $L = 4$ stacked layers with feature dimension $d = 256$. At layer l, the visual feature $\mathbf{F}_v^l \in \mathbb{R}^{H_l W_l \times d}$ is enhanced by SETM using projected semantic priors. The category prior is broadcast over spatial tokens after projection, the scene prior is repeated over all locations as global context, and the detail prior is resized to (H_l, W_l) by bilinear interpolation before 1×1 projection. The enhanced features are then fused by FCDF and passed to the detection head for final predictions.

3.2. Semantic Prior Components

The core insight behind SPC is that ship detection benefits from domain-specific semantic knowledge at multiple granularities. Rather than relying solely on data-driven feature extraction, we leverage CLIP-encoded textual knowledge to provide category-level guidance, complemented by visual priors for scene context and spatial details. We design three complementary prior extraction branches:

3.2.1. Ship Category Prior

Unlike conventional approaches that derive category representations from visual features alone, we leverage a frozen CLIP text encoder [21] to generate category-level semantic embeddings from textual descriptions of ship types. For each ship category c_i $(i = 1, \ldots, N_c)$, we construct a descriptive text prompt t_i that encodes the structural and functional characteristics of the category, e.g., "a remote sensing image of a destroyer, characterized by an elongated hull with visible weapon systems and radar equipment." The frozen CLIP text encoder Φ_{text} maps these prompts into a high-dimensional semantic space, and a learnable projection layer aligns the text embeddings with the visual feature space:

$$\mathbf{F}_{\text{cat}} = \mathbf{W}_p \cdot \Phi_{\text{text}}(\mathcal{T}) + \mathbf{b}_p \in \mathbb{R}^{N_c \times d_s}, \quad (1)$$

where $\mathcal{T} = \{t_1, \ldots, t_{N_c}\}$ denotes the set of category-specific text prompts, Φ_{text} is the frozen CLIP text encoder that outputs embeddings in $\mathbb{R}^{d_t}$, and $\mathbf{W}_p \in \mathbb{R}^{d_s \times d_t}$, $\mathbf{b}_p \in \mathbb{R}^{d_s}$ are learnable projection parameters. The CLIP text encoder is kept frozen during training to preserve its pre-trained vision-language semantic knowledge, while only the projection layer is optimized. The resulting $\mathbf{F}_{\text{cat}}$ encodes category-level semantic features that represent the prototypical structural patterns of each ship type, providing richer and more generalizable category representations than purely data-driven embeddings.

3.2.2. Global Scene Prior

The global scene prior encodes the overall contextual information of the scene (e.g., harbor vs. open sea), which provides important cues for detection. We compute this by applying global average pooling followed by a two-layer MLP:

$$\mathbf{F}_{\text{scene}} = \text{MLP}(\text{GAP}(C_5)) \in \mathbb{R}^{1 \times d_s}, \tag{2}$$

where GAP denotes global average pooling. The scene prior captures holistic contextual features that help distinguish challenging scenarios such as densely packed harbor scenes from open-water detection.

3.2.3. Detail-Aware Prior

The detail-aware prior preserves fine-grained spatial details necessary for distinguishing visually similar ship categories. We extract it from the higher-resolution feature C_3 through a lightweight detail extraction network:

$$\mathbf{F}_{\text{detail}} = \sigma(\text{Conv}_{3\times3}(\text{Conv}_{1\times1}(C_3))) \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times d_s}, \tag{3}$$

where σ denotes the GELU activation function. The detail-aware prior retains high-resolution spatial information that captures subtle structural differences (e.g., antenna configurations, deck layouts) among fine-grained ship categories.

The three priors $\{\mathbf{F}_{\text{cat}}, \mathbf{F}_{\text{scene}}, \mathbf{F}_{\text{detail}}\}$ are then fed into SETM at different encoder layers, providing hierarchical guidance throughout the feature enhancement process. We use "semantic prior" primarily for the CLIP category prior; the scene and detail branches are visual contextual priors that complement the text prior.

3.3. Structure-Enhanced Transformer Module

SETM is designed to enhance encoder features by selectively amplifying discriminative channel responses associated with ship structures. As shown in Figure 2, each SETM block consists of two sub-modules: a Cross-level Channel Selection Module (CCSM) and a Multi-Head Cross-Attention (MHCA) mechanism, connected through layer normalization and residual connections. Unlike methods that explicitly transform features into the frequency domain using FFT, DCT, or wavelets, SETM operates in the learned feature space. We therefore use the term structure-enhanced to emphasize that CCSM selects and modulates channels whose responses are correlated with hull contours, deck textures, edges, and other structural patterns.

3.3.1. Cross-Level Channel Selection Module

The key motivation behind CCSM is that different learned channels respond to different structural cues: some channels emphasize the overall hull shape, while others respond to edges, decks, wakes, and fine details. Rather than using all channels equally, CCSM learns to select the most relevant correlated channels guided by cross-level feature similarity.

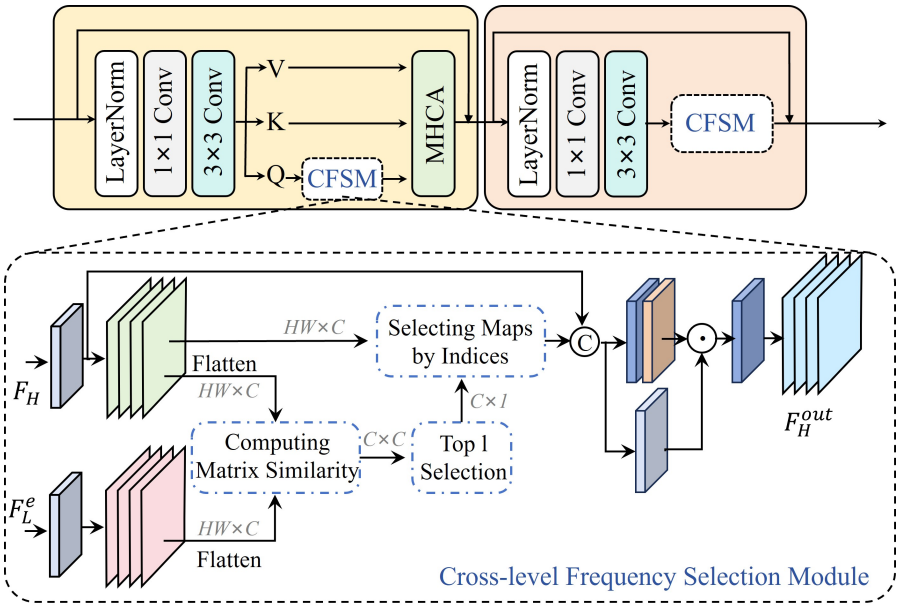

Figure 2. Architecture of the Structure-Enhanced Transformer Module (SETM). The upper portion shows the overall SETM block with LayerNorm, convolutions, MHCA, and CCSM. The lower portion details the Cross-level Channel Selection Module (CCSM), which selects the most relevant structural channel responses from a low-level enhanced feature F_L^e to enhance the high-level feature F_H.

Given a high-level feature $\mathbf{F}_H \in \mathbb{R}^{H' \times W' \times C}$ from the current encoder layer and an enhanced low-level feature $\mathbf{F}_L^e \in \mathbb{R}^{H' \times W' \times C}$ from semantic priors, CCSM operates as follows. Both features are first flattened along the spatial dimension:

$$\hat{\mathbf{F}}_H = \text{Flatten}(\mathbf{F}_H) \in \mathbb{R}^{HW \times C}, \quad \hat{\mathbf{F}}_L = \text{Flatten}(\mathbf{F}_L^e) \in \mathbb{R}^{HW \times C}. \tag{4}$$

A channel-wise similarity matrix is then computed to measure the correlation between structural channel responses:

$$\mathbf{S} = \hat{\mathbf{F}}_L^\top \hat{\mathbf{F}}_L \in \mathbb{R}^{C \times C}, \tag{5}$$

The top-1 most similar channel index for each channel is selected:

$$\mathbf{idx} = \arg\max_j \mathbf{S}_{i,j} \in \mathbb{R}^{C \times 1}, \quad j \neq i. \tag{6}$$

The selected channel maps from $\hat{\mathbf{F}}_H$ are concatenated with the original maps and modulated through element-wise multiplication:

$$\mathbf{F}_H^{\text{out}} = \text{Conv}_{1\times1}\left(\text{Concat}\left[\hat{\mathbf{F}}_H, \hat{\mathbf{F}}_H[\mathbf{idx}]\right]\right) \odot \text{Conv}_{1\times1}(\hat{\mathbf{F}}_H) \in \mathbb{R}^{H' \times W' \times C}, \tag{7}$$

where $\odot$ denotes element-wise multiplication. This mechanism amplifies the most semantically relevant structural responses while suppressing less informative or noisy channels. We use top-1 selection because it provides a deterministic one-to-one correspondence between source and target channels, avoids introducing additional Top-K hyperparameters, and keeps the module lightweight. In preliminary design, weighted multi-channel aggregation increased computation and sometimes smoothed discriminative fine details; therefore, the compact top-1 variant was adopted for the final model.

3.3.2. Multi-Head Cross-Attention

The MHCA module enables bidirectional information exchange between structure-enhanced visual features and semantic prior features. As illustrated in Figure 3, MHCA takes two feature inputs $\mathbf{F}_{\text{PT}}$ (prior-to-transformer) and $\mathbf{F}_{\text{PI}}$ (prior-input) and computes

cross-attention in both directions. At encoder layer l, $\mathbf{F}_{\mathrm{PT}} \in \mathbb{R}^{N_l \times d_s}$ is obtained by projecting and flattening the SETM-enhanced visual feature, where $N_l = H_l W_l$. $\mathbf{F}_{\mathrm{PI}} \in \mathbb{R}^{M_l \times d_s}$ is the concatenation of projected category, scene, and detail priors, where the category and scene tokens are repeated or pooled to match the attention interface and the detail prior is resized to the current scale. If the feature dimensions differ, a 1×1 convolution or linear projection maps them to the shared dimension $d_s = 256$.

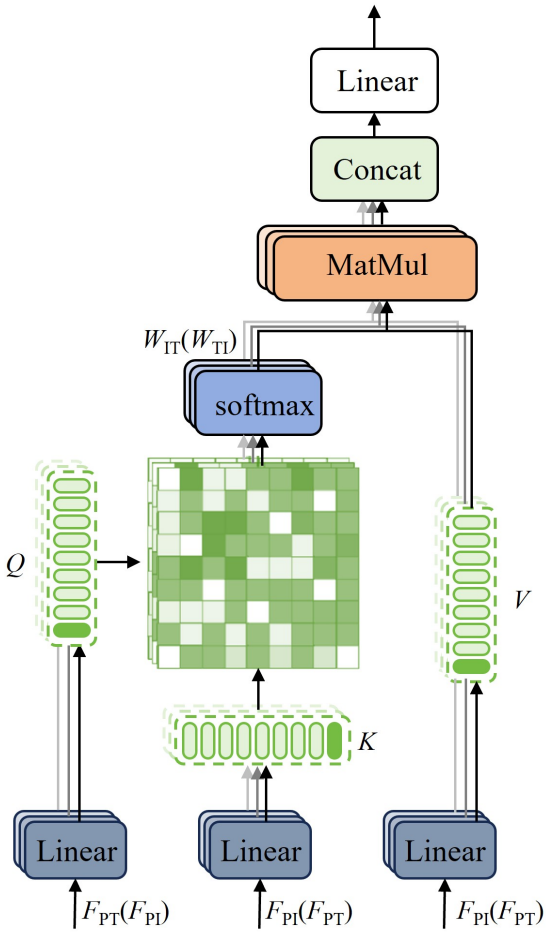

Figure 3. Structure of Multi-Head Cross-Attention (MHCA). The module computes bidirectional cross-attention between prior features $F_{\mathrm{PT}}/F_{\mathrm{PI}}$, generating attention weight matrices W_{IT} and W_{TI} for comprehensive feature interaction. The colored arrows indicate the two complementary information-flow directions and the corresponding feature-projection paths used in bidirectional cross-attention.

For the forward direction (prior-to-transformer), the query, key, and value projections are:

$$\mathbf{Q} = \mathbf{F}_{\mathrm{PT}}\mathbf{W}_Q, \quad \mathbf{K} = \mathbf{F}_{\mathrm{PI}}\mathbf{W}_K, \quad \mathbf{V} = \mathbf{F}_{\mathrm{PI}}\mathbf{W}_V, \tag{8}$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d_s \times d_h}$ are learnable projection matrices and $d_h = d_s/n_h$ is the per-head dimension with n_h attention heads. The attention weight matrix and output are computed as:

$$\mathbf{W}_{\mathrm{IT}} = \mathrm{softmax}_{M_l}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_h}}\right), \quad \mathbf{O}_{\mathrm{IT}} = \mathbf{W}_{\mathrm{IT}}\mathbf{V}. \tag{9}$$

where $\mathrm{softmax}_{M_l}$ denotes normalization along the key/prior-token dimension. Similarly, the reverse direction (transformer-to-prior) computes $\mathbf{W}_{\mathrm{TI}}$ and $\mathbf{O}_{\mathrm{TI}}$ by swapping the roles

of $\mathbf{F}_{\mathrm{PT}}$ and $\mathbf{F}_{\mathrm{PI}}$, with softmax normalization along the visual-token dimension N_l. The bidirectional outputs are concatenated and projected:

$$\mathbf{F}_{\mathrm{MHCA}} = \mathrm{Linear}(\mathrm{Concat}[\mathbf{O}_{\mathrm{IT}}, \mathbf{O}_{\mathrm{TI}}]). \tag{10}$$

This bidirectional design ensures that semantic priors are not only used to modulate visual features but also enriched by visual context, creating a synergistic enhancement loop.

3.4. Fused Category-Dominated Feature Generation

The FCDF mechanism integrates the outputs of SETM with the original encoder features through a category-guided attention scheme. As shown in the upper-right portion of Figure 1, given the visual features from the encoder layer and the SETM-enhanced semantic features, FCDF performs the following operations.

For each encoder layer l, the visual feature $\mathbf{F}_v^l \in \mathbb{R}^{H_l \times W_l \times d}$ and the fused semantic feature $\mathbf{F}_s^l \in \mathbb{R}^{H_l \times W_l \times d_s}$ from SETM are first projected through 1×1 convolutions:

$$\varphi_m^i = \mathrm{Conv}_{1\times1}^m(\mathbf{F}_v^l), \quad \varphi_q^i = \mathrm{Conv}_{1\times1}^q(\mathbf{F}_v^l), \quad \omega_q^i = \mathrm{Conv}_{1\times1}^{wq}(\mathbf{F}_s^l). \tag{11}$$

A category-guided attention map is computed through softmax normalization:

$$\varphi_f^{ic} = \mathrm{softmax}\left(\varphi_q^i \otimes \omega_q^i\right), \tag{12}$$

where $\otimes$ denotes the outer product operation. The intermediate fused feature is obtained by: Here, the visual tensor is flattened to $N_l = H_l W_l$ spatial tokens before attention. The softmax in Equation (12) is applied along the semantic/category-token dimension so that each spatial token receives normalized category-prior weights.

$$\varphi_o^{ic} = \omega_v^c \otimes \varphi_f^{ic} \oplus \varphi_m^i, \tag{13}$$

where $\omega_v^c = \mathrm{Conv}_{1\times1}^{wv}(\mathbf{F}_s^l)$ and $\oplus$ denotes element-wise addition. This process is repeated at the second stage to produce the final Fused Category-Dominated Feature:

$$\mathbf{F}_{\mathrm{FCDF}}^l = \varphi_w^{icl} \otimes \mathrm{softmax}\left(\varphi_q^{ic} \otimes \omega_k^l\right) \oplus \varphi_m^{ic}, \tag{14}$$

where $\varphi_m^{ic}, \varphi_q^{ic}, \omega_k^l$ are projections of the intermediate features and semantic features at subsequent levels. The second-stage softmax is also normalized along the key-token dimension. After attention, the output is reshaped back to $\mathbb{R}^{H_l \times W_l \times d}$ and aligned with the original visual feature by residual addition. The multi-scale FCDF features $\{\mathbf{F}_{\mathrm{FCDF}}^l\}_{l=1}^L$ are then passed to the detection encoder for final bounding box regression and classification.

3.5. Loss Function

SPFDet employs a multi-task loss consisting of three components: classification loss, regression loss, and a novel Category-Semantic Consistency (CSC) loss. The overall training procedure is summarized in Algorithm 1.

Classification Loss. We adopt the focal loss [5] to handle the class imbalance between foreground ship instances and background regions:

$$\mathcal{L}_{\mathrm{cls}} = -\frac{1}{N_{\mathrm{pos}}} \sum_{i=1}^{N} \alpha_t (1 - p_t)^\gamma \log(p_t), \tag{15}$$

where p_t is the predicted probability for the ground-truth class, α_t is the balancing factor, and γ is the focusing parameter.

Algorithm 1 Training Procedure of SPFDet

Require: Training set $\mathcal{D} = \{(\mathbf{I}_k, \mathbf{G}_k)\}_{k=1}^{K}$, backbone Φ_b, frozen CLIP text encoder Φ_{text}, SPC module Φ_s, SETM module Φ_f, FCDF module Φ_c, detection head Φ_d, text prompts $\mathcal{T}$, learning rate η, total epochs T
Ensure: Trained model parameters Θ^*

```
Initialize model parameters Θ with pretrained backbone weights
F_cat ← W_p · Φ_text(T) + b_p                          ▷ Pre-compute CLIP text-prior embeddings
for epoch = 1 to T do
    for each mini-batch (I, G) ∈ D do
        {C_l}_{l=2}^5 ← Φ_b(I)                          ▷ Extract backbone features
        {P_l}_{l=2}^5 ← FPN({C_l})                      ▷ Feature pyramid construction
        F_scene, F_detail ← Φ_s({C_l})                  ▷ Visual prior extraction
        for each encoder layer l = 1 to L do
            F_v^l ← EncoderLayer_l(P_l)                 ▷ Visual feature encoding
            F_enh^l ← Φ_f(F_v^l, F_cat, F_scene, F_detail)   ▷ SETM enhancement
            F_FCDF^l ← Φ_c(F_v^l, F_enh^l)              ▷ FCDF generation
        end for
        P̂ ← Φ_d({F_FCDF^l})                             ▷ Detection predictions
        Compute L_cls, L_reg, L_CSC via Equations (15)–(17)
        L ← λ_cls L_cls + λ_reg L_reg + λ_CSC L_CSC
        Θ ← Θ − η ∇_Θ L                                 ▷ Update parameters
    end for
    Adjust learning rate with cosine annealing schedule
end for
return Θ* ← Θ
```

Regression Loss. The bounding box regression loss combines the ℓ_1 loss and Generalized IoU loss [54] for robust localization. For HRSC2016, SPFDet predicts oriented bounding boxes represented as $\mathbf{b} = (x_c, y_c, w, h, \theta)$, where (x_c, y_c) is the box center, w and h are width and height, and θ is the rotation angle. The IoU term is computed using rotated box overlap, and the GIoU formulation follows the minimum enclosing rotated rectangle used in MMRotate. For ShipRSImageNet, the same detection head is used with the dataset's bounding-box annotation format.

$$\mathcal{L}_{\text{reg}} = \frac{1}{N_{\text{pos}}} \sum_{i=1}^{N_{\text{pos}}} \Big[\lambda_{L1} \|\hat{\mathbf{b}}_i - \mathbf{b}_i\|_1 + \lambda_{\text{GIoU}} \mathcal{L}_{\text{GIoU}}(\hat{\mathbf{b}}_i, \mathbf{b}_i) \Big], \tag{16}$$

where $\hat{\mathbf{b}}_i$ and $\mathbf{b}_i$ denote the predicted and ground-truth bounding boxes, respectively. We set $\lambda_{L1} = 5.0$ and $\lambda_{\text{GIoU}} = 2.0$ in all experiments.

Category-Semantic Consistency Loss. To enforce alignment between the predicted detection features and the semantic priors from SPC, we introduce a CSC loss based on cosine similarity:

$$\mathcal{L}_{\text{CSC}} = 1 - \frac{1}{N_{\text{pos}}} \sum_{i=1}^{N_{\text{pos}}} \frac{\mathbf{e}_i^{\text{pred}} \cdot \mathbf{e}_{c_i}^{\text{prior}}}{\|\mathbf{e}_i^{\text{pred}}\| \cdot \|\mathbf{e}_{c_i}^{\text{prior}}\|}, \tag{17}$$

where $\mathbf{e}_i^{\text{pred}} \in \mathbb{R}^{d_s}$ is the category embedding extracted from the FCDF feature at the query assigned to the i-th positive prediction, $\mathbf{e}_{c_i}^{\text{prior}} \in \mathbb{R}^{d_s}$ is the corresponding semantic prior embedding for ground-truth class c_i, and the summation runs over all N_{pos} positive matches. Positive matches are determined by the same Hungarian assignment used by the DETR-style detection head, with classification, box regression, and IoU costs. This loss encourages the model to learn detection features that are semantically consistent with the category priors, improving fine-grained classification accuracy.

Total Loss. The overall training objective is:

$$\mathcal{L} = \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}} + \lambda_{\text{CSC}}\mathcal{L}_{\text{CSC}}, \tag{18}$$

where λ_{cls}, λ_{reg}, and λ_{CSC} are balancing coefficients set to 2.0, 5.0, and 1.0, respectively.

4. Experiments

4.1. Datasets

We evaluate SPFDet on two widely-used ship detection benchmarks:

HRSC2016 [26] is a high-resolution ship collection dataset containing 1061 images collected from Google Earth, with resolutions ranging from 300 × 300 to 1500 × 900 pixels. The dataset focuses on single-class ship detection with oriented bounding box annotations. Following standard practice, we use 436 images for training and 444 images for testing, and report mean Average Precision (mAP) under both VOC2007 and VOC2012 evaluation protocols.

ShipRSImageNet [8] is a large-scale fine-grained ship detection dataset comprising 3435 images with 17,573 ship instances across 50 categories organized in a four-level hierarchy. We evaluate on the Level-3 categorization with 20 classes including Warship, Cargo, Cruiser, Submarine, Destroyer, and others. The official split provides 2507 training images and 928 testing images. We report mAP at IoU thresholds of 0.5 (mAP_{50}), 0.75 (mAP_{75}), and 0.5:0.95 (mAP).

4.2. Implementation Details

SPFDet is implemented based on PyTorch (version 2.1.0), MMDetection (version 3.2.0), and MMRotate (version 1.0.0) [55]. We adopt ResNet-50 [52] pretrained on ImageNet as the backbone. The encoder consists of $L = 4$ layers with $d = 256$ feature channels. The semantic embedding dimension d_s is set to 256, and MHCA uses $n_h = 8$ attention heads. For the CLIP-based category prior, we employ the CLIP ViT-B/32 text encoder [21] (output dimension $d_t = 512$), which is kept frozen throughout training. Category-specific text prompts are manually designed with structural descriptions for each ship type (e.g., "a remote sensing image of a submarine, characterized by a streamlined cylindrical hull with a conning tower"). Only the projection layer ($\mathbf{W}_p$, $\mathbf{b}_p$) is trainable, adding negligible parameters (∼0.13 M). Input images are resized to 800 × 800 for HRSC2016 and 1024 × 1024 for ShipRSImageNet.

We train the model using the AdamW optimizer [56] with an initial learning rate of 1×10^{-4}, weight decay of 1×10^{-4}, and batch size of 8. A cosine annealing schedule is employed over 120 epochs for HRSC2016 and 72 epochs for ShipRSImageNet. The focal loss parameters are set to $\alpha = 0.25$ and $\gamma = 2.0$. Data augmentation includes random horizontal flipping, random rotation ($\pm 90^\circ$), and multi-scale training. Unless otherwise stated, $\lambda_{\text{cls}} = 2.0$, $\lambda_{\text{reg}} = 5.0$, $\lambda_{\text{CSC}} = 1.0$, $\lambda_{L1} = 5.0$, and $\lambda_{\text{GIoU}} = 2.0$. All experiments are conducted on 4 NVIDIA A100 GPUs with 80 GB memory, and random seeds are fixed for dataset splitting and model initialization. FPS is measured on a single A100 GPU with batch size 1 after 50 warm-up iterations and 500 timed iterations, using FP32 inference and including model forward propagation and post-processing but excluding image loading from disk.

Prompt design. To reduce prompt-dependent variation, all category prompts follow the same template: "a remote sensing image of a [ship category], characterized by [category-specific structural attributes]." The category name anchors the semantic concept, while the attribute phrase describes observable geometry, deck layout, and functional components that are visible from overhead imagery. Table 1 lists the prompts used for

the 20-class ShipRSImageNet Level-3 setting. For HRSC2016, which is a single-class ship detection dataset, the category text prior mainly provides generic ship-shape guidance for background suppression and oriented localization rather than fine-grained classification.

Table 1. CLIP text prompts used to generate category priors for ShipRSImageNet Level-3.

Category	Prompt
Warship	a remote sensing image of a warship, characterized by an elongated hull, deck weapons, and radar equipment
Aircraft carrier	a remote sensing image of an aircraft carrier, characterized by a flat flight deck and island superstructure
Cruiser	a remote sensing image of a cruiser, characterized by a long combat hull and multiple radar or missile systems
Destroyer	a remote sensing image of a destroyer, characterized by an elongated hull with visible weapon systems and radar equipment
Frigate	a remote sensing image of a frigate, characterized by a compact escort hull and central superstructure
Corvette	a remote sensing image of a corvette, characterized by a small patrol hull and compact deck equipment
Submarine	a remote sensing image of a submarine, characterized by a streamlined cylindrical hull with a conning tower
Landing ship	a remote sensing image of a landing ship, characterized by a broad deck and ramp-like bow structure
Patrol ship	a remote sensing image of a patrol ship, characterized by a narrow hull and lightweight deck facilities
Auxiliary ship	a remote sensing image of an auxiliary ship, characterized by support equipment and a service-oriented deck layout
Cargo ship	a remote sensing image of a cargo ship, characterized by a large rectangular cargo area and long hull
Container ship	a remote sensing image of a container ship, characterized by stacked container blocks and a regular deck grid
Tanker	a remote sensing image of a tanker, characterized by a long smooth deck and cylindrical storage layout
Bulk carrier	a remote sensing image of a bulk carrier, characterized by large cargo hatches distributed along the deck
Fishing boat	a remote sensing image of a fishing boat, characterized by a small hull and visible working deck equipment
Tug	a remote sensing image of a tug boat, characterized by a compact hull and strong towing superstructure
Engineering ship	a remote sensing image of an engineering ship, characterized by cranes or construction equipment on deck
Dock ship	a remote sensing image of a dock ship, characterized by a large open well deck or loading structure
Civil ship	a remote sensing image of a civil ship, characterized by non-military deck layout and transport-oriented structure
Other ship	a remote sensing image of another ship type, characterized by generic hull contours and visible deck structures

Baseline reporting. For reproduced methods, we use the same dataset split, input resolution, training schedule, augmentation strategy, and evaluation protocol as SPFDet whenever the original implementation permits. Results not marked with † are taken from the corresponding papers or official reports under the closest available setting, and the backbone reported in the table is kept consistent with the cited source. All FPS values in our tables are measured under the protocol above when reproduced; literature FPS values are retained only when reproduction code or checkpoints are unavailable.

4.3. Comparison with State-of-the-Art

4.3.1. Results on HRSC2016

Table 2 presents comparisons with 17 state-of-the-art methods on HRSC2016, organized by detector category. SPFDet achieves the highest mAP under both VOC2007 (93.8%) and VOC2012 (97.2%) protocols among the compared methods.

Among two-stage detectors, Oriented R-CNN achieves strong performance (90.5%/96.5%) due to its oriented proposal generation, and SPFDet further improves the VOC2007/VOC2012 results by 3.3%/0.7%. Among one-stage methods, R3Det attains competitive 89.3%/96.1% through its rotation refinement mechanism. YOLO-based detectors show progressively improving performance with YOLOv9-C reaching 90.2%/90.7%. Among transformer-based approaches, RT-DETR-L achieves 91.2%/96.8% with its hybrid encoder design. These results indicate that semantic prior guidance and structure-aware channel enhancement are beneficial for ship detection, while the improvement on HRSC2016 is relatively moderate under the VOC2012 protocol.

Figure 4 presents qualitative detection results on HRSC2016, showing that SPFDet accurately localizes ships across diverse scenarios including densely packed harbors, isolated vessels in open water, and ships with varying orientations and scales.

Table 2. Comparison with state-of-the-art methods on HRSC2016. The best results are in **bold** and the second best are underlined. † denotes results reproduced by us.

Category	Method	Backbone	mAP (%, 07)	mAP (%, 12)	Params (M)
Two-Stage	Faster R-CNN [4]	ResNet-50	84.6	85.3	41.5
	Cascade R-CNN [9]	ResNet-50	87.2	88.0	69.2
	Sparse R-CNN [27]	ResNet-50	86.5	87.1	106.1
	Oriented R-CNN [43]	ResNet-50	90.5	96.5	41.1
	RoI Transformer [44]	ResNet-50	86.2	87.4	55.3
One-Stage	SSD [28] †	VGG-16	79.5	80.8	26.3
	RetinaNet [5]	ResNet-50	82.9	84.3	36.5
	FCOS [10]	ResNet-50	84.1	85.6	32.1
	R3Det [42]	ResNet-50	89.3	96.1	41.9
YOLO-Based	YOLOX-L [30] †	CSPDarknet	87.8	88.5	54.2
	YOLOv7 [11] †	E-ELAN	88.5	89.2	37.2
	YOLOv8-L [12] †	CSPDarknet	89.6	90.1	43.6
	YOLOv9-C [13] †	GELAN	90.2	90.7	25.5
Transformer	DETR [6]	ResNet-50	83.5	84.2	41.3
	Deformable DETR [14]	ResNet-50	87.3	88.6	40.0
	DINO [36]	ResNet-50	90.1	90.8	47.5
	RT-DETR-L [15]	HGNetv2	91.2	96.8	32.0
	SPFDet (Ours)	ResNet-50	**93.8**	**97.2**	38.7

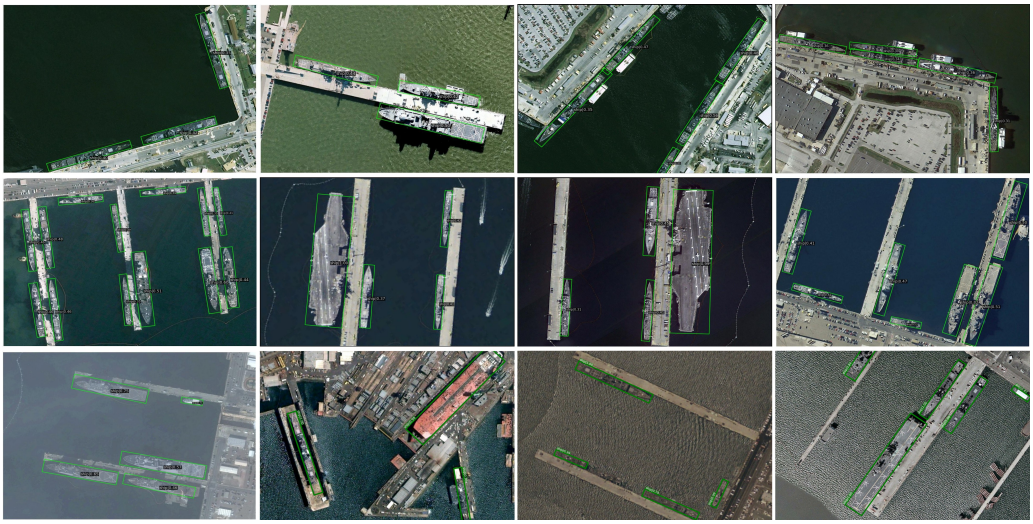

Figure 4. Qualitative detection results of SPFDet on the HRSC2016 dataset. The green bounding boxes indicate detected ships. Our method successfully handles challenging scenarios including dense ship clusters, varying scales, and complex harbor backgrounds.

4.3.2. Results on ShipRSImageNet

Table 3 reports results on the more challenging ShipRSImageNet dataset with 20 fine-grained ship categories. SPFDet achieves the best performance across all metrics with 72.8% mAP_{50}, 48.5% mAP_{75}, and 43.2% mAP, demonstrating strong fine-grained classification capability.

The performance gains on ShipRSImageNet are more pronounced than on HRSC2016, which is expected since the fine-grained multi-class nature of ShipRSImageNet better demonstrates the advantages of our semantic prior guidance and category-dominated feature fusion. Compared to the second-best RT-DETR-L, SPFDet improves mAP_{50} by 3.5%, mAP_{75} by 2.7%, and mAP by 1.7%, highlighting the effectiveness of category-aware detection.

To further analyze the per-category performance, Figure 5 presents the AP_{50} of each ship category for four representative methods. SPFDet consistently outperforms all competitors across every category. The improvements are particularly pronounced for challenging categories with fewer training samples (e.g., Tug, Other, Civil Ship) and categories with

high inter-class similarity (e.g., Cruiser vs. Destroyer, Frigate vs. Corvette), confirming that our semantic prior guidance effectively enhances fine-grained discrimination.

Table 3. Comparison with state-of-the-art methods on ShipRSImageNet (Level-3, 20 classes). The best results are in **bold** and the second best are underlined.

Category	Method	mAP_{50} (%)	mAP_{75} (%)	mAP (%)	FPS
Two-Stage	Faster R-CNN [4]	58.3	34.7	32.1	18.2
	Cascade R-CNN [9]	61.5	37.8	34.6	14.5
	Sparse R-CNN [27]	60.2	36.3	33.5	22.8
	Oriented R-CNN [43]	65.8	41.2	37.9	16.7
	ReDet [45]	66.3	42.0	38.5	13.2
One-Stage	SSD [28]	48.7	26.3	24.8	62.5
	RetinaNet [5]	55.4	32.1	29.8	32.1
	FCOS [10]	57.8	33.9	31.5	35.4
	R3Det [42]	63.7	39.5	36.2	21.5
YOLO-Based	YOLOX-L [30]	61.2	37.4	34.3	52.6
	YOLOv7 [11]	63.5	39.1	35.8	55.8
	YOLOv8-L [12]	65.1	40.6	37.2	58.3
	YOLOv9-C [13]	66.7	42.3	38.7	48.7
Transformer	DETR [6]	54.6	31.2	28.7	28.3
	Deformable DETR [14]	63.8	40.1	36.8	25.6
	DINO [36]	68.5	44.7	40.6	22.4
	RT-DETR-L [15]	69.3	45.8	41.5	45.2
	SPFDet (Ours)	**72.8**	**48.5**	**43.2**	35.6

Figure 6 illustrates detection results on ShipRSImageNet, where SPFDet correctly classifies diverse ship types including Cruiser, Destroyer, Submarine, Cargo, and Dock ships with high confidence scores, even in cluttered harbor environments with multiple ship types.

Figure 7 provides a visual comparison between SPFDet and representative methods from different categories (Faster R-CNN, RT-DETR, and YOLOv9-C). Our method produces accurate bounding boxes with fewer missed detections and false positives in many small and densely arranged ship scenes.

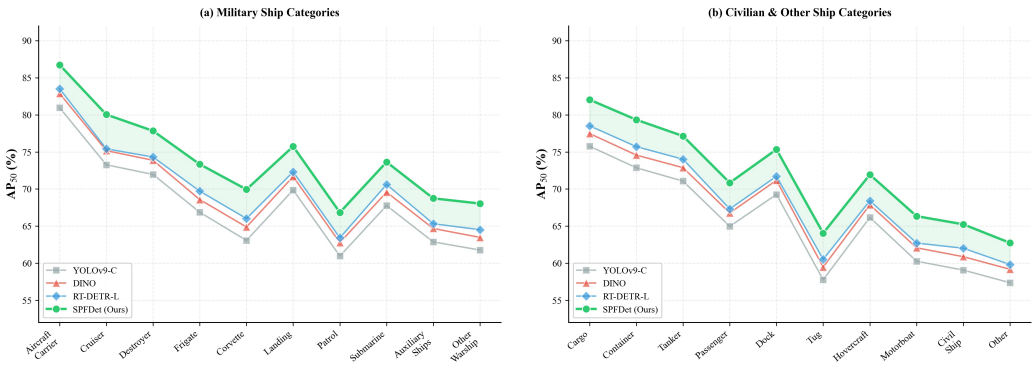

Figure 5. Per-category AP_{50} (%) comparison on ShipRSImageNet. (**a**) Military ship categories. (**b**) Civilian and other ship categories. The green shaded area highlights the performance margin of SPFDet over the second-best method. Our method achieves consistent improvements across all 20 categories, with pronounced gains on fine-grained military types and data-scarce civilian categories.

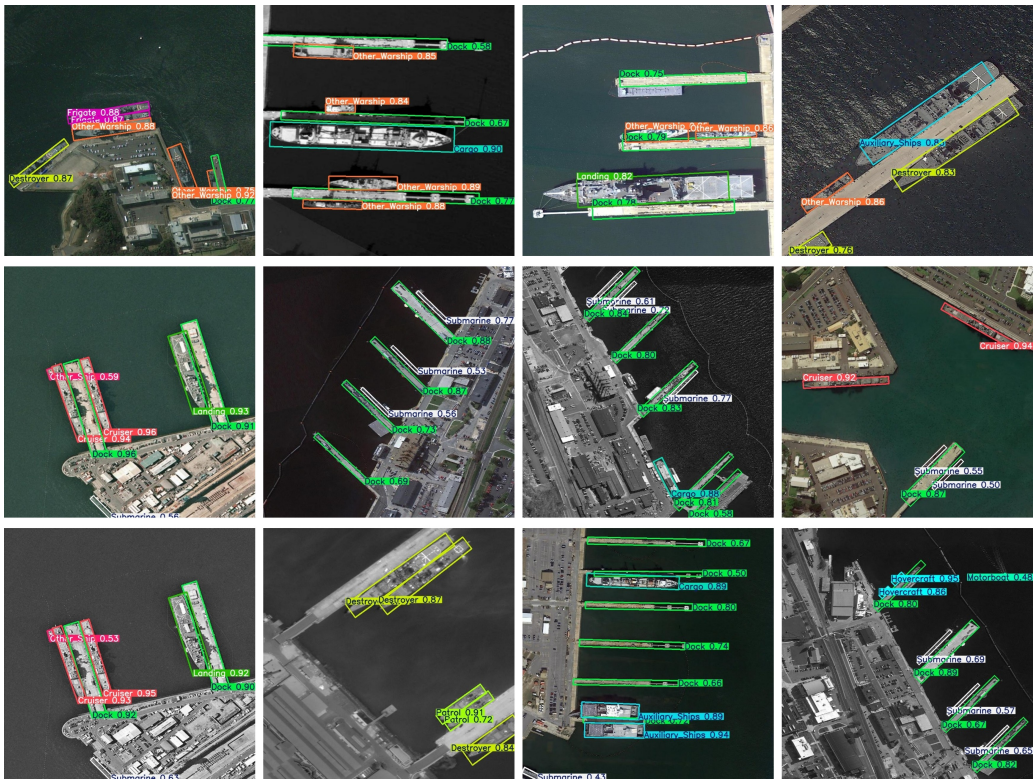

Figure 6. Qualitative detection results of SPFDet on the ShipRSImageNet dataset. Different colors represent different ship categories. Our method demonstrates strong fine-grained classification capability across diverse ship types including Cruiser, Destroyer, Submarine, Landing, and Cargo ships.

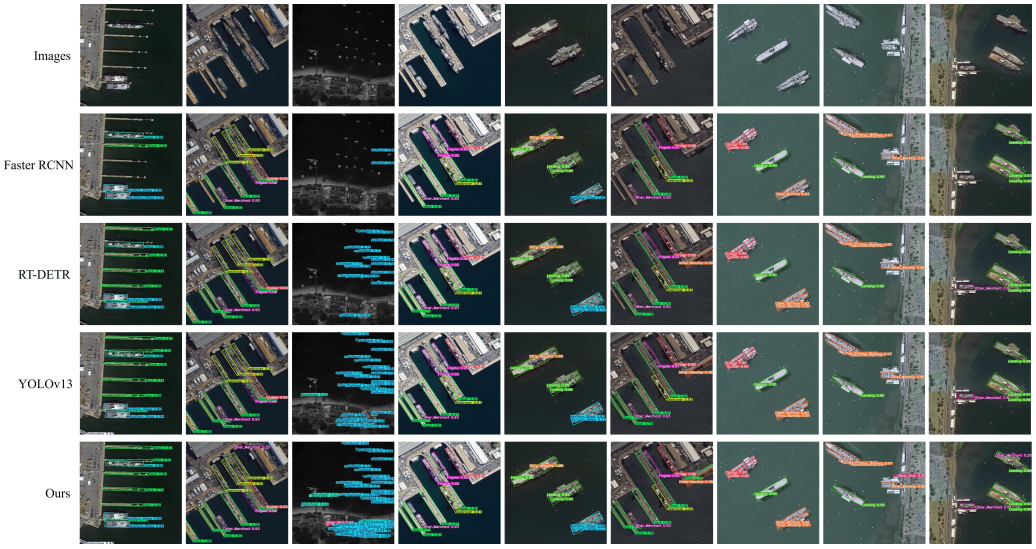

Figure 7. Visual comparison of detection results on ShipRSImageNet between different methods. From top to bottom: original images, Faster R-CNN [4], RT-DETR [15], YOLOv9-C [13], and SPFDet (Ours). Our method achieves accurate classification and localization with fewer false alarms in these examples.

4.4. Ablation Studies

4.4.1. Effectiveness of Each Component

To validate the contribution of each proposed component, we conduct ablation experiments on ShipRSImageNet by progressively adding components to a baseline detector (Deformable DETR with ResNet-50 backbone). The results are summarized in Table 4.

Table 4. Ablation study on the effectiveness of each component on ShipRSImageNet. Baseline: Deformable DETR with ResNet-50.

Configuration	SPC	SETM	MHCA	FCDF	mAP_{50} (%)	mAP (%)
Baseline					63.8	36.8
+ SPC	✓				65.9	38.3
+ SETM	✓	✓			68.4	40.1
+ MHCA	✓	✓	✓		70.6	41.8
+ FCDF (SPFDet)	✓	✓	✓	✓	**72.8**	**43.2**

Adding SPC alone improves mAP_{50} by 2.1%, confirming that semantic priors provide valuable guidance. SETM further boosts performance by 2.5%, demonstrating the effectiveness of structure-aware channel enhancement. MHCA contributes an additional 2.2% improvement through bidirectional cross-attention. Finally, FCDF brings another 2.2% gain by generating category-dominated features. The cumulative improvement of 9.0% mAP_{50} suggests that the components contribute synergistically.

4.4.2. Impact of Loss Components

Table 5 examines the contribution of each loss component. Removing the CSC loss causes a notable 1.8% drop in mAP_{50}, confirming that semantic consistency supervision is essential for fine-grained classification. Using only ℓ_1 or GIoU for regression yields suboptimal results, while their combination provides the best localization accuracy.

Table 5. Ablation study on loss components on ShipRSImageNet.

$\mathcal{L}_{cls}$	$\mathcal{L}_{L1}$	$\mathcal{L}_{GIoU}$	$\mathcal{L}_{CSC}$	mAP_{50} (%)	mAP (%)
✓	✓			69.5	40.3
✓		✓		70.1	41.0
✓	✓	✓		71.0	41.9
✓	✓	✓	✓	**72.8**	**43.2**

To further understand the training dynamics, Figure 8 visualizes the convergence behavior of different model configurations. As shown in Figure 8a, progressively adding components consistently improves the final mAP while maintaining stable convergence. The full SPFDet converges smoothly to the highest mAP_{50} of 72.8%. Figure 8b compares the total training loss, showing that the CSC loss contributes to faster and more stable convergence by providing additional semantic supervision signals.

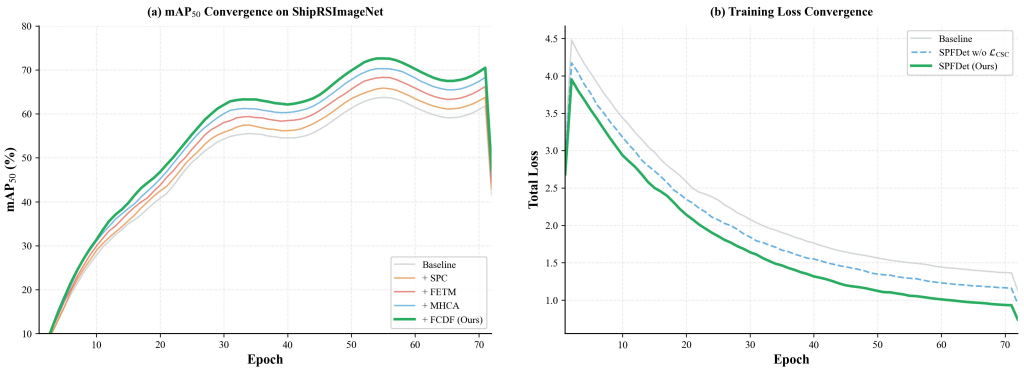

Figure 8. Training convergence analysis on ShipRSImageNet. (**a**) mAP_{50} convergence curves for progressive component additions, demonstrating that each component contributes to higher final performance with stable training. (**b**) Total loss convergence comparison, showing that the Category-Semantic Consistency loss ($\mathcal{L}_{CSC}$) enables faster and more stable convergence.

4.4.3. Impact of Semantic Prior Types

Table 6 investigates the contribution of each semantic prior type. Using only the ship category prior already provides a 1.5% improvement over the baseline without priors, validating our hypothesis that category-level patterns are important for ship detection. Adding the global scene prior further improves performance by helping the model contextualize detection scenarios. The detail-aware prior contributes the most among individual priors on the fine-grained mAP metric, as it preserves spatial details crucial for distinguishing similar ship types.

Table 6. Ablation study on semantic prior types on ShipRSImageNet.

Category Prior	Scene Prior	Detail Prior	mAP_{50} (%)	mAP (%)
			69.8	41.0
✓			71.3	42.2
	✓		71.0	42.0
		✓	71.4	42.5
✓	✓		71.8	42.6
✓		✓	72.1	42.8
✓	✓	✓	**72.8**	**43.2**

4.4.4. Impact of Category Prior Generation Strategies

A key design choice in SPFDet is the use of a frozen CLIP text encoder for generating category-level semantic priors. To validate this choice, Table 7 compares different strategies for generating the ship category prior $\mathbf{F}_{cat}$ while keeping all other components identical.

Table 7. Ablation study on category prior generation strategies on ShipRSImageNet. "Learned Emb." denotes randomly initialized learnable category embeddings optimized during training. "PPF" denotes Pyramid Pooled Features extracted from backbone feature C_5.

Prior Source	Text Encoder	Frozen	mAP_{50} (%)	mAP (%)
Random Embeddings	–	–	69.1	40.5
Learned Emb.	–	–	71.0	42.0
PPF (visual only)	–	–	71.3	42.2
Word2Vec	–	✓	71.5	42.3
BERT	bert-base	✓	72.1	42.7
CLIP Text Encoder	ViT-B/32	✓	**72.8**	**43.2**

Several observations can be drawn. First, random embeddings yield the lowest performance, confirming that meaningful category representations are essential. Second, learned embeddings and PPF-based visual priors achieve comparable results (71.0% vs. 71.3% mAP_{50}), but both are limited by the training data distribution. Third, among language model-based approaches, Word2Vec provides marginal improvement due to its limited contextual understanding, while BERT offers stronger performance through its contextualized representations. Finally, the CLIP text encoder achieves the best results, outperforming the learned embedding baseline by 1.8% mAP_{50} and 1.2% mAP. We attribute this superiority to CLIP's vision-language alignment pre-training, which produces text embeddings that are inherently compatible with visual feature spaces, facilitating more effective cross-modal semantic guidance.

4.4.5. Visualization Analysis

To provide deeper insights into how each component contributes to detection, we visualize Grad-CAM [57] activation maps for the ablation configurations in Figure 9. The baseline model produces diffuse attention that often extends to background regions. Adding SETM concentrates attention more precisely on ship regions by amplifying relevant structural channel responses. MHCA further refines the attention to focus on discriminative parts

of ships. The full SPFDet (with FCDF) achieves concentrated attention maps that highlight ship boundaries and internal structures while suppressing background distractions.

These visualizations indicate that: (1) SETM enhances channel responses corresponding to ship structural patterns; (2) MHCA enables the model to attend to discriminative ship regions through semantic-guided cross-attention; and (3) FCDF integrates these enhancements into category-dominated representations for ship instances.

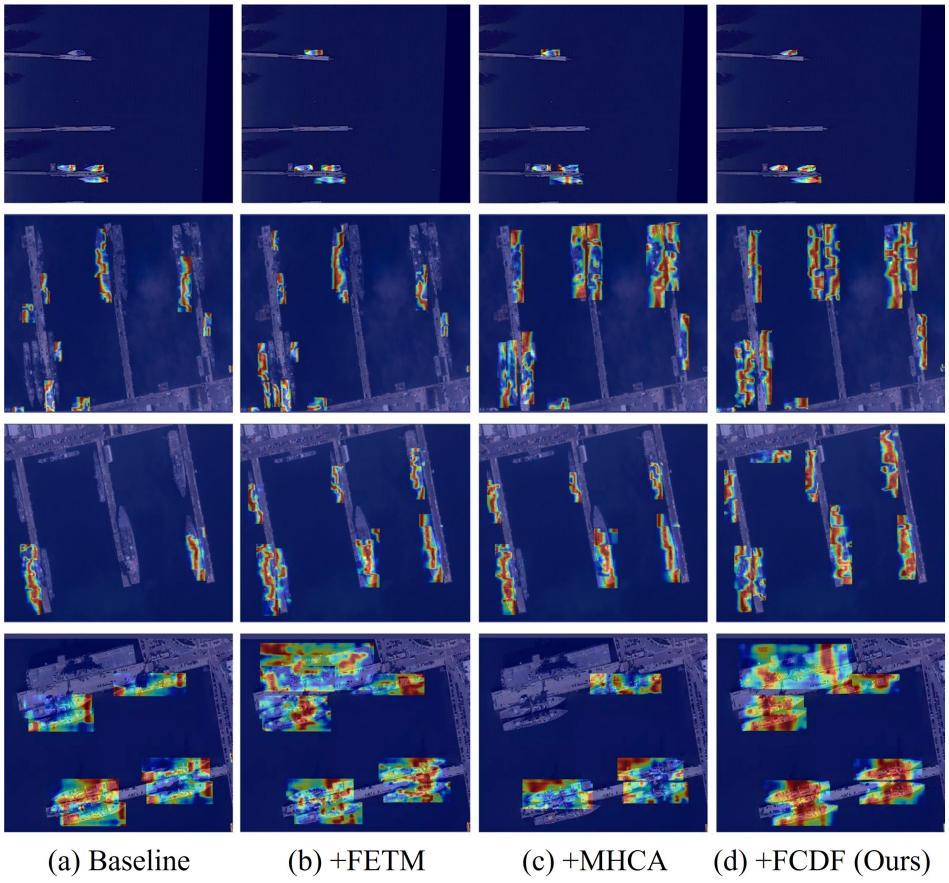

Figure 9. Grad -CAM visualization of ablation configurations on ShipRSImageNet. (**a**) Baseline shows diffuse activation. (**b**) +SETM produces more concentrated attention on ship structures. (**c**) +MHCA further refines discriminative regions. (**d**) +FCDF (Ours, full SPFDet) produces focused attention around ship boundaries and distinctive features.

4.4.6. Computational Efficiency Analysis

Table 8 reports the computational costs of SPFDet compared with representative methods. Despite incorporating additional modules, SPFDet maintains reasonable computational overhead with 38.7M parameters and 156.3 GFLOPs, comparable to RT-DETR-L while achieving higher accuracy in our setting. The inference speed of 35.6 FPS suggests practical potential for offline or near-real-time remote sensing analysis, although deployment speed may vary across hardware platforms.

Table 8. Computational efficiency comparison.

Method	Params (M)	GFLOPs	FPS	mAP_{50}(%)
Faster R-CNN [4]	41.5	134.4	18.2	58.3
DINO [36]	47.5	179.2	22.4	68.5
RT-DETR-L [15]	32.0	110.6	45.2	69.3
YOLOv9-C [13]	25.5	102.8	48.7	66.7
SPFDet (Ours)	38.7	156.3	35.6	**72.8**

5. Discussion

The experimental results demonstrate that SPFDet consistently outperforms state-of-the-art methods on both HRSC2016 and ShipRSImageNet benchmarks. We attribute this success to several factors and discuss key observations:

CLIP text priors help bridge the domain gap. Generic object detectors treat all categories uniformly, missing the opportunity to exploit domain-specific knowledge. Our SPC module introduces ship-specific inductive biases through CLIP category embeddings, visual scene priors, and detail-aware priors. The ablation study on category prior generation strategies (Table 7) shows that CLIP text embeddings outperform learned embeddings and purely visual priors in our setting, suggesting that vision-language pre-training provides useful category representations for fine-grained ship detection. The ablation study on prior types (Table 6) further shows that combining all three priors yields the best performance, as they capture complementary aspects of ship semantics. This finding aligns with recent observations in multi-modal remote sensing detection [16] and intelligent recognition [18], where incorporating external knowledge and domain priors consistently improves accuracy.

Channel selection enhances structural discrimination. Ships possess distinctive structural patterns such as hull shapes, deck configurations, and superstructures. CCSM selectively amplifies correlated structural channel responses rather than processing all channels uniformly. The Grad-CAM visualizations (Figure 9) indicate that SETM focuses attention on ship structures, producing sharper feature responses compared to the baseline.

Cross-attention enables semantic-visual synergy. The bidirectional design of MHCA creates a mutual enhancement loop: semantic priors guide visual feature refinement, while visual context enriches prior representations. This is particularly effective for handling challenging cases such as partially occluded ships or ships in unusual poses, where either semantic or visual information alone may be insufficient.

Category-dominated features improve fine-grained classification. The largest performance gains are observed on ShipRSImageNet with its 20 fine-grained categories, where SPFDet improves over the second-best method by 3.5% mAP_{50}. This demonstrates that the FCDF mechanism effectively generates category-discriminative features. In contrast, on the single-class HRSC2016, the gains are relatively smaller since fine-grained classification is not required; the semantic prior mainly helps background suppression and structure-aware localization.

Multi-dimensional comparison. To provide a holistic comparison, Figure 10 presents a radar chart evaluating SPFDet against three top-performing competitors across six dimensions: mAP_{50}, mAP_{75}, and mAP on ShipRSImageNet, mAP(VOC12) on HRSC2016, inference speed (FPS), and parameter efficiency. Each axis is min-max normalized among the compared methods, and parameter efficiency is computed as the inverse normalized parameter count so that a larger value consistently indicates a more favorable result. SPFDet obtains a large coverage area because it improves the accuracy metrics while maintaining competitive speed and model size. RT-DETR-L achieves higher FPS, YOLOv9-C has strong speed and parameter efficiency, and DINO provides competitive accuracy; SPFDet offers a balanced accuracy-efficiency trade-off in our experimental setting.

Limitations and future work. While SPFDet achieves strong performance, several directions warrant future exploration. First, the current text prompts are manually designed, and performance may be affected by prompt wording. Future work will investigate prompt ensembling, automatic prompt optimization, and stronger text encoders under controlled parameter budgets. Second, SETM does not perform an explicit FFT, DCT, or wavelet transform; it enhances learned structural channel responses. Introducing explicit spectral transforms may further improve frequency-sensitive representation. Third, the model focuses on optical remote sensing images; extending to SAR (Synthetic Aperture Radar) imagery would broaden its applicability, where recent advances in efficient SAR

recognition [24] and privacy-preserving federated learning for SAR targets [23,25] provide promising complementary techniques. Finally, recent advances in lightweight model design [58] and efficient detection architectures [32] suggest opportunities for reducing computational overhead while maintaining accuracy.

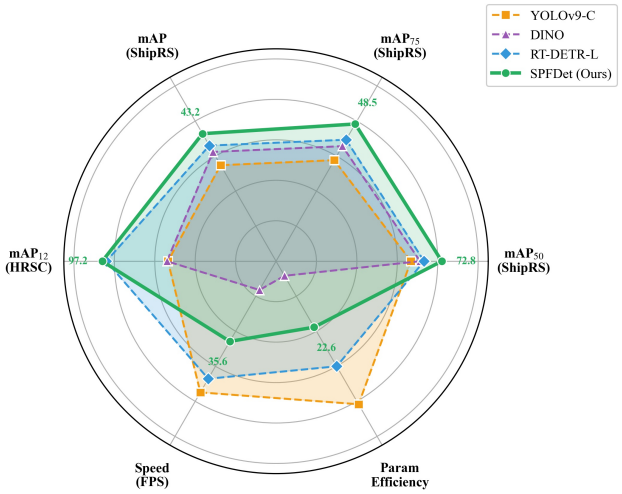

Figure 10. Multi-dimensional performance comparison among top-performing methods via radar chart. Each axis is min-max normalized among the compared methods, and parameter efficiency uses the inverse normalized parameter count. SPFDet (green, solid) provides a balanced accuracy-efficiency trade-off across detection accuracy, HRSC2016 mAP_{12}, inference speed, and parameter efficiency.

6. Conclusions

In this paper, we proposed SPFDet, a CLIP text-prior-guided structure-enhanced detector for fine-grained ship detection in remote sensing images. By leveraging a frozen CLIP text encoder to generate category-level semantic embeddings from textual descriptions of ship types and combining them with visual scene and detail-aware priors, SPFDet addresses the challenges of complex backgrounds, scale variations, and fine-grained inter-class similarities. The Structure-Enhanced Transformer Module with cross-level channel selection and multi-head cross-attention selectively amplifies discriminative structural responses guided by semantic priors, while the Fused Category-Dominated Feature generation mechanism integrates these priors with multi-scale visual features through category-guided attention for precise detection. Extensive experiments on HRSC2016 and ShipRSImageNet benchmarks demonstrate competitive state-of-the-art performance with 97.2% and 72.8% mAP respectively among the compared methods. Ablation studies confirm the benefit of CLIP-based semantic priors over conventional learned embeddings (1.8% mAP_{50} improvement) and validate the contribution of each component through quantitative analysis and Grad-CAM visualizations.

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

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The HRSC2016 dataset is publicly available at https://www.kaggle.com/datasets/guofeng/hrsc2016 (accessed on 10 June 2026). The ShipRSImageNet dataset is publicly

available at https://github.com/zzndream/ShipRSImageNet (accessed on 10 June 2026). The code, configuration files, prompt list, and evaluation scripts will be made available upon publication to facilitate reproducibility.

Acknowledgments: The authors thank the anonymous reviewers for their constructive comments that helped improve this manuscript.

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

Abbreviations

The following abbreviations are used in this manuscript:

SPFDet	CLIP Text-Prior-Guided Structure-Enhanced Detector
VLM	Vision-Language Model
CLIP	Contrastive Language-Image Pre-training
SPC	Semantic Prior Component
SETM	Structure-Enhanced Transformer Module
CCSM	Cross-level Channel Selection Module
MHCA	Multi-Head Cross-Attention
FCDF	Fused Category-Dominated Feature
CSC	Category-Semantic Consistency
FPN	Feature Pyramid Network
mAP	mean Average Precision
IoU	Intersection over Union
GIoU	Generalized Intersection over Union

References

1. Cheng, G.; Han, J. A survey on object detection in optical remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2016**, *117*, 11–28.
2. Li, K.; Wan, G.; Cheng, G.; Meng, L.; Han, J. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS J. Photogramm. Remote Sens.* **2020**, *159*, 296–307.
3. Sun, X.; Wang, P.; Yan, Z.; Xu, F.; Wang, R.; Diao, W.; Chen, J.; Li, J.; Feng, Y.; Xu, T.; et al. FAIR1M: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *184*, 116–130.
4. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28* , 91–99.
5. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 318–327.
6. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision*; Springer: Cham, Switzerland, 2020; pp. 213–229.
7. Chen, J.; Li, K.; Diao, W.; Sun, X. Ship detection in remote sensing images via selective attention and multi-scale feature fusion. *Remote Sens.* **2024**, *16*, 542.
8. Zhang, Z.; Zhang, L.; Wang, Y.; Feng, P.; He, R. ShipRSImageNet: A large-scale fine-grained dataset for ship detection in high-resolution optical remote sensing images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 8458–8472.
9. 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, 18–22 June 2018; pp. 6154–6162.
10. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: Fully convolutional one-stage object detection. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9627–9636.
11. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 7464–7475.
12. Jocher, G.; Chaurasia, A.; Qiu, J. Ultralytics YOLOv8. 2023. Available online: https://github.com/ultralytics/ultralytics (accessed on 10 June 2026).
13. Wang, C.Y.; Yeh, I.H.; Liao, H.Y.M. YOLOv9: Learning what you want to learn using programmable gradient information. In *Proceedings of the European Conference on Computer Vision*; Springer: Cham, Switzerland, 2024; pp. 1–21.

14. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable DETR: Deformable transformers for end-to-end object detection. In Proceedings of the International Conference on Learning Representations, Virtual Event, 3–7 May 2021.
15. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. RT-DETR: DETRs beat YOLOs on real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 16965–16974.
16. Yan, Z.; Li, Y. AMSRDet: An adaptive multi-scale UAV infrared-visible remote sensing vehicle detection network. *Sensors* **2026**, *26*, 817.
17. Qin, Z.; Li, Y. DCAM-DETR: Dual cross-attention mamba detection transformer for RGB–infrared anti-UAV detection. *Information* **2026**, *17*, 103.
18. Li, Y.; Wang, T.; Luo, N.; Zhou, L.; Chen, Q. Cgmamba: Intelligent identification of counterfeit goods based on state space models. *Int. J. Intell. Syst.* **2025**, *2025*, 9939880.
19. Li, L.; Wen, M.; He, L.; Deng, Y. Frequency-aware feature fusion for dense image prediction. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 10763–10780.
20. Wu, Z.; Guo, J.; Gao, J. Remote sensing ship detection with frequency-spatial interaction network. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 1–5.
21. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning, Virtual Event, 18–24 July 2021; pp. 8748–8763.
22. Li, B.; Weinberger, K.Q.; Belongie, S.; Koltun, V.; Ranftl, R. Language-driven semantic segmentation. In Proceedings of the International Conference on Learning Representations, Virtual Event, 25–29 April 2022.
23. Hou, Y.; Yu, B.; Yang, Z.; Wang, J.; Xiang, W.; Wu, D.; Liwang, M.; Xia, X.; Li, Z.; Tian, Y.; et al. Privacy-Preserving Federated SAR Image Target Recognition with Adaptive Resource Management in Space-Air-Ground Integrated Networks. *Pattern Recognit.* **2026**, *177*, 113253.
24. Hou, Y.; Zhang, L.; Wang, J.; Li, H.; Duan, X.; Huang, K.; Tian, Y. EfficientSARNet: A Light-Weight, High Performance and Fast SAR Image Target Recognition Network. *IEEE Sens. J.* **2026**, *26*, 11413–11431. https://doi.org/10.1109/JSEN.2026.3666203.
25. Hou, Y.; Zhao, S.; Xia, X.; Liwang, M.; Li, Z.; Xu, N.; Wu, D.; Tian, Y.; Quek, T.Q. FedC-DAC: A Federated Clustering with Dynamic Aggregation and Calibration Method for SAR Image Target Recognition. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *19*, 3726–3745.
26. Liu, Z.; Yuan, L.; Weng, L.; Yang, Y. A high resolution optical satellite image dataset for ship recognition and some new baselines. In *Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods—ICPRAM*; SciTePress: Setúbal, Portugal, 2017; Volume 1, pp. 324–331.
27. Sun, P.; Zhang, R.; Jiang, Y.; Kong, T.; Xu, C.; Zhan, W.; Tomizuka, M.; Li, L.; Yuan, Z.; Wang, C.; et al. Sparse R-CNN: End-to-end object detection with learnable proposals. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Virtual Event, 19–25 June 2021; pp. 14454–14463.
28. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 21–37.
29. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 779–788.
30. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. YOLOX: Exceeding YOLO series in 2021. *arXiv* **2021**, arXiv:2107.08430.
31. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W.; et al. YOLOv6: A single-stage object detection framework for industrial applications. *arXiv* **2022**, arXiv:2209.02976.
32. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. YOLOv10: Real-time end-to-end object detection. *Adv. Neural Inf. Process. Syst.* **2024**, *37*. https://doi.org/10.52202/079017-3429.
33. Li, Y.; Chen, Q.; Zhu, J.; Li, Z.; Wang, M.; Zhang, Y. PM-YOLO: A powdery mildew automatic grading detection model for rubber tree. *Insects* **2024**, *15*, 937.
34. Liu, W.; Li, Y.; Zhang, S.; Mao, R. Towards smart city supervision: A detection pipeline for illegal buildings. *Eng. Appl. Artif. Intell.* **2026**, *163*, 113052.
35. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
36. Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L.M.; Shum, H.Y. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In Proceedings of the International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023.
37. Yuan, H.; Li, Y. CSFADet: Dual-modal anti-UAV detection via cross-spectral feature alignment and adaptive multi-scale refinement. *Algorithms* **2026**, *19*, 254.

38. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. In Proceedings of the International Conference on Learning Representations, Virtual Event, 3–7 May 2021.
39. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE International Conference on Computer Vision, Virtual Event, 11–17 October 2021; pp. 10012–10022.
40. Liu, Z.; Hu, H.; Lin, Y.; Yao, Z.; Xie, Z.; Wei, Y.; Ning, J.; Cao, Y.; Zhang, Z.; Dong, L.; et al. Swin transformer v2: Scaling up capacity and resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 12009–12019.
41. Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; et al. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *Proceedings of the European Conference on Computer Vision, Milan, Italy, 29 September–4 October 2024*; Springer: Cham, Switzerland, 2024; pp. 38–55.
42. Yang, X.; Yan, J.; Feng, Z.; He, T. R3Det: Refined single-stage detector with feature refinement for rotating object. *AAAI Conf. Artif. Intell.* **2021**, *35*, 3163–3171.
43. Xie, X.; Cheng, G.; Wang, J.; Yao, X.; Han, J. Oriented R-CNN for object detection. In Proceedings of the IEEE International Conference on Computer Vision, Virtual Event, 11–17 October 2021; pp. 3520–3529.
44. Ding, J.; Xue, N.; Long, Y.; Xia, G.S.; Lu, Q. Learning RoI transformer for oriented object detection in aerial images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2849–2858.
45. Han, J.; Ding, J.; Xue, N.; Xia, G.S. ReDet: A rotation-equivariant detector for aerial object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Virtual Event, 19–25 June 2021; pp. 2786–2795.
46. Yang, X.; Yan, J. Detecting rotated objects as gaussian distributions and its 3-D generalization. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 4335–4354.
47. Liu, L.; Bai, Y.; Li, Y. Locality-aware rotated ship detection in high-resolution remote sensing imagery based on multi-scale convolutional network. *arXiv* **2020**, arXiv:2007.12326.
48. Zhang, W.; Li, H.; Sun, J. Ship detection in SAR images based on feature enhancement and semantic guidance. *Remote Sens.* **2024**, *16*, 1423.
49. Akyon, F.C.; Altinuc, S.O.; Temizel, A. Slicing Aided Hyper Inference and Fine-tuning for Small Object Detection. In *Proceedings of the 2022 IEEE International Conference on Image Processing (ICIP), Bordeaux, France, 16–19 October 2022*; IEEE: Piscataway, NJ, USA, 2022; pp. 966–970.
50. Liu, Y.; Che, S.; Ai, L.; Song, C.; Zhang, Z.; Zhou, Y.; Yang, X.; Xian, C. Camouflage detection: Optimization-based computer vision for alligator sinensis with low detectability in complex wild environments. *Ecol. Inform.* **2024**, *83*, 102802.
51. Lu, X.; Li, Z.; Zhang, Z.; Liu, R.; Liu, Y. IRT-DETR: A lightweight transformer with ConvNextV2 and Cascaded Group Attention for real-time military aircraft detection in aerial imagery. *J. Supercomput.* **2026**, *82*, 69.
52. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
53. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
54. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized intersection over union: A metric and a loss for bounding box regression. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 658–666.
55. Zhou, Y.; Yang, X.; Zhang, G.; Wang, J.; Liu, Y.; Hou, L.; Jiang, X.; Liu, X.; Yan, J.; Lyu, C.; et al. MMRotate: A rotated object detection benchmark using PyTorch. In Proceedings of the ACM International Conference on Multimedia, Lisbon, Portugal, 10–14 October 2022; pp. 7331–7334.
56. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
57. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.
58. Li, Y.; Ye, L.; Zhang, Y.; Zhou, L.; Chen, Q. Lightweight grading segmentation network for powdery mildew based on model pruning and knowledge distillation. *Smart Agric. Technol.* **2025**, *12*, 101319.