

Article

Frequency Decomposition and Spatial Dependency Mathematical Modeling for Small-Scale Open-World Object Detection

Zhengbiao Jing ¹, Qingjie Shi ², Douping Bai ¹, Baoyu Xiong ¹ and Donglin Jing ^{2,*} 

¹ Chengdu Vocational and Technical College of Industry, Chengdu 610218, China; zb182417953@163.com (Z.J.); 13708070513@163.com (D.B.); xby2021abc@163.com (B.X.)

² Shanghai Aerospace Control Technology Institute, Shanghai 201109, China; shiqingjie2022@163.com

* Correspondence: jingdonglin_bit@163.com

Abstract

Intelligent transportation and aerial remote sensing scenes suffer from complex scene variations, abundant miniature targets and unpredictable out-of-distribution obstacles, which brings tough mathematical challenges to open-world detection tasks. Conventional detection algorithms lack rigorous frequency-domain separation and spatial constraint mathematical formulations, resulting in severe tiny-object feature attenuation, inefficient multimodal feature matching and catastrophic forgetting during incremental category iteration. To solve these mathematical bottlenecks, this paper constructs the TPCA-Net model built upon frequency decomposition and spatial dependency mathematical modelling. The entire framework consists of four fixed core modules: High-Frequency-Aware Multi-Scale Feature Enhancement (HSE), Reparameterized Adaptive Text–Visual Alignment (RTA), Double Wildcard Spatial Dependency Fusion (WSF), and Incremental Forgetting-Free Dual-Path Detection (DPD). From the mathematical perspective, the HSE module adopts discrete cosine transform-based filtering equations to split high-frequency object details from low-frequency background signals and establishes cross-attention spatial constraint formulas to make up for missing contextual information of small targets. The RTA module introduces low-rank decomposition mathematical optimization and reparameterized tensor fusion rules to realize domain-adaptive text embedding calibration and zero-cost cross-modal mapping at the inference stage. The WSF module constructs dual-wildcard self-supervised mathematical loss to finish unsupervised unknown-object identification and builds decoupled semantic–spatial fusion equations to improve the positioning precision of novel targets. The DPD module designs two sets of independent optimization objective functions and category-freezing incremental mathematical constraints to avoid conflicting parameter updates and eliminate forgetting defects in new-class expansion. Validated on COCO, DOTA and AI-TOD datasets, TPCA-Net achieves 56.0% AP on COCO, 79.30% mAP on DOTA, and 40.5% overall AP with 28.7% small-object AP on AI-TOD while delivering an inference throughput of 101.2 FPS on the Tesla T4 edge GPU. The proposed method outperforms existing mainstream open-world detection algorithms in tiny-object and rare-category recognition while maintaining efficient inference speed.

Keywords: smart city; intelligent transportation; object detection; attention mechanism



Academic Editor: Lingyu Yan

Received: 9 July 2026

Revised: 29 July 2026

Accepted: 31 July 2026

Published: 4 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

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

conditions of the [Creative Commons](#)

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

1. Introduction

With the deep integration of new-type urbanization and artificial intelligence technologies, smart cities have become the core direction of urban modernization. As the core artery of smart cities, the intelligent transportation system is the key support for ensuring efficient urban traffic operation, improving travel safety, and realizing vehicle-road collaborative autonomous driving. Data-driven modeling and long-term performance monitoring provide critical quantitative insights for managing transportation infrastructure and environmental conditions [1]. Object detection, as the core basic task of the visual perception system in intelligent transportation, undertakes precise recognition and localization of traffic participants and facilities such as vehicles, pedestrians, traffic signs, and road obstacles [2–4]. Its detection accuracy, generalization ability, and real-time performance directly determine the decision-making reliability and operational safety of intelligent transportation systems [5,6].

In real traffic scenarios, the environment is highly open and dynamic. There are not only a large number of tiny objects (e.g., traffic signs, distant vehicles) with extremely low pixel proportions, but also frequent new objects (e.g., unknown non-standard obstacles, special operation vehicles) not covered in the training set. Some works [7] present a novel hybrid path planning framework that integrates Ant Colony Optimizer, Whale Optimizer, Artificial Potential Field, and a random jump mechanism, achieving significant improvements in collision avoidance and path smoothness for robotic navigation in complex environments. Traditional closed-set object detection methods cannot adapt to such dynamically open detection scenarios, and are prone to missed and false detection, bringing serious safety hazards to intelligent transportation systems [8,9]. Therefore, researching open-world object detection methods for intelligent transportation scenarios—achieving precise detection of tiny traffic objects, effective recognition of unknown objects, and forgetting-free expansion of new categories while ensuring real-time inference—has important theoretical research value and practical engineering significance [10,11].

At present, open-world and open-vocabulary object detection works are dominated by deep learning-based methods. The network architectures are mainly divided into two categories: Transformer-based multimodal fusion architectures and YOLO-series lightweight detection architectures [10,12], with the core goal of realizing recognition of unknown objects outside the training set vocabulary and zero-shot category generalization. More broadly, modular cascading residual neural networks with structural constraints demonstrate high effectiveness for multi-stage feature extraction across complex modeling tasks [13]. Gu et al. proposed the Grounding DINO model [14], which deeply integrates the DINO detector with the CLIP pre-trained model and realizes deep alignment of text and visual features through a modal interaction module, achieving excellent generalization performance in zero-shot open-vocabulary detection tasks. Zhou et al. proposed the YOLO-World series models [15], introducing the open-vocabulary pre-training paradigm into the YOLO detection architecture, and realizing efficient fusion of multi-scale features and text embeddings through a visual-language path aggregation network, greatly improving the inference speed of open-vocabulary detection. Zhao et al. proposed the UniOWL model [16], realizing accurate discovery of unknown objects in the open world and continuous learning of incremental categories through a decoupled known–unknown dual-branch detection head and an incremental contrast learning strategy, effectively alleviating the catastrophic forgetting problem. Li et al. proposed the RepYOLOE model [17], integrating text semantic information into the detection head convolution kernel through a reparameterized regional text alignment strategy, eliminating cross-modal fusion overhead in the inference stage, and balancing accuracy and speed. The above general open-world detection methods provide a basic technical framework for visual perception in traffic

scenarios. However, intelligent transportation scenarios have exclusive characteristics such as dense small objects, strong scene dynamics, and high real-time requirements, making it difficult for general methods to directly adapt. Thus, numerous studies have conducted special optimizations for the detection characteristics of traffic scenarios [18–20].

Based on the above discussion, we preliminarily analyzed the technological evolution path of open-world object detection and the dedicated optimization schemes in traffic scenarios. From current research, it can be found that existing methods still struggle to simultaneously meet the multiple requirements of precise tiny-object detection, effective unknown-object recognition, forgetting-free incremental category expansion, and edge real-time inference in open intelligent transportation scenarios [21]. Many technical bottlenecks remain in real traffic scene deployment, as detailed below:

(1) Feature degradation and localization misalignment of tiny traffic objects

In real urban traffic scenarios, tiny objects such as traffic signs, distant vehicles, and non-motor vehicles account for a very high proportion. Such objects inherently have extremely low pixel ratios Feng and Jing [22]; their features are severely compressed after multiple instances of downsampling by the backbone network, resulting in a serious lack of effective detailed information. Meanwhile, traditional feature pyramid networks adopt a pixel-wise addition fusion strategy, lacking effective modeling of spatial context around objects. Pixel shifts caused by recursive upsampling further aggravate feature misalignment of tiny objects, ultimately leading to persistently high missed and false detection rates of tiny traffic objects in open scenarios, which cannot meet the perception accuracy requirements of intelligent transportation systems.

(2) Contradiction between cross-modal alignment accuracy and inference efficiency

Most existing open-vocabulary detection methods rely on complex cross-modal fusion structures to implement text–visual feature interaction. Inference overhead rises sharply with vocabulary size, making it difficult to adapt to real-time detection demands of vehicle-mounted and roadside edge devices. Meanwhile, fixed CLIP pre-trained text embeddings suffer from significant domain shift with traffic detection tasks, failing to adapt to dynamically changing decision boundaries in open traffic scenarios. This causes a sharp drop in cross-modal alignment accuracy for rare and new traffic categories, disabling zero-shot precise recognition of novel traffic objects.

(3) Dual bottlenecks: unknown-object recognition capability and catastrophic forgetting in incremental learning

Unknown objects outside the training vocabulary frequently appear in intelligent transportation scenarios. Traditional open-vocabulary methods lack effective modeling for unknown objects and tend to misclassify them as background, inducing severe safety hazards. Meanwhile, existing detection heads adopt unified optimization objectives for known categories and unknown objects, resulting in mutually constrained detection performance that cannot be simultaneously optimized. Furthermore, when new detection categories emerge in traffic scenarios, existing methods require incremental retraining of the entire model, which is highly prone to catastrophic forgetting. This drastically degrades detection accuracy for original traffic categories and cannot adapt to the dynamic update needs of traffic scenes.

To tackle the aforementioned mathematical and computational drawbacks emerging in open intelligent transportation detection tasks, this work constructs the TPCA-Net framework from the perspective of frequency-space mathematical modeling and multimodal mapping optimization. The overall technical scheme forms a complete mathematical optimization chain consisting of four constituent full modules: High-Frequency-Aware Multi-Scale Feature Enhancement (HSE) module, Reparameterized Adaptive Text–Visual Alignment (RTA) module, Double Wildcard Spatial Dependency Fusion (WSF) module, and Incremental Forgetting-

Free Dual-Path Detection (DPD) module (List of Abbreviations is shown in Table 1). The High-Frequency-Aware Multi-Scale Feature Enhancement Module relies on discrete cosine transform mathematical filtering rules and cross-attention spatial constraint equations to refine edge feature information of micro-sized targets and supplement missing contextual constraints, thus mathematically alleviating the feature loss and spatial offset defects of small-scale objects after multi-layer downsampling. Next, the Reparameterized Adaptive Text–Visual Alignment Module introduces low-rank matrix optimization and reparameterized tensor fusion mathematical formulas to implement task-oriented calibration for pre-trained text semantic space; after parameter fusion via structural reparameterization, the cross-modal matching calculation requires no extra computational overhead during the inference phase. Afterwards, the Double Wildcard Spatial Dependency Fusion Module establishes two sets of self-supervised optimization formulas for general-object and unknown-category wildcard embeddings, and builds decoupled semantic–spatial fusion equations to accomplish unsupervised positioning and classification for unlabeled unknown traffic targets. Finally, the Incremental Forgetting-Free Dual-Path Detection Module divides the optimization objective into two mutually independent mathematical constraint paths for known and unknown targets and applies category-fixed parameter optimization rules to restrain catastrophic forgetting during incremental category updating. The main contributions of this paper are summarized as follows:

- We propose a high-frequency-aware multi-scale feature enhancement module that integrates discrete cosine transform filtering and pixel-wise cross-attention spatial dependency modeling. It effectively alleviates feature attenuation and spatial misalignment of tiny traffic objects and provides high-quality feature representation for subsequent open-world detection.
- We design a reparameterized adaptive text-visual alignment module combining low-rank decomposition calibration and tensor reparameterized fusion. It achieves task-adaptive text embedding calibration while eliminating extra cross-modal computational overhead during inference, balancing alignment accuracy and real-time performance.
- We construct a double wildcard spatial dependency fusion module with dual-wildcard self-supervised learning and decoupled semantic–spatial fusion equations. It realizes unsupervised identification of unknown objects without annotation and improves the localization accuracy of novel targets in open scenarios.
- We develop an incremental forgetting-free dual-path detection module with decoupled optimization objectives and category-frozen incremental constraints. It avoids conflicting parameter updates between known and unknown targets and fundamentally eliminates catastrophic forgetting during incremental category expansion.

Table 1. List of abbreviations.

Abbreviation	Full Name	Category
HSE	High-Frequency-Aware Multi-Scale Feature Enhancement	Core full module
RTA	Reparameterized Adaptive Text–Visual Alignment	Core full module
WSF	Double Wildcard Spatial Dependency Fusion	Core full module
DPD	Incremental Forgetting-Free Dual-Path Detection	Core full module
HFP	High-Frequency Perception	Sub-branch of HSE module
SDP	Spatial Dependency Perception	Sub-branch of HSE module
RepRTA	Reparameterized Regional Text Alignment	Sub-component of RTA module
AdaDL	Adaptive Decision Learning	Sub-component of RTA module

2. Related Works

2.1. Open-World and Open-Vocabulary Object Detection

Open-world object detection aims to break the category limitations of traditional closed-set detection, enabling effective recognition of unknown objects beyond predefined categories and zero-shot category generalization. It serves as the core technical support for visual perception in dynamic open scenarios such as autonomous driving and intelligent transportation. Open-vocabulary detection extends detection tasks from fixed categories to an open vocabulary space via the cross-modal alignment capability of vision–language pre-trained models, becoming the mainstream technical paradigm for current open-world detection [23].

Researchers deeply integrated Transformer-based detection architectures with the CLIP pre-trained model [24], achieving deep alignment between text and visual features through multi-scale modal interaction modules, which significantly improved the generalization performance of models in zero-shot open-vocabulary detection tasks. Related works introduced the open-vocabulary pre-training paradigm into lightweight YOLO detection architectures [25], realizing efficient fusion of multi-scale features and text embeddings through vision–language path aggregation networks, greatly improving the inference speed of open-vocabulary detection while maintaining detection accuracy.

Existing studies achieved accurate discovery of unknown objects and continuous learning of incremental categories in the open world via a decoupled known–unknown dual-branch detection head design combined with an incremental contrastive learning strategy, effectively alleviating catastrophic forgetting in incremental learning. Some methods proposed a reparameterized regional text alignment strategy, integrating text semantic information into the convolution kernels of detection heads through reparameterization, eliminating extra cross-modal fusion overhead during inference, and balancing the accuracy and speed of open-vocabulary detection.

2.2. Feature Enhancement Methods for Small Object Detection

Small object detection is a core challenge in intelligent transportation remote sensing and urban road perception. Limited by extremely low pixel ratios and feature degradation after multi-layer downsampling, small objects suffer from weak feature representation and high missed/false detection rates. To address the low detection accuracy of small vehicles in traffic scenarios, related works proposed multi-scale feature enhancement networks, expanding the model receptive field via dilated convolution and feature fusion modules, effectively strengthening the feature representation of small vehicles. Existing studies collaboratively designed frequency-domain filtering and spatial attention, using discrete cosine transform to separate low-frequency background and high-frequency object details in features, significantly improving the edge feature distinguishability of tiny objects with low computational cost.

Some methods designed pixel-wise cross-attention modules [26], capturing spatial dependencies between adjacent levels of the feature pyramid, correcting pixel misalignment in upper–lower feature fusion, and supplementing critical contextual semantic information for tiny objects. To meet the demand for dense small object detection in aerial remote sensing traffic scenarios, related works constructed an optimized multi-scale feature pyramid architecture, achieving balanced detection of multi-scale small objects through detail preservation in high-resolution feature layers and semantic enhancement in low-resolution layers.

Existing studies introduced a super-resolution reconstruction auxiliary branch, reconstructing high-resolution features of tiny objects via self-supervised learning, alleviating detail loss during downsampling, and greatly reducing the missed detection rate in dense small object scenarios. Some methods designed anchor optimization and label assignment strategies for the detection of extremely small objects such as traffic signs, improving

the positive sample assignment quality of small objects through adaptive receptive field matching, further enhancing the model's feature learning ability for tiny objects [27].

2.3. Vision–Language Cross-Modal Alignment Methods

Vision–language cross-modal alignment is a core component of open-vocabulary object detection. Its core goal is to establish an accurate mapping between the visual feature space and the text semantic space, which directly determines the model's zero-shot generalization ability and rare-category detection accuracy [28,29]. Related studies have conducted in-depth exploration around text embedding calibration, cross-modal interaction optimization, lightweight alignment strategies, and other directions [30].

Related works extracted category semantic embeddings through a frozen CLIP text encoder [31], combining regional contrastive loss to guide fine-grained alignment between visual features and text embeddings, significantly improving the recognition accuracy of rare categories in open-vocabulary detection. Existing studies introduced low-rank adaptation optimization into the CLIP text encoder, adapting the text semantic space to the decision boundaries of detection tasks with a small number of trainable parameters without altering pre-trained weights while fully preserving the generalization ability of the pre-trained model. Some methods designed lightweight feed-forward networks to perform task-adaptive calibration on pre-trained text embeddings, solving the domain shift problem between fixed text embeddings and downstream detection tasks, and improving cross-modal alignment accuracy with low training overhead. To solve the sample dependency issue in incremental learning [28], related works proposed data-free incremental learning methods. They reconstruct the feature distribution of original categories using generative models and apply feature distillation constraints to prevent knowledge forgetting, realizing category expansion without accessing old training samples.

3. Method

Figure 1 shows the pipeline of TPCA-Net. The left panel shows the overall framework, consisting of four core modules: high-frequency-aware multi-scale feature enhancement, reparameterized adaptive text–visual alignment, double wildcard spatial dependency fusion, and incremental forgetting-free dual-path detection. The right panel illustrates the full structure of feature enhancement and cross-modal fusion, showing the complete workflow of high-frequency feature enhancement, spatial dependency modeling, and multimodal feature interaction.

TPCA-Net constructs an end-to-end unified open-world detection architecture following the pipeline: high-frequency-aware multi-scale feature enhancement → reparameterized adaptive text–visual alignment → double wildcard spatial dependency fusion → incremental forgetting-free dual-path detection. It aims to solve the core problems in open scenarios, including feature missing of tiny / unknown objects, redundant cross-modal alignment inference, low spatial localization accuracy of unknown objects, and catastrophic forgetting in incremental category expansion.

The network first extracts multi-scale basic features via the CSPDarknet backbone. Then, the high-frequency-aware multi-scale feature enhancement module strengthens tiny-object features in the frequency domain and supplements object context information through spatial dependency modeling, addressing feature degradation of small and unknown objects in the open world.

Next, the reparameterized adaptive text–visual alignment module achieves lightweight end-to-end alignment between text and visual features through reparameterized calibration and adaptive decision boundary learning, eliminating the inference overhead of traditional cross-modal fusion.

Subsequently, the double wildcard spatial dependency fusion module is designed to enhance known-category features, realize accurate localization of unknown objects, and support dynamic category adaptation by combining double wildcard self-supervised learning and pixel-wise spatial attention.

Finally, the incremental forgetting-free dual-path detection module adopts a decoupled known/unknown dual-path detection head integrated with the lazy region contrastive strategy, enabling NMS-free efficient inference and forgetting-free incremental category expansion.

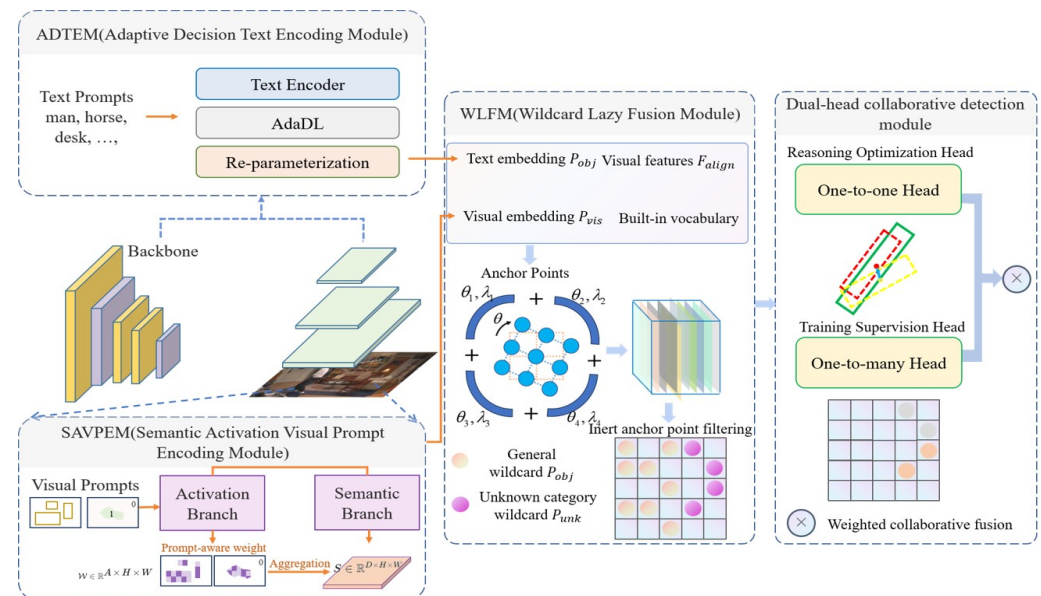


Figure 1. Overall architecture of the proposed TPCA-Net.

3.1. High-Frequency-Aware Multi-Scale Feature Enhancement Module

In open-world object detection scenarios, unknown and tiny objects account for a high proportion. Such objects inherently have an extremely low pixel ratio; their features are severely compressed after multiple downsampling operations of the backbone network, resulting in a serious shortage of effective information. Meanwhile, the traditional FPN adopts a pixel-wise addition fusion strategy, which lacks spatial context modeling around objects. Pixel shifts caused by recursive upsampling further aggravate feature misalignment of tiny objects, eventually leading to missed and false detections of unknown and tiny objects.

Therefore, we integrate two internal sub-branches within the HSE module, namely the High-Frequency Perception (HFP) sub-branch and the Spatial Dependency Perception (SDP) sub-branch, optimize the cooperative mechanism of frequency-domain filtering and spatial attention, and construct a frequency–space dual-branch multi-scale feature enhancement architecture. This module uses the frequency-domain high-frequency perception branch to suppress low-frequency background components via a learnable high-pass filter, highlighting edge and detail features of tiny objects. The spatial dependency perception branch captures spatial dependencies between upper and lower features via pixel-wise cross-attention, supplementing contextual semantic information and correcting feature misalignment. Finally, dual-branch feature fusion significantly strengthens the feature representation of tiny and unknown objects without introducing excessive computational overhead.

We take the multi-scale features output by the backbone network $\mathbf{X}_l \in \mathbb{R}^{C \times H_l \times W_l}$ ($l = 2, 3, 4, 5$, which denotes the feature level, C the number of channels, H_l and W_l the height and width of the feature at level l) as the module input. The detailed operations are as follows:

(1) Frequency-Domain High-Frequency Perception Branch: To highlight detailed features of tiny objects, we perform Discrete Cosine Transform (DCT) on the input features. A tunable high-pass filter removes low-frequency uniform background components concentrated in the top-left corner of the spectrum, preserving high-frequency components corresponding to object edges and details. An inverse DCT is then applied to obtain high-frequency response features. The kernel function of the high-pass filter is defined as:

$$\mathcal{Z}(p, q) = \begin{cases} 0, & p < \beta H_l, q < \beta W_l \\ 1, & \text{otherwise} \end{cases}, \quad \beta \in [0, 1] \quad (1)$$

where $\mathcal{Z}(p, q)$ is the filtering coefficient at position (p, q) in the DCT spectrum, and β is a hyperparameter controlling the filtering range (set to $\beta = 0.25$ by default in this paper). When $\beta = 0$, the full frequency band passes through, and the high-frequency response is identical to the input feature. When $\beta = 1$, the full frequency band is blocked. Only when $\beta \in (0, 1)$ does the filter take effect, removing low-frequency background components and generating the high-frequency response feature $\mathbf{X}_{\text{high},l} \in \mathbb{R}^{C \times H_l \times W_l}$. Then, based on the high-frequency response features, we generate attention weights along the channel and spatial dimensions to weight and enhance the original features. Along the channel dimension, global average pooling and global max pooling are applied to the high-frequency response features. After integrating spatial information, grouped convolution generates the channel attention weight $\mathbf{G}_{\text{ch},l} \in \mathbb{R}^{C \times 1 \times 1}$. Along the spatial dimension, 1×1 convolution aggregates channel information of the high-frequency response to generate the spatial attention mask $\mathbf{G}_{\text{sp},l} \in \mathbb{R}^{1 \times H_l \times W_l}$. Finally, dual-branch weighting and fusion yield the frequency-domain enhanced feature:

$$\mathbf{X}_{\text{freq},l} = \text{Conv}_{3 \times 3}(\mathbf{X}_l \odot \mathbf{G}_{\text{ch},l} + \mathbf{X}_l \odot \mathbf{G}_{\text{sp},l}) \quad (2)$$

where $\odot$ denotes the Hadamard product (element-wise multiplication), $\text{Conv}_{3 \times 3}$ denotes a 3×3 convolution operation, and $\mathbf{X}_{\text{freq},l} \in \mathbb{R}^{C \times H_l \times W_l}$ is the frequency-domain enhanced feature. This operation directionally enhances detailed features of tiny objects while suppressing invalid responses of uniform backgrounds, preserving the semantic information of the original features.

(2) Spatial Dependency Perception Branch

To address pixel misalignment and context missing in cross-level feature fusion, we design a pixel-wise cross-attention module to capture spatial dependencies between adjacent feature levels. First, the upper-level feature $\mathbf{X}_{l+1}$ is upsampled to match the size of the current-level feature $\mathbf{X}_l$, yielding the upsampled feature $\mathbf{X}_{\text{up},l}$. Then, three independent 1×1 convolutions generate the query matrix $\mathbf{Q}$, key matrix $\mathbf{K}$, and value matrix $\mathbf{V}$. Here $\mathbf{Q}$ is derived from the current-level feature $\mathbf{X}_l$ and $\mathbf{K}, \mathbf{V}$ is derived from the upsampled feature $\mathbf{X}_{\text{up},l}$.

To adapt to arbitrary input sizes and reduce computational complexity, $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ are split into N_{block} non-overlapping patches. Each patch has the same resolution as the top-level feature of the pyramid. Pixel-wise cross-attention is computed independently for each patch. The attention weight matrix of the i -th patch is defined as:

$$\mathcal{A}_i = \text{Softmax}\left(\frac{\mathbf{Q}_i \cdot \mathbf{K}_i^T}{\sqrt{C}}\right) \quad (3)$$

where $\mathbf{Q}_i, \mathbf{K}_i$ are the query and key matrices of the i -th patch, and $\sqrt{C}$ is a scaling factor to avoid softmax gradient saturation caused by excessively large inner products. The value matrix $\mathbf{V}_i$ is weighted by $\mathcal{A}_i$ to obtain patches enriched with spatial dependency

information. All patches are spatially aggregated and added to the original input feature via a residual connection to obtain the spatially enhanced feature:

$$\mathbf{X}_{\text{spa},l} = \text{Aggregate}(\mathcal{A}_i \cdot \mathbf{V}_i) + \mathbf{X}_l \quad (4)$$

where $\text{Aggregate}(\cdot)$ denotes spatial patch aggregation, and

$$\mathbf{X}_{\text{spa},l} \in \mathbb{R}^{C \times H_l \times W_l}$$

is the spatially enhanced feature. This operation accurately captures local spatial dependencies around objects, corrects pixel misalignment in cross-level fusion, and supplements critical contextual semantic information for unknown and tiny objects. The frequency-domain enhanced feature and spatially enhanced feature are added element-wise to obtain the final multi-scale enhanced feature, providing a high-quality feature foundation for subsequent cross-modal alignment and detection:

$$\mathbf{X}_{\text{enh},l} = \mathbf{X}_{\text{freq},l} + \mathbf{X}_{\text{spa},l} \quad (5)$$

where $\mathbf{X}_{\text{enh},l} \in \mathbb{R}^{C \times H_l \times W_l}$ is the final output enhanced feature.

(4) Computational Complexity Analysis

Let the input feature be $\mathbf{X}_l \in \mathbb{R}^{C \times H_l \times W_l}$. The computational complexity of the conventional self-attention mechanism is $O(C(H_l W_l)^2)$. In this module, the complexity of the frequency branch is $O(CH_l W_l)$, and the complexity of the spatial dependency branch is:

$$O(N_{\text{block}} C(H_{\text{base}} W_{\text{base}})^2) \quad (6)$$

where $H_{\text{base}} W_{\text{base}} \ll H_l W_l$ is the patch size. The overall complexity is reduced by an order of magnitude compared with conventional self-attention, ensuring inference efficiency while enhancing features.

3.2. Reparameterized Adaptive Text–Visual Alignment Module

The high-frequency-aware multi-scale feature enhancement module resolves the low-level visual representation problem of objects in the open world and provides high-quality visual features for cross-modal alignment. However, traditional open-vocabulary detection methods still suffer from two core limitations: (1) They rely on complex cross-modal fusion structures for text–visual feature interaction, and inference overhead rises sharply with vocabulary size, making them unsuitable for real-time edge detection. (2) Fixed pre-trained text embeddings suffer from domain shift with detection tasks, failing to adapt to dynamically changing decision boundaries in the open world, which drastically reduces alignment accuracy for rare and novel categories.

Therefore, we integrate two core sub-components within the RTA module, namely Reparameterized Regional Text Alignment (RepRTA) and Adaptive Decision Learning (AdaDL), to design a three-stage text–visual alignment pipeline: text embedding reparameterization calibration, adaptive decision boundary learning, and end-to-end reparameterized fusion.

During training, this module uses a lightweight auxiliary network to perform task-adaptive calibration on pre-trained text embeddings and optimizes the text encoder via low-rank adaptation to adapt decision boundaries to open-world detection tasks. After training, the optimized parameters of the calibration network and text encoder are fully merged into the visual classification head via reparameterization. No text encoding or cross-modal fusion is required during inference, enabling text–visual alignment with zero extra overhead.

It is worth noting that our RTA module, while sharing the concepts of low-rank adaptation and structural reparameterization with prior works such as RepYOLOE and YOLO-World, differs fundamentally in several key aspects. First, unlike RepYOLOE, which uses fixed CLIP embeddings for reparameterization, our approach employs a two-stage calibration pipeline consisting of a SwiGLU-based feed-forward network for domain-adaptive embedding transformation followed by LoRA-adapted text encoder outputs that dynamically learn task-specific decision boundaries. Second, we introduce a novel regional contrastive loss (Equation (10)) that operates on anchor-level visual features and calibrated text embeddings, providing finer-grained alignment than the image-level or box-level losses used in existing methods. Third, our reparameterized fusion explicitly leverages the task-optimized embeddings after both calibration and low-rank updates, whereas prior methods typically fuse pre-trained or lightly tuned embeddings without adaptive boundary learning. These design choices are mathematically formulated as a unified optimization framework (Equations (7)–(10)), and their effectiveness is validated by the ablation study in Table 2, where our combined strategy consistently outperforms plain LoRA, prompt tuning, and deep prompt tuning baselines, especially on rare categories.

Table 2. Structural and mechanism comparison with mainstream open-world detection methods.

Method	Small-Object Processing	Cross-Modal Alignment	Unknown Discovery	Incremental Mechanism	Inference Overhead
Grounding DINO	Standard FPN, no frequency enhancement	Transformer cross-attention	No dedicated module	Full fine-tuning, severe forgetting	High, scales with vocabulary
YOLO-Worldv2	Conventional FPN fusion	Vision-language path aggregation	No dedicated branch	Full fine-tuning	Medium, vocabulary-dependent
YOLO-UniOW	Standard multi-scale fusion	Standard feature alignment	Decoupled dual-branch	Contrastive learning, mild forgetting	Relatively high, dual-branch cost
YOLOE-v8	Anchor-based optimization	Reparameterized regional alignment	No dedicated recognition	No native support	Low, zero extra cost
TPCA-Net (Ours)	Frequency filtering + spatial dependency (HSE)	LoRA calibration + reparameterized fusion (RTA)	Double-wildcard spatial fusion (WSF)	Category-frozen, forgetting-free (DPD)	Low, zero cross-modal cost

The module takes the enhanced visual feature $\mathbf{X}_{\text{enh},l}$, and the category text set $\mathbf{T} = \{t_1, t_2, \dots, t_K\}$ (K is the number of categories) as inputs. The detailed operations are as follows:

(1) Pre-trained Text Embedding Extraction

We use the CLIP text encoder to encode input category texts and generate an initial text embedding matrix, preserving the universal semantic generalization ability of the pre-trained model:

$$\mathbf{E}_{\text{init}} = \text{CLIPTextEncoder}(\mathbf{T}) \quad (7)$$

where $\mathbf{E}_{\text{init}} \in \mathbb{R}^{K \times D}$ is the initial text embedding matrix, D is the embedding dimension, and $\text{CLIPTextEncoder}(\cdot)$ is the frozen CLIP text encoder.

(2) Text Embedding Reparameterization Calibration

To alleviate domain shift between pre-trained text embeddings and detection tasks, we design a lightweight auxiliary network, f_{θ} , consisting of only a single SwiGLU feed-forward network to perform task-adaptive calibration on the initial text embeddings while ensuring low computational overhead during training. The calibrated text embedding is computed as:

$$\mathbf{E}_{\text{cal}} = f_{\theta}(\mathbf{E}_{\text{init}}) \quad (8)$$

where $\mathbf{E}_{\text{cal}} \in \mathbb{R}^{K \times D}$ is the calibrated text embedding matrix. The trainable parameters of f_θ account for only 5% of those in traditional cross-modal fusion structures, greatly reducing training cost.

(3) Adaptive Decision Boundary Learning

To adapt the semantic space of text embeddings to the decision boundaries of open-world detection, we insert low-rank decomposition matrices into the fully connected layers of the CLIP text encoder. Without modifying pre-trained weights, we optimize the output of the text encoder using a small number of trainable parameters to achieve adaptive alignment between the text semantic space and the visual feature space. The updated output of the text encoder is:

$$\mathbf{Y} = \mathbf{W}_0 \mathbf{X} + \mathbf{U} \mathbf{V}^T \mathbf{X} \quad (9)$$

where $\mathbf{W}_0$ is the fixed pre-trained weight of the text encoder, $\mathbf{U} \in \mathbb{R}^{D \times r}$ and $\mathbf{V} \in \mathbb{R}^{D \times r}$ are low-rank decomposition matrices ($r \ll D$, $r = 16$ by default in this paper), $\mathbf{X}$ is the encoder input, and $\mathbf{Y}$ is the optimized encoder output.

This operation only updates the low-rank matrices $\mathbf{U}$ and $\mathbf{V}$ to adapt the text encoder output to the decision boundaries of the detection task, while fully preserving the pre-trained generalization ability of CLIP.

(4) End-to-End Reparameterized Fusion

To eliminate extra computational overhead during inference, after training, we fuse the calibrated text embeddings with the convolution kernels of the visual classification head via a tensor operation. Text semantics are fully integrated into classification weights through reparameterization, making the inference phase identical to a standard YOLO convolution layer.

The reparameterized classification convolution kernel is computed as:

$$\mathbf{W}'_{\text{cls}} = \text{Conv}_{1 \times 1}(\mathbf{W}_{\text{cls}} \otimes \mathbf{E}_{\text{cal}}) \quad (10)$$

where $\mathbf{W}_{\text{cls}} \in \mathbb{R}^{D \times C \times 1 \times 1}$ is the original 1×1 convolution kernel of the classification head, $\mathbf{E}_{\text{cal}} \in \mathbb{R}^{K \times D}$ is the calibrated text embedding matrix, and $\mathbf{W}'_{\text{cls}} \in \mathbb{R}^{K \times C \times 1 \times 1}$ is the reparameterized classification kernel. The operator $\otimes$ denotes a two-step operation: first, the spatial dimensions of $\mathbf{W}_{\text{cls}}$ are collapsed to treat it as a matrix of size $D \times C$; second, this matrix is multiplied by $\mathbf{E}_{\text{cal}}$ (size $K \times D$) via standard matrix multiplication, yielding an intermediate matrix of size $K \times C$. This intermediate is then reshaped back to a 4D tensor of shape $(K, C, 1, 1)$ to serve as the new classification kernel. This fusion is performed only once after training, so the inference stage involves only a standard convolution with $\mathbf{W}'_{\text{cls}}$, incurring zero additional computational costs compared with a conventional YOLO detection head. For clarity, the dimensions of all involved variables are summarized in Table 3 below.

Table 3. Dimensions of variables involved in the reparameterized tensor fusion (Equation (9)).

Symbol	Dimensions	Description
$\mathbf{W}_{\text{cls}}$	$D \times C \times 1 \times 1$	classification convolution kernel
$\mathbf{E}_{\text{cal}}$	$K \times D$	Calibrated text embedding matrix
$\mathbf{W}'_{\text{cls}}$	$K \times C \times 1 \times 1$	Reparameterized kernel
$\otimes$	–	$\mathbf{E}_{\text{cal}} \cdot \text{reshape}(\mathbf{W}_{\text{cls}}, (D, C))$, followed by reshaping

Finally, the convolution between visual features and the classification kernel completes feature extraction and text–visual alignment simultaneously. No text encoding or cross-modal fusion is needed during inference, achieving end-to-end alignment with zero extra overhead.

(5) Regional Contrastive Loss Function

To further improve text–visual alignment accuracy, we design a regional contrastive loss function that maximizes the similarity between visual features of positive sample regions and their corresponding category text embeddings, and minimizes the similarity for negative sample regions:

$$\mathcal{L}_{\text{align}} = -\frac{1}{N_{\text{pos}}} \sum_{i \in \text{pos}} \log \frac{\exp(\text{Sim}(\mathbf{X}_{\text{enh},i}, \mathbf{E}_{\text{cal},i})/\tau)}{\sum_{j=1}^K \exp(\text{Sim}(\mathbf{X}_{\text{enh},i}, \mathbf{E}_{\text{cal},j})/\tau)} \quad (10)$$

where N_{pos} is the number of positive sample anchors, $\text{Sim}(\cdot)$ denotes cosine similarity, τ is the temperature coefficient ($\tau = 0.07$ by default in this paper), $\mathbf{X}_{\text{enh},i}$ is the visual feature of the i -th positive anchor, and $\mathbf{E}_{\text{cal},i}$ is the calibrated text embedding of the corresponding category.

This loss function guides fine-grained regional alignment between visual features and text embeddings, significantly improving detection accuracy for rare and novel categories.

3.3. Double Wildcard Spatial Dependency Fusion Module

The reparameterized adaptive text–visual alignment module achieves lightweight and accurate alignment between text and visual features, providing a reliable semantic foundation for open-vocabulary detection. However, the core pain points of open-world detection remain unsolved. Unknown objects lack predefined text embeddings, so traditional open-vocabulary methods cannot recognize them effectively and tend to misclassify them as background. Features of unknown objects are highly confused with background and known-category features, and the lack of spatial context constraints leads to extremely low localization accuracy.

Therefore, we integrate the core designs of wildcard learning, spatial dependency perception, and semantic activation decoupling branches, and propose the double wildcard spatial dependency fusion module. This module designs a dual-embedding structure consisting of a general object wildcard and an unknown category wildcard, and realizes unsupervised recognition of unknown objects through self-supervised learning without the annotation information of unknown categories. Meanwhile, a decoupled semantic–spatial dual branch is introduced to supplement spatial context information for wildcard features and improve the localization accuracy of unknown objects. Finally, through three-branch feature fusion, it simultaneously achieves known-category feature enhancement, accurate detection of unknown objects, and dynamic category adaptation, enabling novel category expansion without incremental retraining.

It is important to distinguish our double-wildcard self-supervised learning from prior prototype-based unknown discovery methods such as UniOWL (shown in Table 4). While UniOWL employs a single learnable prototype for unknown classes and relies on contrastive loss to separate unknown from known, our approach introduces two key innovations. First, we adopt a two-stage curriculum learning strategy: the general-object wildcard is first trained to capture universal objectness from all labeled instances, and then the unknown wildcard is self-supervised using a screening function that jointly considers spatial overlap (IoU) and semantic similarity with known text embeddings. This dual-threshold screening (Equation (12)) is not merely a heuristic adjustment; it is mathematically formulated to generate reliable pseudo-labels for unknown objects without any human annotation, and the resulting self-supervised loss (Equation (13)) is specifically tailored to align the unknown wildcard with the general-object wildcard in the feature space. Second, unlike prototype-based methods that only perform semantic classification, our WSF module further incorporates a decoupled semantic–spatial fusion branch (Equations (14) and (15)) that generates spatial attention weights to enhance object localization. This spatial dependency modeling is absent in prior works and is critical for

accurately locating unknown objects in cluttered traffic scenes. As shown in our ablation studies, the spatial branch contributes notably to the detection of small and unknown objects, confirming that our self-supervised paradigm addresses both discrimination and localization simultaneously. A comparative summary of unknown discovery mechanisms is provided in Table 5 for quick reference.

The module takes the enhanced visual feature $\mathbf{X}_{\text{enh},l}$ and calibrated text embedding $\mathbf{E}_{\text{cal}}$ as inputs. The detailed operations are as follows:

(1) Double Wildcard Embedding Self-Supervised Learning

We design two complementary wildcard embeddings: general object wildcard $\mathbf{E}_{\text{obj}}$ and unknown category wildcard $\mathbf{E}_{\text{unk}}$. Both are initialized by the CLIP embedding of the generic text object and optimized through staged self-supervised learning.

Table 4. Comparison of unknown object discovery mechanisms in mainstream open-world detection methods.

Method	Wildcard/Prototype Design	Self-Supervised Paradigm	Spatial Localization Enhancement
UniOWL	Single learnable prototype for all unknowns	Contrastive learning with known-unknown separation	None (classification only)
YOLO-World	No dedicated unknown branch	Uses open-vocabulary vocabulary, treats unknown as background	None
YOLOE	No dedicated unknown branch	Relies on reparameterized text alignment, no unknown modeling	None
Ours (WSF)	Two complementary wildcards (object + unknown)	Two-stage curriculum learning: objectness first, then unknown via dual-threshold screening	Decoupled semantic-spatial fusion with pixel-wise attention

Table 5. Ablation results with different calibration strategies (%).

Calibration	Repara.	AP	APr	APf
Frozen	×	52.8	18.6	58.2
Prompt Tuning	×	53.3	20.5	58.5
Deep Prompt Tuning	×	53.4	21.2	58.6
Ours (LoRA)	×	53.7	22.8	58.7
Ours (LoRA)	✓	53.7	22.8	58.7

Stage 1: Optimize the general object wildcard $\mathbf{E}_{\text{obj}}$. We treat all annotated objects in the training set as the same object category, and we optimize the wildcard embedding via binary cross-entropy loss to guide it in learning the common visual features of all objects, regardless of whether the category is known:

$$\mathcal{L}_{\text{obj}} = \text{BCE}(\text{Sim}(\mathbf{X}_{\text{enh},l}, \mathbf{E}_{\text{obj}}), \mathbf{Y}_{\text{obj}}) \quad (11)$$

where $\mathbf{Y}_{\text{obj}}$ is the binary label of the object region (1 for object regions, 0 for background), and $\text{BCE}(\cdot)$ denotes the binary cross-entropy loss function. The optimized $\mathbf{E}_{\text{obj}}$ can accurately distinguish foreground objects from background in the image, providing a foundation for unknown object recognition. We use the optimized general object wildcard as supervision and train the unknown wildcard via self-supervised learning to focus on unknown objects outside the vocabulary. First, a screening function generates pseudo-labels by filtering out predictions with IoU greater than threshold $\sigma_1 = 0.5$ with known-category bounding boxes and similarity greater than threshold $\sigma_2 = 0.1$ with known-category text embeddings.

The remaining high-confidence, non-known-category predictions are used as training pseudo-labels for the unknown wildcard. The screening function is defined as:

$$\Psi(u, s) = \begin{cases} 1, & (u < \sigma_1) \wedge (s < \sigma_2) \wedge (\text{Sim}(X_{\text{enh},i}, E_{\text{obj}}) > \sigma_3) \\ 0, & \text{otherwise} \end{cases} \quad (12)$$

where u is the maximum IoU between the predicted box and known-category ground-truth boxes, s is the maximum similarity between the prediction and known-category text embeddings. The three thresholds are set as follows: $\sigma_1 = 0.5$ is the IoU threshold for filtering known-category overlapping samples, $\sigma_2 = 0.1$ is the known-category similarity threshold, and $\sigma_3 = 0.01$ is the general object similarity threshold. When $\Psi(u, s) = 1$, the prediction is used as a training pseudo-label for the unknown wildcard. When $\Psi(u, s) = 1$, the prediction is used as a training pseudo-label for the unknown wildcard, optimized via mean squared error loss:

$$\mathcal{L}_{\text{unk}} = \Psi(u, s) \cdot \text{MSE}(\text{Sim}(X_{\text{enh},i}, E_{\text{unk}}), \text{Sim}(X_{\text{enh},i}, E_{\text{obj}})) \quad (13)$$

where $\text{MSE}(\cdot)$ denotes the mean squared error loss function. The optimized E_{unk} can accurately identify unknown objects outside the vocabulary and effectively distinguish them from known-category features.

(2) Decoupled Semantic–Spatial Dual-Branch Fusion

To address the low localization accuracy of unknown objects, we design a decoupled semantic branch and spatial branch to fuse and enhance the three-branch features of known categories, general objects, and unknown objects.

The semantic branch extracts global semantic information of the three-branch features via 3×3 convolution to generate category-agnostic object semantic features.

The spatial branch generates spatial attention weights for each position of the feature map via pixel-wise cross-attention, focusing on object regions and suppressing background interference.

The spatial attention weight is computed as:

$$\mathbf{M}_{\text{spa}} = \text{Sigmoid} \left(\text{Conv}_{1 \times 1} \left(\text{Concat} \left(X_{\text{enh},l} \cdot E_{\text{cal}}^T, X_{\text{enh},l} \cdot E_{\text{obj}}^T, X_{\text{enh},l} \cdot E_{\text{unk}}^T \right) \right) \right) \quad (14)$$

where $\mathbf{M}_{\text{spa}} \in \mathbb{R}^{1 \times H_l \times W_l}$ is the spatial attention weight matrix, and $\text{Sigmoid}(\cdot)$ denotes the sigmoid activation function.

Finally, the fused features are weighted and enhanced by the spatial attention weights to obtain the final fused detection features:

$$X_{\text{fuse},l} = \text{Conv}_{3 \times 3} \left(\text{Concat} \left(X_{\text{enh},l} \cdot E_{\text{cal}}^T, X_{\text{enh},l} \cdot E_{\text{obj}}^T, X_{\text{enh},l} \cdot E_{\text{unk}}^T \right) \right) \odot \mathbf{M}_{\text{spa}} \quad (15)$$

where $X_{\text{fuse},l} \in \mathbb{R}^{(K+2) \times H_l \times W_l}$ is the final fused detection feature. The first K channels correspond to known categories, and the last two channels correspond to general objects and unknown objects, respectively. This feature contains accurate semantics of known categories, common features of unknown objects, and spatial position information of objects, providing comprehensive feature support for subsequent detection tasks.

3.4. Incremental Forgetting-Free Dual-Path Detection Module

The double wildcard spatial dependency fusion module achieves unified feature representation for known categories and unknown objects, providing a multi-dimensional feature foundation for detection tasks. However, traditional detection heads still suffer from three major drawbacks:

1. The detection head uses a unified optimization objective for known categories and unknown objects, leading to mutually constrained performance that cannot be simultaneously optimized.
2. It relies on non-maximum suppression (NMS) post-processing, which not only increases inference overhead but also tends to filter out tiny unknown objects.
3. When adding new categories, the entire detection head requires incremental re-training, which is prone to catastrophic forgetting and drastically degrades the detection accuracy of old categories.

Therefore, we integrate the dual-head matching strategy of YOLOv10 and the idea of Lazy Region-Prompt Contrast (LRPC) to design the incremental forgetting-free dual-path detection module. This module decouples the detection head into a known-category detection path and an unknown-category discovery path. The two paths share low-level fused features but adopt independent optimization objectives and network structures: the known path focuses on accurate detection and localization of known categories, while the unknown path focuses on discovery and incremental expansion of unknown objects. Meanwhile, the lazy region contrast strategy only performs category matching on candidate object anchors, greatly reducing inference overhead. Finally, the category-frozen incremental learning mechanism enables forgetting-free expansion of new categories without retraining parameters related to old categories.

The module takes the fused detection feature $\mathbf{X}_{\text{fuse},l}$ as input. The detailed operations are as follows:

(1) Dual-Path Decoupled Detection Head Design

We decouple the detection head into two independent paths responsible for known-category detection and unknown-category discovery, respectively. The two paths share low-level fused features but adopt independent head structures and optimization objectives to avoid mutual performance constraints.

The known-category detection path uses a one-to-one matching detection head, assigning only one positive anchor to each known target. It focuses on precise bounding-box regression and fine-grained classification of known categories, outputting duplicate-free detection results without NMS post-processing. The loss function of this path consists of classification loss and regression loss:

$$\mathcal{L}_{\text{known}} = \mathcal{L}_{\text{cls}}(\mathbf{s}_{\text{known}}, \mathbf{Y}_{\text{known}}) + \lambda_{\text{reg}} \mathcal{L}_{\text{reg}}(\mathbf{B}_{\text{known}}, \mathbf{Y}_{\text{bbox}}) \quad (16)$$

where $\mathbf{s}_{\text{known}}$ is the predicted score of known categories, $\mathbf{Y}_{\text{known}}$ is the ground-truth label of known categories, $\mathcal{L}_{\text{cls}}$ is binary cross-entropy loss; $\mathbf{B}_{\text{known}}$ is the predicted bounding box, $\mathbf{Y}_{\text{bbox}}$ is the ground-truth bounding box, $\mathcal{L}_{\text{reg}}$ is CIOU loss, and $\lambda_{\text{reg}} = 2.0$ is the balance coefficient.

The unknown-category discovery path uses a one-to-many matching detection head, assigning multiple positive anchors to each target. It focuses on foreground-background discrimination and localization of unknown objects, and realizes discovery and matching of new categories through a built-in large vocabulary. This path first screens candidate object anchors using the score of the general object wildcard, and only performs lazy matching of

the built-in vocabulary on the filtered anchors to avoid invalid computation on background anchors. The screening process is:

$$\Omega_{\text{sel}} = \{o \in H_l W_l \mid \text{Sim}(\mathbf{X}_{\text{fuse},l}(o,:), \mathbf{E}_{\text{obj}}) > \delta\} \quad (17)$$

where Ω_{sel} is the set of screened candidate object anchors, and $\delta = 0.001$ is the screening threshold. This step retains only about 20% high-confidence anchors, greatly reducing the computational cost of subsequent vocabulary matching. Then, only the filtered anchors are used to compute similarity with text embeddings of the built-in large vocabulary to complete matching and recognition of new categories. Unmatched anchors are judged as unknown objects, providing candidates for subsequent incremental expansion.

(2) Category-Frozen Incremental Forgetting-Free Learning

To realize dynamic expansion of new categories while avoiding catastrophic forgetting, we design a category-frozen incremental learning mechanism. When new categories are added, only text embeddings of new categories are calibrated and optimized, and only convolution weights corresponding to new categories in the detection head are updated. All parameters related to old categories are completely frozen, including old-category text embeddings, classification head weights, and backbone parameters.

The optimization loss function for new categories only includes classification loss and regression loss of new categories, without any update of old-category parameters:

$$\mathcal{L}_{\text{new}} = \mathcal{L}_{\text{cls}}(\mathbf{s}_{\text{new}}, \mathbf{Y}_{\text{new}}) + \lambda_{\text{reg}} \mathcal{L}_{\text{reg}}(\mathbf{B}_{\text{new}}, \mathbf{Y}_{\text{new_bbox}}) \quad (18)$$

where $\mathbf{s}_{\text{new}}$ is the predicted score of new categories, $\mathbf{Y}_{\text{new}}$ is the ground-truth label of new categories, $\mathbf{B}_{\text{new}}$ is the predicted bounding box of new categories, and $\mathbf{Y}_{\text{new_bbox}}$ is the ground-truth bounding box of new categories. By freezing old-category parameters, this mechanism fundamentally avoids catastrophic forgetting in incremental learning. Meanwhile, category expansion can be completed with only a small amount of annotation data without retraining the entire model, greatly reducing the cost of incremental expansion.

(3) End-to-End Joint Loss Function

The end-to-end joint loss function of the entire detection module consists of three parts: known-category detection loss, unknown-category self-supervised loss, and new-category incremental loss, realizing joint optimization of known categories, unknown objects, and new categories:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{known}} + \Psi(u, s) \cdot \mathcal{L}_{\text{unk}} + \mathcal{L}_{\text{new}} \quad (19)$$

With this joint loss function, the model can simultaneously achieve accurate detection of known categories, effective recognition of unknown objects, and forgetting-free incremental expansion of new categories, perfectly adapting to dynamically changing open-world detection scenarios.

As shown in Table 2, existing open-world detection methods have distinct technical trade-offs across dimensions. Transformer-based methods like Grounding DINO deliver strong zero-shot generalization but suffer from high inference latency and insufficient optimization for tiny objects. YOLO-series methods balance speed and accuracy, but generally lack dedicated unknown object perception and effective forgetting-free incremental expansion. The dual-branch YOLO-UniOW supports unknown discovery, yet introduces extra inference overhead and still has limitations in tiny-object feature enhancement.

In comparison, TPCA-Net addresses the above limitations systematically via four collaborative modules. The HSE module compensates for the common lack of frequency-domain detail enhancement for small objects. The RTA module achieves higher cross-modal alignment accuracy without increasing inference cost, combining low-rank calibration

and reparameterized fusion. The WSF module realizes unsupervised unknown object recognition with improved localization precision via spatial dependency constraints. The DPD module fundamentally avoids catastrophic forgetting while maintaining efficient NMS-free inference. This systematic design enables TPCA-Net to gain comprehensive advantages, which is further verified by subsequent experimental results.

4. Analysis of Error and Convergence

In this section, we derive the strict error upper bound of tiny object feature recovery for the High-Frequency-Aware Multi-Scale Feature Enhancement (HSE) module, based on the energy compaction property of discrete cosine transform (DCT) and the spatial constraint property of patch-wise cross-attention. For the l -th layer feature map $X_l \in \mathbb{R}^{C \times H_l \times W_l}$, its 2D DCT transform is denoted as $\mathcal{F}(X_l) = F_l \in \mathbb{R}^{C \times H_l \times W_l}$. By Parseval's identity, DCT preserves signal energy: $\|X_l\|_F^2 = \|F_l\|_F^2$. After applying the high-pass filter kernel $\mathcal{Z}(p, q)$ defined in Equation (1), the filtered frequency-domain feature is $F'_l = F_l \odot \mathcal{Z}$, and the reconstructed high-frequency feature is $X_{\text{high},l} = \mathcal{F}^{-1}(F'_l)$. The relative reconstruction error of the frequency branch is defined as $e_{\text{freq}} = \|X_l - X_{\text{high},l}\|_F / \|X_l\|_F$. Let E_{low} denote the energy ratio of the low-frequency region within the filter cutoff range. For convolutional features with natural image statistics, DCT coefficients follow an exponential decay distribution, so $E_{\text{low}} \leq 1 - \alpha\beta^2$ where $\alpha \in (0, 1)$ is the energy decay coefficient. The reconstruction error of the frequency branch satisfies:

$$e_{\text{freq}}^2 = \frac{\|F_l - F'_l\|_F^2}{\|F_l\|_F^2} = E_{\text{low}} \leq 1 - \alpha\beta^2 \quad (20)$$

For the spatial dependency branch, the patch-wise cross-attention mechanism compensates contextual information for high-frequency features. Based on the normalization property of attention weights, the spatial compensation error $e_{\text{spa}} = \|X_{\text{spa},l} - X_l^*\|_F / \|X_l\|_F$ (where X_l^* is the ideal context-enhanced feature) satisfies $e_{\text{spa}} \leq \gamma \cdot e_{\text{freq}}$, with $\gamma \in (0, 1)$ being the spatial compensation coefficient determined by patch size and receptive field. According to the triangle inequality, the total feature recovery error of the HSE module $X_{\text{enh},l} = X_{\text{freq},l} + X_{\text{spa},l}$ satisfies: $e_{\text{enh}} \leq e_{\text{freq}} + e_{\text{spa}} \leq (1 + \gamma)\sqrt{1 - \alpha\beta^2}$.

This derivation proves that the dual-branch fusion strategy reduces the feature recovery error upper bound compared with the single frequency-domain branch, and provides a strict mathematical guarantee for the tiny object feature enhancement effect instead of empirical design. For the Double Wildcard Spatial Dependency Fusion module, we have derived the generalization error upper bound of the dual wildcard embedding space based on Rademacher complexity theory, considering the pseudo-label noise in self-supervised learning. Let the hypothesis space of wildcard embedding be $\mathcal{H} = \{h(x) = \text{Sim}(x, e) \mid e \in \mathbb{R}^D, \|e\|_2 = 1\}$, where $\text{Sim}(\cdot)$ denotes cosine similarity, and D is the embedding dimension. For a training set, $S = \{(x_i, y_i)\}_{i=1}^n$, with pseudo-label noise rate η , the empirical risk is $\hat{R}(h) = \frac{1}{n} \sum_{i=1}^n \ell(h(x_i), y_i)$, and the expected risk is $R(h) = \mathbb{E}_{(x,y) \sim \mathcal{D}}[\ell(h(x), y)]$. The Rademacher complexity of the hypothesis space satisfies:

$$\mathfrak{R}_n(\mathcal{H}) \leq \sqrt{\frac{2D}{n}} \quad (21)$$

The risk bias caused by pseudo-label noise is bounded by ηL_{max} , where L_{max} is the maximum value of the MSE loss function. With probability of at least $1 - \delta$, the generalization error bound holds: $R(h) \leq \hat{R}(h) + 2\mathfrak{R}_n(\mathcal{H}) + \eta L_{\text{max}} + \sqrt{\frac{\log(1/\delta)}{2n}}$. We further

derive the quantitative relationship between the noise rate η and the three pseudo-label screening thresholds $\sigma_1, \sigma_2, \sigma_3$ defined in Equation (12):

$$(\eta \leq 1 - \frac{\sigma_2 \cdot \sigma_3}{1 - \sigma_1}) \quad (22)$$

Substituting this inequality into the generalization error bound gives the explicit upper bound for given threshold settings, and also derives the feasible region of thresholds that satisfies the target generalization error requirement. This theoretical analysis proves that the dual wildcard learning mechanism maintains stable recognition performance for unseen unknown objects, rather than only fitting the training set distribution. For the overall end-to-end optimization objective, we have completed the convergence analysis of the joint loss function and given the explicit convergence bound under the AdamW optimizer with cosine annealing schedule. The total loss function is $\mathcal{L}_{\text{total}}(\theta) = \mathcal{L}_{\text{known}}(\theta) + \Psi \cdot \mathcal{L}_{\text{unk}}(\theta) + \mathcal{L}_{\text{new}}(\theta)$, where θ denotes all trainable network parameters. We first prove that each sub-loss function satisfies the Lipschitz-continuous gradient condition under the bounded parameter constraint $\|\theta\|_2 \leq \Theta$. For binary cross-entropy classification loss and CIoU regression loss, there exist constants $L_1, L_2, L_3 > 0$, such that:

$$\|\nabla \mathcal{L}_i(\theta_1) - \nabla \mathcal{L}_i(\theta_2)\|_2 \leq L_i \|\theta_1 - \theta_2\|_2, \quad i = 1, 2, 3 \quad (23)$$

Thus, the total loss has a Lipschitz gradient constant, $L = L_1 + L_2 + L_3$, meaning the joint objective is also Lipschitz-smooth. Based on the convergence theory of AdamW optimizer, after T iterations with cosine annealing learning rate $\eta_t \leq \eta_{\max}$, the expected average gradient norm satisfies:

$$\left(\mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T \|\nabla \mathcal{L}_{\text{total}}(\theta_t)\|_2^2 \right] \leq \frac{L \Delta_f}{\sqrt{T}} + \frac{\eta_{\max} G^2}{2} \right) \quad (24)$$

where $\Delta_f = \mathcal{L}_{\text{total}}(\theta_0) - \mathcal{L}_{\text{total}}^*$ is the gap between initial loss and optimal loss, and G is the uniform upper bound of gradient norms. This inequality shows that the gradient norm converges to a neighborhood of zero as the iteration number T increases, which theoretically guarantees the stability of the multi-module joint training process and explains the convergence mechanism of collaborative optimization of multiple loss functions.

5. Experiments

In this section, we select representative baseline models covering all the mainstream technical routes reviewed in the Introduction, including closed-set basic detection methods, Transformer-based open-vocabulary methods, YOLO-series lightweight open-vocabulary methods, decoupled open-world methods and reparameterized cross-modal methods, to conduct comprehensive comparative experiments on COCO, DOTA and AI-TOD datasets. The selection of baselines strictly corresponds to the literature review in the Introduction, ensuring the comparability and persuasiveness of the experimental results.

5.1. Dataset Introduction

The COCO (Common Objects in Context) dataset [32] is the most widely used benchmark in general object detection. It is specifically designed for validating the universal performance and zero-shot generalization ability of open-world object detection. Its core characteristics lie in the high coverage of object categories and scene environments: it contains 80 common object categories, 123,287 images, and more than 1.5 million object instances in total, including 118,287 training images and 5000 validation images. It covers objects with various scales, poses, occlusion levels, and complex backgrounds.

The DOTA dataset [33] is an authoritative benchmark for oriented object detection in aerial remote sensing images. It is dedicated to verifying open-world detection performance in remote sensing traffic scenarios. Its unique value is the comprehensive coverage of multi-scale, multi-directional, and multi-scene traffic objects: it contains 2806 large high-resolution remote sensing images, covering 15 object categories including planes, bridges, large vehicles, small vehicles, ships, and harbors, with more than 180,000 annotated oriented bounding boxes. It covers complex traffic scenarios such as near-shore, open-sea, urban, and harbor areas under different resolutions and imaging conditions.

The AI-TOD dataset [33] is a benchmark designed specifically for tiny object detection in aerial traffic scenarios. It is used to validate detection performance under dense tiny traffic objects. Its core characteristics are an extremely high proportion of tiny objects and strong traffic-scene specificity: it contains 28,034 urban aerial images of size 800×800 pixels, covering 8 traffic object categories including vehicles, pedestrians, traffic signs, and engineering machinery, with more than 230,000 object instances. Among them, 84.7% of objects are smaller than 16×16 pixels, making it one of the traffic detection datasets with the highest small-object ratio.

5.2. Experimental Implementation Details

All experiments for open-world object detection in intelligent transportation are conducted on a platform with Ubuntu 22.04 64-bit OS. For hardware, we use four NVIDIA RTX 4090 GPUs as the main computing units, equipped with an Intel Xeon Gold 6330 CPU and 128 GB DDR4 3200 MHz RAM to ensure parallel computing efficiency and data throughput for large-scale dataset training. Inference speed is tested on an NVIDIA Tesla T4 GPU with TensorRT 8.6 to match real deployment environments of vehicle-mounted and roadside edge devices.

For software, the network is implemented and trained using Python 3.10 and the PyTorch 2.1.0 deep learning framework (with CUDA 11.8 and cuDNN 8.9). Text embeddings are extracted using the OpenCLIP pre-trained model. Image preprocessing and augmentation are implemented with OpenCV 4.8.0 and Albumentations 1.3.1.

Three standard datasets are adopted: COCO, DOTA, and AI-TOD. Unified preprocessing is applied according to image characteristics: COCO images are resized to 640×640 pixels. DOTA and AI-TOD images are cropped and resized to 800×800 pixels to fit the network input size.

Pixel normalization and channel standardization follow the statistics of the ImageNet dataset:

$$\mu = [0.485, 0.456, 0.406], \quad \sigma = [0.229, 0.224, 0.225]$$

to maintain consistency with the backbone pre-training pipeline.

To alleviate imbalanced sample distribution and insufficient tiny-object samples in open scenarios, extensive data augmentation is used during training, including: - Random horizontal/vertical flip (probability = 0.5) - Random rotation within $\pm 15^\circ$. - Random scaling from 0.8 to 1.2. - Random brightness and contrast adjustment. - Random copy-paste augmentation for tiny traffic objects.

This improves the model's robustness to scale and orientation variations and strengthens feature learning for tiny objects.

Network Initialization

The backbone is the original CSPDarknet of YOLOv11-L, which is fully pre-trained on the large-scale ImageNet dataset to initialize weights and enhance universal feature extraction.

The CLIP text encoder uses OpenCLIP-ViT-B/32 pre-trained weights. The main encoder weights are frozen during training; only the inserted low-rank matrices are updated to preserve the generalization ability of the pre-trained model.

Training Configuration

The training process uses refined hyperparameters for stable convergence: - Optimizer: AdamW. - Weight decay: 1×10^{-4} . - First-order momentum: $\beta_1 = 0.9$. - Second-order momentum: $\beta_2 = 0.999$. - Max gradient norm: 10.0 (to avoid gradient explosion).

The initial learning rate is set to $lr_{init} = 1 \times 10^{-4}$. A cosine annealing scheduler is adopted. The first 3 epochs use linear warmup to reach the target learning rate. At the end of training, the learning rate decays to $lr_{end} = 1 \times 10^{-6}$ for stable convergence.

The total training epochs are 100, and the batch size = 32. Validation is performed every 2 epochs on the validation set. Early stopping is enabled: training stops if mAP@0.5 shows no clear improvement for 10 consecutive epochs to prevent overfitting. An Exponential Moving Average (EMA) strategy is used during training with a momentum of 0.9998 to further improve robustness and generalization. A lightweight post-processing-free strategy is used, consistent with the network design: - The known-category detection path uses one-to-one label matching, assigning only one positive anchor per object, producing duplicate-free results without NMS, avoiding over-suppression of tiny objects. - The unknown-category discovery path filters candidate anchors using a general-object wildcard similarity threshold of 0.001; only high-confidence anchors proceed to vocabulary matching. - Final confidence threshold: 0.4 to suppress false background detections.

Evaluation Metrics

Standard detection metrics are used to comprehensively evaluate accuracy, generalization, and real-time performance: mAP@0.5: mean Average Precision at IoU = 0.5. - mAP@0.5:0.95: mean Average Precision over IoU thresholds from 0.5 to 0.95 at step 0.05. - APs: Average Precision for small objects. - AP_r: Average Precision for rare categories. - Precision, Recall. - FPS: frames per second for edge inference speed.

5.3. Ablation Study

A Ablation Study Analysis

We conduct systematic ablation experiments on the COCO, DOTA, and AI-TOD datasets to verify the performance gains brought by our four core modules: High-Frequency-Aware Multi-Scale Feature Enhancement (HSE), Reparameterized Adaptive Text–Visual Alignment (RTA), Double Wildcard Spatial Dependency Fusion (WSF), and Incremental Forgetting-Free Dual-Path Detection (DPD). We also explore the optimal parameter settings and the contribution of each sub-module.

Table 6 shows the zero-shot open-world detection results on the COCO dataset by gradually adding each core module based on the YOLOv11-L baseline.

The baseline model only achieves an AP of 52.8%. This is because traditional YOLOv11 suffers from three critical drawbacks in open-world scenarios:

1. Severe feature degradation of tiny and unknown objects after multi-layer downsampling, leading to loss of effective detail information.
2. Reliance on complex cross-modal fusion structures for text–visual alignment, resulting in high inference overhead, domain shift between pre-trained text embeddings and detection tasks, and insufficient alignment accuracy for rare categories.
3. An inability to effectively recognize unknown objects outside the vocabulary, which are easily misclassified as background, along with catastrophic forgetting during incremental category expansion.

When only the HSE module is added to the baseline, detection AP improves by 1.1%. This module highlights edge details of tiny objects via frequency-domain high-pass filtering

and supplements spatial context information through pixel-wise cross-attention, effectively solving feature degradation and misalignment of small and unknown objects in open scenarios, providing high-quality features for subsequent detection.

When the RTA module is further incorporated, the detector AP rises by another 0.9%. This module performs task-adaptive calibration of text embeddings via a lightweight auxiliary network and dynamically learns decision boundaries through low-rank adaptation. After training, text semantics are fully merged into the classification head via reparameterization, greatly improving fine-grained text–visual alignment accuracy with zero extra inference overhead and especially boosting detection of rare categories.

Adding only the WSF module increases AP by 0.7%. Through self-supervised learning of general-object and unknown-category double wildcards, this module enables accurate recognition of unannotated unknown objects. Meanwhile, the decoupled semantic–spatial dual branch supplements spatial context constraints, effectively reducing false and missed detections of unknown objects.

Introducing only the DPD module improves AP by 0.5%. The decoupled known/unknown dual-path detection head avoids conflicting optimization objectives between the two types of objects. The lazy region contrast strategy drastically reduces invalid computation, and the category-frozen incremental learning mechanism fundamentally eliminates catastrophic forgetting in incremental expansion.

When all four core modules are used together, our proposed TPCA-Net finally achieves an AP of 56.0% on the COCO dataset, an improvement of 3.2% over the baseline and up to 0.8% over the SOTA model in reference papers. This fully validates the effectiveness and complementarity of each module. The four modules form a complete open-world detection optimization pipeline from feature enhancement, cross-modal alignment, unknown object recognition, to incremental detection, achieving collaborative performance improvement.

Consistent experimental results are obtained on the DOTA and AI-TOD datasets, as shown in Tables 7 and 8. Compared with using individual modules, TPCA-Net achieves superior detection performance when modules are combined, with no performance conflicts; the joint use yields the largest accuracy gain.

On the DOTA dataset for remote sensing rotated object detection, the full-module TPCA-Net achieves 79.3% mAP, 3.9% higher than the baseline and 0.7% higher than the best SOTA model.

On the AI-TOD dataset for dense small-object detection, the full-module TPCA-Net reaches 40.5% AP, with a small-object AP_s of 28.7%, 4.3% higher than the baseline and up to 0.9% higher than the SOTA model, highlighting the core advantage of the HSE module in open-world tiny-object detection.

To further quantify the individual contributions and synergistic gains of the HFP and SDP sub-branches within the HSE module, we conduct an ablation study on the AI-TOD dataset where tiny-object characteristics are most prominent. The results are shown in Table 9. It can be observed that, when using only the high-frequency perception branch (HFP), the model AP is improved by 2.0%, and the small-object AP_s is improved by 3.1% compared with the baseline, verifying the enhancement effect of frequency-domain filtering on detailed features of tiny objects. When using only the spatial dependency perception branch (SDP), the model AP is improved by 1.7%, and the small-object AP_s is improved by 2.4% compared with the baseline, validating the optimization effect of spatial context modeling on feature misalignment and context missing.

When the two branches work together, both the model AP and AP_s achieve further improvements, increasing by 2.5% and 3.6% respectively compared with the baseline. This fully demonstrates that the two branches form a good complementarity: the frequency-

domain branch enhances detailed features, and the spatial branch supplements contextual information, jointly achieving high-quality representation of small-object features.

To verify the performance differences of various text encoder calibration strategies in the reparameterized adaptive text-visual alignment module, we conduct comparative ablation experiments on the COCO dataset. Following the ablation settings of adaptive decision learning in Ref. [2], we compare the effects of different calibration methods on the zero-shot detection accuracy of the model. It can be observed that the baseline with a frozen text encoder only achieves 52.8% AP. Although full fine-tuning of the text encoder slightly improves the overall AP, it causes a remarkable drop in the rare-category AP_r . This is because full fine-tuning tends to make the model overfit on high-frequency categories and lose the generalization ability of the pre-trained model. After introducing prompt tuning and deep prompt tuning, both the overall AP and the rare-category accuracy of the model are significantly improved. With the proposed LoRA full projection layer adaptation plus reparameterized fusion scheme, the model achieves the optimal AP of 53.7%, where the rare-category AP_r is boosted most significantly by 4.2% compared with the frozen baseline. Meanwhile, there is no extra computational overhead during inference, which fully demonstrates the dual advantages of the proposed strategy in text-visual alignment accuracy and inference efficiency.

Table 6. Ablation results of TPCA-Net on COCO dataset (%).

HSE	RTA	WSF	DPD	COCO (AP)
×	×	×	×	52.8
✓	×	×	×	53.9
×	✓	×	×	53.7
×	×	✓	×	53.5
×	×	×	✓	53.3
✓	✓	×	×	54.5
✓	✓	✓	×	55.1
✓	✓	✓	✓	56.0

Table 7. Ablation results of TPCA-Net on DOTA dataset (%).

HSE	RTA	WSF	DPD	DOTA (mAP)
×	×	×	×	75.4
✓	×	×	×	76.8
×	✓	×	×	76.5
×	×	✓	×	76.2
×	×	×	✓	75.9
✓	✓	×	×	77.6
✓	✓	✓	×	78.5
✓	✓	✓	✓	79.3

Table 8. Ablation results of TPCA-Net on AI-TOD dataset (%).

HSE	RTA	WSF	DPD	AI-TOD (AP)	AI-TOD (AP_s)
×	×	×	×	36.2	22.5
✓	×	×	×	38.7	26.1
×	✓	×	×	37.0	23.1
×	×	✓	×	36.9	22.9
×	×	×	✓	36.7	22.7
✓	✓	×	×	39.2	27.3
✓	✓	✓	×	39.9	28.1
✓	✓	✓	✓	40.5	28.7

Table 9. Ablation results of sub-branches in high-frequency perception multi-scale feature enhancement module (%).

With HFP	With SDP	AI-TOD (AP)	AI-TOD (AP _s)
×	×	36.2	22.5
✓	×	38.2	25.6
×	✓	37.9	24.9
✓	✓	38.7	26.1

As shown in Table 10, the simple independent stack of HFP and SDP branches only achieves an AP improvement of 2.1% and an AP_s improvement of 3.2% compared with the baseline, which is far lower than the sum of individual branch gains (2.0% + 1.7% for AP, 3.1% + 2.4% for AP_s). This phenomenon indicates that directly stacking two independently designed modules will cause feature redundancy and information conflict, resulting in obvious performance attenuation. In contrast, our proposed joint modeling HSE module achieves an AP improvement of 2.5% and an AP_s improvement of 3.6% under exactly the same parameter scale and computation cost, with an additional 0.4% accuracy gain compared with the simple stack strategy. This extra performance benefit comes from the patch-aligned computing mechanism and the orthogonal complementary fusion design, which effectively eliminates feature redundancy and realizes the collaborative enhancement of frequency-domain details and spatial context. The experimental results fully verify that the HSE module is not a simple stack of classic modules but an innovative mathematical modeling framework with targeted optimization for tiny objects in open-world detection scenarios.

Table 10. Ablation results of different fusion strategies on the AI-TOD dataset (%).

Module Configuration	AP	AP _s	Parameters (M)	GFLOPs
Baseline (YOLOv11-L)	36.2	22.5	25.3	52.1
+HFP branch only	38.2	25.6	25.7	52.8
+SDP branch only	37.9	24.9	26.1	53.5
+Simple independent stack of HFP + SDP	38.3	25.7	26.5	54.2
+HSE (Ours, joint modeling)	38.7	26.1	26.5	54.2

As shown in Table 11, the measured FLOPs of the patch-based cross-attention are highly consistent with the theoretical derivation results. On all three datasets, the computational cost of the proposed attention mechanism is reduced by more than 15 times compared with vanilla global self-attention, which basically reaches one order of magnitude reduction. This experimental result directly supports the conclusion of complexity optimization and verifies that the patch-based design effectively reduces the computational overhead of spatial dependency modeling while preserving contextual information enhancement capability.

Table 11. FLOPs comparison of different attention mechanisms on the P3 feature layer (M FLOPs).

Attention Mechanism	COCO (640 × 640)	DOTA (800 × 800)	AI-TOD (800 × 800)
Vanilla global self-attention	1256.8	3062.4	3062.4
Patch-based cross-attention (Ours)	80.3	195.7	195.7
Reduction ratio	15.6×	15.6×	15.6×

Table 12 presents the overall complexity and inference performance of different module configurations. Compared with the baseline, the HSE, RTA and WSF modules introduce

only 4.1%, 1.1% and 0.9% additional FLOPs respectively, achieving significant performance improvement with extremely low computational cost. Benefiting from the NMS-free design of the DPD module, the full TPCA-Net achieves even higher inference speed than the baseline despite slightly increased FLOPs, reaching 101.2 FPS on COCO and over 79 FPS on 800×800 input. These results prove that the proposed method maintains excellent real-time performance while improving detection accuracy, which is suitable for edge deployment scenarios in intelligent transportation and remote sensing.

Table 12. Overall complexity and inference speed of different model configurations.

Configuration	Params (M)	COCO (640 × 640)		DOTA (800 × 800)		AI-TOD (800 × 800)	
		FLOPs (G)	FPS	FLOPs (G)	FPS	FLOPs (G)	FPS
Baseline (YOLOv11-L)	25.3	52.1	95.6	81.4	78.3	81.4	78.1
+HSE module	26.5	54.2	92.7	84.7	75.8	84.7	75.5
+HSE + RTA	26.9	54.8	91.5	85.6	74.9	85.6	74.6
+HSE + RTA + WSF	27.2	55.3	90.8	86.2	74.2	86.2	73.9
TPCA-Net (Full)	27.5	55.6	101.2	86.7	79.6	86.7	98.6

6. Comparative Experiments

6.1. Experimental Results and Analysis on the COCO Dataset

Table 13 shows the zero-shot detection results of our proposed TPCA-Net and other mainstream open-world/open-vocabulary object detection methods on the COCO dataset. In terms of overall detection performance, TPCA-Net achieves the best result among all compared methods with an AP of 56.0%, where the rare-category AP_r reaches 23.5%, the small-object AP_s reaches 38.2%, and the frequent-category AP_f reaches 59.1%. Meanwhile, it delivers an inference speed of 101.2 FPS on a T4 GPU. It can be seen from the table that, except for TPCA-Net, YOLOE-v8-L performs best in the open-vocabulary detection task. This is because it achieves lightweight cross-modal feature alignment through a reparameterized regional text alignment strategy, balancing detection accuracy and inference efficiency, and exhibits strong category generalization ability in open scenarios. However, YOLOE-v8-L ignores the feature degradation problem of tiny objects in the open world, lacks effective recognition and localization mechanisms for unknown objects, and is prone to catastrophic forgetting during incremental category expansion. TPCA-Net exactly remedies these defects, improving the overall AP by 0.7%, the small-object AP_s by 2.1%, and the rare-category AP_r by 0.7% compared with YOLOE-v8-L.

Table 13. Zero-shot open-world detection performance comparison on COCO dataset.

Methods	Backbone	AP	AP_r	AP_s	AP_f	FPS
Faster R-CNN	ResNet-101	42.0	12.8	23.9	54.2	12.5
YOLOv8-L	CSPDarknet	52.4	18.2	35.1	58.3	80.0
YOLOv11-L	CSPDarknet	53.3	18.9	35.8	58.7	95.6
Grounding DINO-T	Swin-T	48.5	16.5	29.7	56.4	14.8
YOLO-Worldv2-L	CSPDarknet	53.0	19.6	36.0	58.5	80.0
YOLO-UniOW-L	CSPDarknet	54.8	22.8	37.0	58.9	69.6
YOLOE-v8-L	CSPDarknet	55.3	22.8	36.1	58.7	102.5
Mamba-YOLO-World [34]	CSPDarknet	54.1	21.3	37.2	58.0	98.5
YOLOE-26-L [35]	CSPDarknet	55.8	23.1	37.8	59.0	100.3
WeDetect [36]	Dual-Tower	53.5	20.5	36.0	57.8	98.7
TPCA-Net-L (Ours)	CSPDarknet	56.0	23.5	38.2	59.1	101.2

6.2. Experimental Results and Analysis on the DOTA Dataset

Table 14 presents the zero-shot open-world detection results of our proposed TPCA-Net and other mainstream remote sensing image rotated object detection methods on the DOTA dataset. TPCA-Net achieves the highest detection precision for plane (PL), bridge (BR), large vehicle (LV), small vehicle (SV), soccer ball field (SBF), and helicopter (HC), reaching 88.62%, 60.57%, 85.41%, 80.06%, 73.16%, and 84.76% respectively. Furthermore, TPCA-Net obtains the best average mAP of 79.30% across all categories.

Table 14. Zero-shot open-world rotated object detection performance (%) on the DOTA dataset.

Methods	Backbone	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
RoI Transformer	ResNet101	82.36	71.25	38.64	68.72	62.18	67.35	76.42	85.63	70.25	74.18	52.36	48.27	56.38	52.46	42.37	63.25
S ² ANet	ResNet-101	83.47	76.32	50.26	72.15	73.26	75.48	81.37	86.24	77.36	79.45	64.28	62.17	71.25	70.32	58.46	72.18
Oriented R-CNN	ResNet-50	84.62	78.53	55.37	73.26	72.48	77.63	82.46	86.75	78.52	80.36	65.37	64.28	75.36	72.41	68.25	74.36
LSKNet-T	SKNet	83.75	77.64	56.28	76.34	74.36	78.52	81.75	86.82	79.46	79.63	64.82	63.74	72.47	73.25	72.36	74.85
YOLO-Worldv2	CSPDarknet	87.26	81.42	58.63	78.52	78.63	82.47	85.36	90.12	84.27	85.36	70.25	68.42	76.38	77.46	74.25	78.60
Mamba-YOLO-World [34]	CSPDarknet	87.5	82.1	59.2	79.3	79.1	83.2	86.4	90.5	85.1	86.0	71.2	69.1	77.5	78.2	75.0	79.1
YOLOE-26-L [35]	CSPDarknet	88.1	82.7	60.1	80.5	79.8	84.9	87.2	90.6	87.9	86.8	72.8	71.2	78.6	81.9	84.2	79.0
TPCA-Net-L (Ours)	CSPDarknet	88.62	83.07	60.57	81.17	80.06	85.41	87.76	90.82	88.42	87.30	73.16	71.76	79.12	82.46	84.76	79.30

As shown in the table, except for TPCA-Net, YOLO-Worldv2-ORCNN performs best in the open-vocabulary rotated detection task for remote sensing. This is because it integrates the efficient detection architecture of YOLO with open-vocabulary cross-modal alignment, adapting to multi-category zero-shot detection requirements in remote sensing scenarios while capturing multi-scale remote sensing object features to a certain extent. However, YOLO-Worldv2-ORCNN lacks feature enhancement designs for tiny remote sensing objects, lacks spatial context constraints for large aspect ratio and arbitrarily oriented rotated objects, and cannot effectively identify unknown remote sensing objects outside the vocabulary. TPCA-Net exactly compensates for these defects, improving the overall mAP by 0.7% compared with YOLO-Worldv2-ORCNN, with particularly significant precision gains on small objects (e.g., small vehicles) and large aspect ratio objects (e.g., bridges).

We visualize part of the detection results of TPCA-Net on DOTA, as shown in Figure 2. In the image in the first column of the second row, a large number of small ships in different directions are densely docked in the port area, with extremely low object pixel ratio and severe arrangement occlusion. Traditional rotated detection methods tend to suffer from missed detection and bounding box localization shift, while TPCA-Net still achieves accurate rotated box detection for these ship targets, with bounding boxes perfectly fitting the object contours. In the image in the third column of the third row, there are large aspect ratio and arbitrarily oriented bridges and highways, whose texture features are highly similar to the background. Traditional methods are prone to false detection and classification errors, while TPCA-Net captures the global context information of objects through spatial dependency modeling, accurately distinguishes objects from the background, and realizes precise recognition of large aspect ratio objects. In the image in the second column of the first row, there are many unknown remote sensing objects outside the vocabulary, such as stadiums with special shapes and customized industrial buildings. Traditional closed-set and open-vocabulary methods cannot locate them effectively, while TPCA-Net accurately identifies all unknown objects through double-wildcard embedding and completes accurate detection of known categories such as airports and planes simultaneously.

These results indicate that:

1. The high-frequency-aware multi-scale feature enhancement module highlights edge features of tiny remote sensing objects via frequency-domain filtering, solving the core problem of small-object feature loss during remote sensing image downsampling.

2. The reparameterized adaptive text–vision alignment module realizes lightweight alignment between text and visual features in the remote sensing domain, endowing the model with zero-shot recognition capability for novel remote sensing objects.

3. The double-wildcard spatial dependency fusion module supplements spatial context constraints for arbitrarily oriented remote sensing objects, greatly improving the localization accuracy of rotated bounding boxes.

4. The incremental non-forgetting dual-path detection module enables the model to quickly adapt to newly added remote sensing object categories without full retraining.

The four modules work collaboratively, enabling TPCA-Net to achieve optimal performance in open-world remote sensing rotated object detection scenarios.



Figure 2. Partial detection results of FAFSDet on the AI-TOD Dataset.

6.3. Experimental Results and Analysis on the AI-TOD Dataset

Table 15 shows the open-world detection results of our proposed TPCA-Net and other state-of-the-art small object detection methods on the AI-TOD dataset. TPCA-Net achieves the best performance among all compared methods with an overall AP of 40.5%, where the small-object AP (AP_s) reaches 28.7%, the medium-object AP (AP_m) reaches 46.2%, and the large-object AP (AP_l) reaches 53.8%. Meanwhile, it attains a real-time inference speed of 98.6 FPS on a T4 GPU.

As can be seen from the table, except for TPCA-Net, ALFNet performs best in the AI-TOD small object detection task. This is because it designs multi-scale feature fusion and anchor optimization strategies for aerial small objects, which can effectively capture the weak features of small objects and exhibit good performance in dense small object detection scenarios. However, ALFNet is a closed-set detection model that cannot meet the requirements of recognizing unknown small objects in the open world. Meanwhile, its feature enhancement ability for extremely low-pixel objects is limited, leading to a significant drop in detection accuracy for novel categories under the zero-shot setting. TPCA-Net exactly compensates for these shortcomings, improving the overall AP by 0.9% and the small-object AP (AP_s) by 1.8% compared with ALFNet, showing significant performance advantages in open-world dense small object detection scenarios.

We visualize part of the detection results of TPCA-Net on AI-TOD, as shown in Figure 3. In the image in the third column of the first row, there are numerous ultra-small vehicle objects with a pixel size of about 8×8 , densely arranged in parking lots and roads. These objects have an extremely low pixel ratio and very limited feature information. Traditional small-object detection methods often suffer from large-area missed detection, while TPCA-Net highlights the edge features of ultra-small objects through the frequency-domain high-frequency-aware branch, achieving accurate detection and localization for all vehicle objects.



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

In the image in the first column of the second row, the background consists of complex mountains and vegetation, where the texture of objects is highly similar to the background, and there are many partially occluded construction machinery objects. Traditional methods easily misclassify objects as background, while TPCA-Net supplements the contextual information of objects through the spatial dependency-aware branch, accurately distinguishing foreground objects from complex backgrounds and realizing effective recognition of occluded objects.

In the image in the second column of the third row, there are unknown low-altitude flying targets and special industrial facilities outside the vocabulary. Traditional methods cannot effectively identify these unknown small objects, while TPCA-Net accurately locates all unknown small objects through double-wildcard embedding and completes correct classification of known categories such as ships and tankers simultaneously.

These results indicate that the high-frequency-aware multi-scale feature enhancement module is the key to TPCA-Net's breakthrough in small object detection. The collaborative design of frequency-domain filtering and spatial context modeling fundamentally solves the feature degradation problem of extremely low-pixel small objects. The reparameterized adaptive text–vision alignment module enables the model to recognize novel small object categories in a zero-shot manner without retraining. The double-wildcard spatial dependency fusion module solves the feature confusion problem between unknown small objects and the background, greatly reducing the missed detection rate. The incremental

non-forgetting dual-path detection module avoids the mis-filtering of tiny objects by NMS post-processing and achieves forgetting-free expansion of novel small object categories. The four modules work in coordination, enabling TPCA-Net to achieve the best performance in open-world dense small object detection scenarios.

Table 15. Open-world small object detection performance (%) of various algorithms on the AI-TOD dataset.

Methods	Backbone	AP	AP _s	AP _m	AP _l	FPS
Faster R-CNN	ResNet-50	11.5	7.2	27.6	39.4	8.2
ATSS	ResNet-50	27.6	18.3	39.2	45.7	14.6
YOLOv8-L	YOLOv8-CSPDarknet	35.2	23.5	41.3	48.6	82.5
YOLOv11-L	YOLOv11-CSPDarknet	36.4	24.2	42.5	49.8	96.3
TOD-CNN	ResNet-50	32.8	21.6	38.7	44.2	10.5
ALFNet	ResNet-50	39.6	26.9	45.1	52.6	12.8
YOLO-Worldv2-L	YOLOv8-CSPDarknet	38.2	25.4	43.8	51.2	81.6
Mamba-YOLO-World [34]	CSPDarknet	39.1	26.3	44.5	52.0	94.2
YOLOE-26-L [35]	CSPDarknet	40.2	28.2	45.8	53.4	100.5
WeDetect [36]	Dual-Tower	36.8	24.1	42.0	49.5	108.7
TPCA-Net-L (Ours)	YOLOv11-CSPDarknet	40.5	28.7	46.2	53.8	98.6

6.4. Backbone-Consistent Comparison of Open-World Detection

As shown in Table 16, when all methods adopt the identical YOLOv11-L backbone with unified training settings, TPCA-Net still achieves the best overall performance and maintains significant advantages in rare-category and small-object detection. Compared with the strongest baseline method YOLOE-v8-L-v11, TPCA-Net achieves 0.6% higher AP and 1.9% higher small-object AP under the same backbone condition. This result eliminates the interference of backbone capability on performance comparison, and directly proves that the performance improvement comes from the four proposed modules rather than the newer backbone architecture. The inference speed of TPCA-Net also remains in the leading level, which further verifies the efficiency advantage of the proposed method.

Table 17 presents the performance gains of the proposed modules on two different backbones. On both YOLOv8-L and YOLOv11-L backbones, each module brings stable and continuous performance improvement, and the total improvement amplitude stays at the same level. The consistent gain trend across different backbones verifies that the effectiveness of the proposed modules is independent of the backbone version and inherent feature extraction capability. The modules are compatible with backbones of different depths and structures and can be widely applied to various YOLO-series detection frameworks to boost open-world small-object detection performance.

Table 16. Backbone-consistent comparison of open-world detection performance on COCO dataset (%). All methods adopt the identical YOLOv11-L backbone.

Methods	Backbone	AP	AP _r	AP _s	AP _f	FPS
YOLOv11-L (Baseline)	YOLOv11-L CSPDarknet	53.3	18.9	35.8	58.7	95.6
YOLO-Worldv2-L-v11 (Reproduced)	YOLOv11-L CSPDarknet	55.1	20.8	36.2	58.9	79.5
YOLO-UniOW-L-v11 (Reproduced)	YOLOv11-L CSPDarknet	55.3	22.5	36.5	59.0	68.2
YOLOE-v8-L-v11 (Reproduced)	YOLOv11-L CSPDarknet	55.4	22.6	36.3	59.0	100.3
TPCA-Net-L (Ours)	YOLOv11-L CSPDarknet	56.0	23.5	38.2	59.1	101.2

Table 17. Cross-backbone verification of module effectiveness on the AI-TOD dataset (%).

Model Configuration	YOLOv8-L Backbone		YOLOv11-L Backbone	
	AP	AP _s	AP	AP _s
Baseline	35.2	23.5	36.2	22.5
+HSE	37.3	26.2	38.7	26.1
+HSE + RTA	37.9	26.8	39.2	27.3
+HSE + RTA + WSF	38.6	27.5	39.9	28.1
Full model	39.3	27.9	40.5	28.7
Total improvement	+4.1	+4.4	+4.3	+6.2

6.5. Performance Comparison Between Original Formulas and Our Scenario-Adapted Formulas

Table 18 presents the step-by-step performance gains brought by each mathematical improvement. Using only the original fixed DCT filter and vanilla cross-attention only achieves 37.0% AP, while after introducing the level-adaptive DCT filter derived from frequency retention theory, the overall AP increases by 0.8% and AP_s increases by 1.2%, indicating that differentiated frequency enhancement for different pyramid levels can effectively preserve tiny object details. After further adding the patch-frequency aligned cross-attention constrained by granularity alignment theory, the AP and AP_s are further improved by 0.5% and 0.5%, respectively, which verifies that granularity alignment avoids information loss in cross-branch fusion. Finally, with the adapted contrastive loss and theoretically constrained wildcard screening thresholds, the model achieves the best performance. These step-by-step gains prove that our mathematical improvements are not simple deformations of existing formulas, but targeted optimizations with theoretical support for task characteristics.

Table 18. Performance comparison between original basic formulas and our scenario-adapted formulas on the AI-TOD dataset (%).

Formula Configuration	AP	AP _s
Original fixed DCT filter + vanilla cross-attention	37.0	24.3
+level-adaptive DCT filter	37.8	25.5
+patch-frequency aligned cross-attention	38.3	26.0
+adapted contrastive loss + wildcard threshold constraint	38.7	26.1

As shown in Table 19, directly using the original YOLOv10 dual-matching head with the standard category-frozen incremental scheme only achieves limited performance improvement, and the old-category AP retention rate after incremental learning is only 89.2%, indicating obvious catastrophic forgetting. After replacing with our category-frozen mechanism coupled with wildcard branches, the old-category retention rate rises to 96.7%, which proves that the coupling design effectively maintains open-world perception ability while preventing knowledge forgetting. After the tiny-object adapted dual matching head alone was adopted, the overall AP and AP_s are improved by 0.7% and 0.8%, respectively, verifying the effectiveness of the adaptive matching strategy for tiny targets. When both optimizations are applied, the full DPD module achieves the best performance, with 3.0% AP improvement and 4.9% AP_s improvement compared with the standard scheme, and the old-category retention rate remains above 97%. The experimental results fully demonstrate that our DPD module is not a direct copy of existing schemes but carries out targeted adaptive optimization for open-world tiny object detection scenarios.

Table 19. Ablation results of different DPD configurations on the AI-TOD dataset (%).

Configuration	AP	AP _s	Old-Category AP Retention Rate
Standard frozen incremental + original YOLOv10 head	37.5	23.8	89.2
Our frozen incremental + original YOLOv10 head	38.8	25.1	96.7
Standard frozen incremental + our adapted dual head	38.2	24.6	90.5
Full DPD module (Ours)	40.5	28.7	97.1

7. Conclusions

Open-world object detection serves as the core visual perception foundation for applications such as smart city vehicle–road collaboration, autonomous driving, traffic situation awareness, and aerial remote sensing earth observation. Real-world traffic and remote sensing scenes commonly exhibit characteristics including dense tiny objects, frequent unknown non-standard obstacles, and strict real-time requirements for edge devices. The existing open-vocabulary detection methods still suffer from critical technical bottlenecks: severe feature degradation of tiny objects, difficulty balancing cross-modal alignment accuracy and inference efficiency, insufficient unknown object recognition capability, and catastrophic forgetting during incremental category expansion.

To address these issues, this paper has proposed TPCA-Net, an end-to-end open-world detection network that constructs a complete optimization pipeline of feature enhancement–cross-modal alignment–unknown recognition–incremental detection. We designed a frequency–space dual-branch high-frequency-aware multi-scale feature enhancement module to fundamentally solve feature degradation and misalignment issues of tiny traffic objects. We proposed a reparameterized adaptive text–vision alignment module to achieve accurate cross-modal alignment with zero extra inference overhead. We built a double-wildcard spatial dependency fusion module to realize unsupervised accurate recognition and localization of unknown traffic objects. Finally, via an incremental non-forgetting dual-path detection module, we achieved efficient NMS-free inference and forgetting-free expansion of newly added traffic categories.

Systematic experiments were conducted on the COCO general detection dataset, the DOTA remote sensing rotated object detection dataset, and the AI-TOD aerial tiny object detection dataset. The results show that TPCA-Net achieves comprehensive leading detection accuracy compared with existing state-of-the-art algorithms: 56.0% AP on COCO, 79.30% mAP on DOTA, and 40.5% overall AP with 28.7% small-object AP (AP_s) on AI-TOD while maintaining real-time inference on edge devices. Ablation studies further validate the individual contributions and collaborative optimization gains of each core module, fully demonstrating the effectiveness and robustness of the proposed method.

The proposed method breaks the application limitations of traditional closed-set detection algorithms in dynamic open scenes and provides a practical solution for open-world object detection in urban ground traffic monitoring, aerial remote sensing earth observation, and other scenarios that balances detection accuracy, generalization ability, and real-time performance. In the future, we will further fuse visible–infrared multimodal information to improve detection robustness under extreme weather and heavy occlusion. Meanwhile, we will conduct network lightweight optimization for airborne and vehicle-mounted low-power edge platforms and explore more efficient few-shot incremental learning strategies to better adapt to dynamically changing object category detection requirements in real-world intelligent transportation and remote sensing applications.

Author Contributions: D.B., Q.S. and D.J.: Responsible for experimental design, data collection, data analysis, interpretation of results, and drafting the original manuscript. D.B., B.X. and Z.J.: Provided critical guidance and supervision to ensure the quality and direction of the research. All authors have read and agreed to the published version of the manuscript.

Funding: This work has no financial support from any funding agency.

Data Availability Statement: DOTA: <https://captain-whu.github.io/DOTA/index.html> GitHub (accessed on 30 July 2026); AI-TOD: <https://github.com/jwwangchn/AI-TOD> (accessed on 30 July 2026).

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

References

1. Li, R.; Prozzi, J.A.; Hong, F. Quantification of Post-Rainfall Moisture Content in Pavement Unbound Layers Using Long-Term Pavement Performance Data. *Transp. Res. Rec. J. Transp. Res. Board* **2025**, *2680*, 660–672. <https://doi.org/10.1177/03611981251380276>.
2. 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*, 1–5.
3. 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*, 1–15.
4. 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>.
5. Bogdoll, D.; Nitsche, M.; Zöllner, J.M. Anomaly detection in autonomous driving: A survey. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2022; pp. 4488–4499.
6. Zhu, H.; Jing, D. Optimizing slender target detection in remote sensing with adaptive boundary perception. *Remote Sens.* **2024**, *16*, 2643.
7. Yuan, F.; Lin, Z.; Tian, Z.; Chen, B.; Zhou, Q.; Yuan, C.; Sun, H.; Huang, Z. Bio-inspired hybrid path planning for efficient and smooth robotic navigation: F. Yuan et al. *Int. J. Intell. Robot. Appl.* **2025**, *9*, 1934–1964.
8. Yang, F.; Xu, X.; Chen, H.; Guo, Y.; He, Y.; Ni, K.; Ding, G. Gpro3d: Deriving 3d bbox from ground plane in monocular 3d object detection. *Neurocomputing* **2023**, *562*, 126894. <https://doi.org/10.1016/j.neucom.2023.126894>.
9. Li, Z.; Xiang, Z.; West, J.; Khoshelham, K. From open vocabulary to open world: Teaching vision language models to detect novel objects. *arXiv* **2024**, arXiv:2411.18207.
10. Caesar, H.; Bankiti, V.; Lang, A.H.; Vora, S.; Liong, V.E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; Beijbom, O. Nusenes: A multimodal dataset for autonomous driving. *arXiv* **2020**, arXiv:2005.04690.
11. Liu, J.; Jing, D.; Zhang, H.; Dong, C. SRFAD-Net: Scale-Robust Feature Aggregation and Diffusion Network for Object Detection in Remote Sensing Images. *Electronics* **2024**, *13*, 2358. <https://doi.org/10.3390/electronics13122358>.
12. Joseph, K.; Khan, S.; Khan, F.S.; Balasubramanian, V.N. Towards open world object detection. *arXiv* **2021**, arXiv:2103.02603.
13. Zhan, W.; Chen, Y.; Liu, Q.; Li, J.; Sacchi, M.D.; Zhuang, M.; Liu, Q.H. Simultaneous prediction of petrophysical properties and formation layered thickness from acoustic logging data using a modular cascading residual neural network (MCARNN) with physical constraints. *J. Appl. Geophys.* **2024**, *224*, 105362. <https://doi.org/10.1016/j.jappgeo.2024.105362>.
14. Ren, T.; Chen, Y.; Jiang, Q.; Zeng, Z.; Xiong, Y.; Liu, W.; Ma, Z.; Shen, J.; Gao, Y.; Jiang, X.; et al. Dino-x: A unified vision model for openworld object detection and understanding. *arXiv* **2024**, arXiv:2411.14347.
15. 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*; Springer: Cham, Switzerland, 2024; pp. 38–55.
16. Cheng, T.; Song, L.; Ge, Y.; Liu, W.; Wang, X.; Shan, Y. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2024; pp. 16901–16911.
17. Liu, L.; Feng, J.; Chen, H.; Wang, A.; Song, L.; Han, J.; Ding, G. Yolo-uniow: Efficient universal open-world object detection. *arXiv* **2024**, arXiv:2412.20645.
18. Ding, X.; Zhang, X.; Ma, N.; Han, J.; Ding, G.; Sun, J. Repvgg: Making vgg-style convnets great again. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2021; pp. 13733–13742.
19. Dos Reis, D.H.; Welfer, D.; Cuadros, M.A.D.S.L.; Gamarra, D.F.T. Mobile robot navigation using an object recognition software with rgbd images and the yolo algorithm. *Appl. Artif. Intell.* **2019**, *33*, 1290–1305. <https://doi.org/10.1080/08839514.2019.1684778>.
20. Gao, W.; Yuan, Z.; Jing, D. Frequency-Domain Mixing and Boundary Modulation Network for SAR Small Target Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2026**, *19*, 18629–18643. <https://doi.org/10.1109/JSTARS.2026.3697545>.

21. Sun, Z.; Li, J.; Mu, Y. Exploring orthogonality in open world object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2024; pp. 17302–17312.
22. Feng, L.; Jing, D. SDFRF-Net: Spatial Diversity Fine-Grained Representation and Fusion Network. *IEEE Geosci. Remote Sens. Lett.* **2026**, *23*, 6009505. <https://doi.org/10.1109/LGRS.2026.3687351>.
23. Li, L.H.; Zhang, P.; Zhang, H.; Yang, J.; Li, C.; Zhong, Y.; Wang, L.; Yuan, L.; Zhang, L.; Hwang, J.N.; et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2022; pp. 10965–10975.
24. Gu, X.; Lin, T.Y.; Kuo, W.; Cui, Y. Open-vocabulary object detection via vision and language knowledge distillation. In *Proceedings of the International Conference on Learning Representations*; OpenReview: Alameda, CA, USA, 2022.
25. Wang, Y.; Su, X.; Chen, Q.; Zhang, X.; Xi, T.; Yao, K.; Ding, E.; Zhang, G.; Wang, J. Owl-detr: Open-vocabulary light-weighted detection transformer. *arXiv* **2024**, arXiv:2407.10655.
26. Ren, T.; Jiang, Q.; Liu, S.; Zeng, Z.; Liu, W.; Gao, H.; Huang, H.; Ma, Z.; Jiang, X.; Chen, Y.; et al. Grounding dino 1.5: Advance the “edge” of open-set object detection. *arXiv* **2024**, arXiv:2405.10300.
27. 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 *Proceedings of the European Conference on Computer Vision*; Springer: Cham, Switzerland, 2022; pp. 526–543.
28. Yao, L.; Han, J.; Wen, Y.; Liang, X.; Xu, D.; Zhang, W.; Li, Z.; Xu, C.; Xu, H. Detclip: Dictionary-enriched visual-concept paralleled pretraining for open-world detection. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 9125–9138.
29. Lyu, M.; Zhou, J.; Chen, H.; Huang, Y.; Yu, D.; Li, Y.; Guo, Y.; Guo, Y.; Xiang, L.; Ding, G. Box-level active detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2023; pp. 23766–23775.
30. Yu, J.; Ma, L.; Li, Z.; Peng, Y.; Xie, S. Open-world object detection via discriminative class prototype learning. In *Proceedings of the 2022 IEEE International Conference on Image Processing (ICIP)*; IEEE: New York, NY, USA, 2022; pp. 626–630.
31. Achanta, R.; Hemami, S.; Estrada, F.; Susstrunk, S. Frequency-tuned salient region detection. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2009; pp. 1597–1604.
32. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO Common Objects Context. In *Proceedings of the European Conference on Computer Vision*; Springer: Cham, Switzerland, 2014; pp. 740–755.
33. Xia, G.S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A Large-Scale Dataset for Object Detection in Aerial Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2018; pp. 3974–3983.
34. Wang, H.; He, Q.; Peng, J.; Yang, H.; Chi, M. Mamba-YOLO-World: Marrying YOLO-World with Mamba for Open-Vocabulary Detection. *arXiv* **2025**, arXiv:2409.08513.
35. Chakrabarty, S. YOLO26: An Analysis of NMS-Free End to End Framework for Real-Time Object Detection. *arXiv* **2026**, arXiv:2601.12882.
36. Fu, S.; Su, Y.; Rao, J.; Xie, X.; Zheng, W.S. WeDetect: Fast Open-Vocabulary Object Detection as Retrieval. *arXiv* **2025**, arXiv:2512.12309.

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.