

Article

A^2 -Det: Dual Asymmetric Architecture for Tiny Object Detection in Remote Sensing Imagery

Shi-Jie Fan ^{1,2} , Fan-Lu Wu ¹ , Ze-Xian Huang ^{1,2} , Xin Gao ^{1,2} , Ao Han ^{1,2}  and Xiao-Nan Jiang ^{1,*} 

¹ Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China; fanshijie22@mails.ucas.ac.cn (S.-J.F.); flwu@ciomp.ac.cn (F.-L.W.); huangzexian20@mails.ucas.ac.cn (Z.-X.H.); gaoxin22@mails.ucas.ac.cn (X.G.); hanao23@mails.ucas.ac.cn (A.H.)

² School of Optoelectronics, University of Chinese Academy of Sciences, Beijing 100049, China

* Correspondence: jiangxn@ciomp.ac.cn

Highlights

What are the main findings?

- A^2 -Det combines high-resolution shallow-feature enhancement with task-specific asymmetric prediction to preserve fine-grained tiny object cues and alleviate classification–regression representation inconsistency.
- On VisDrone, A^2 -Det improves mAP₅₀ by 5.8 percentage points over YOLO11n, while consistent gains on USOD and RSOD demonstrate favorable cross-scene generalization.

What are the implications of the main findings?

- The results show that preserving shallow fine-grained information and differentiating classification and localization representations are complementary strategies for remote-sensing tiny object detection.
- The framework is particularly suitable for accuracy-oriented remote-sensing and UAV applications involving dense tiny objects and complex backgrounds, where sufficient computational resources are available.

Abstract

To address the attenuation of shallow fine-grained structural information during deep feature propagation and the representational conflict between classification and regression in remote sensing tiny object detection, this paper proposes a dual asymmetric detection framework, termed A^2 -Det. Built upon YOLO11n, the detection pyramid is shifted toward the high-resolution P2–P4 levels by introducing a P2 detection branch and removing the original P5 stage, thereby reducing the loss of fine-grained spatial details caused by successive downsampling. On this high-resolution feature basis, a Query–Key–Value (QKV)-guided Asymmetric Spatial Feature Enhancement module (Q-ASFE) is deployed at the P2 and P3 stages, where direction-sensitive asymmetric convolutions are combined with lightweight QKV-based contextual modulation to strengthen tiny object structural responses while suppressing complex background interference. Furthermore, an Asymmetric Coordinate–Semantic Decoupled Head (ACS-Head) performs differentiated modeling for the semantic selection required by classification and the coordinate-sensitive representation required by regression, thereby alleviating task-specific representational conflict. Through the progressive coordination of high-resolution feature preservation, shallow-feature enhancement, and task-specific prediction, A^2 -Det improves mAP₅₀ by 5.8 percentage points over YOLO11n on the VisDrone dataset. Consistent improvements on the USOD and RSOD datasets further demonstrate its effectiveness and cross-scene generalization.



Academic Editor: Salah Bourennane

Received: 29 August 2026

Revised: 19 September 2026

Accepted: 22 September 2026

Published: 30 September 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

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

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: tiny object detection; remote sensing; YOLO; asymmetric feature enhancement; QKV-guided modulation

1. Introduction

Remote sensing object detection plays an important role in a wide range of applications, including urban traffic monitoring, maritime inspection, disaster assessment, and intelligent security surveillance [1,2]. Compared with natural-image scenarios, targets in aerospace and UAV remote sensing imagery are often characterized by small scales, dense spatial distributions, complex backgrounds, and substantial variations in imaging conditions. These characteristics make existing general-purpose detectors prone to noticeable performance degradation when applied to remote sensing scenes [2–4]. This issue is particularly pronounced for tiny objects, which occupy only a limited image area and contain very limited texture, edge, and shape information. For quantitative evaluation in this study, tiny objects are operationally defined following the MS COCO small-object criterion as instances with a bounding-box area smaller than 32×32 pixels. As a result, their weak visual cues can be easily overwhelmed by background patterns during feature extraction and cross-layer fusion, or further attenuated by successive downsampling, leading to missed detections, false alarms, and unstable localization [5]. Therefore, how to reliably preserve fine-grained structural information of tiny objects while suppressing complex background interference remains a critical challenge in remote sensing vision [5,6].

In recent years, tiny object detection has evolved from scale-robust modeling in generic scenarios toward task-oriented architectural designs tailored to remote sensing images with complex backgrounds, large scale variations, and dense object distributions [7,8]. Feature pyramid structures represented by FPN and PANet have laid the foundation for unified multi-scale object representation through cross-level information propagation and multi-scale feature fusion [9,10]. With the development of one-stage detectors, the YOLO series has become a commonly adopted base detection framework for remote sensing tiny object detection, owing to its mature architecture, stable performance, and ease of extension.

Beyond conventional feature-pyramid and one-stage CNN designs, recent deep-learning-based remote sensing object detection has evolved toward increasingly specialized representation, prediction, and deployment paradigms. At the feature-representation level, multi-scale and fine-grained modeling remains fundamental for preserving discriminative cues of tiny and densely distributed objects, while Transformer- and DETR-based detectors further complement local convolutional representations through long-range dependency modeling, deformable attention, and query-based end-to-end prediction. For example, PR-Deformable DETR combines adaptive multi-scale feature fusion, deformable encoding, and dynamic reference-point modeling to enhance both object representation and localization in complex remote sensing scenes [11]. Beyond conventional spatial object detection, task-specific deep representation learning has also been explored in other remote sensing detection tasks. Lin et al. integrated dynamic low-rank and sparse priors with deep autoencoders to enhance discriminative representation for hyperspectral anomaly detection [12]. At the prediction level, task-aligned and decoupled designs have also been introduced to better accommodate the different representation and optimization requirements of classification and localization. DCYOLO, for instance, employs cross-layer decoupled prediction together with task-aligned label assignment to improve the utilization of multi-scale information in remote sensing detection [13]. Meanwhile, practical remote sensing applications have motivated increasing attention to accuracy–efficiency trade-offs. EL-Net combines lightweight feature fusion, feature enhancement, and structured pruning

to reduce model complexity while maintaining competitive detection performance for UAV imagery [14]. More recently, open-vocabulary remote sensing detectors such as LAE-DINO have exploited vision–language representations to extend conventional closed-set detection toward previously unseen object categories [15]. Collectively, these developments indicate that modern remote sensing detection increasingly adopts task-specific representation and prediction strategies to address scale variation, long-range contextual dependencies, target–background discrimination [16,17], classification–localization heterogeneity, computational constraints, and semantic openness.

Among these broader developments, multi-scale and fine-grained feature modeling is particularly relevant to remote sensing tiny object detection, where weak target responses are highly sensitive to spatial-resolution degradation. Building upon feature pyramids and multi-scale detection frameworks such as YOLO, existing studies on remote sensing tiny object detection have mainly improved detectors by introducing high-resolution branches, redesigning feature pyramids, enhancing cross-level interactions, and embedding task-specific functional modules, with the aim of alleviating insufficient multi-scale representation for tiny objects [7,8,18,19]. Tang et al. proposed HIC-YOLOv5, which introduces a high-resolution prediction head for tiny objects and incorporates a feature enhancement module between the backbone and the neck to strengthen the representation of shallow fine-grained features [20]. Li et al. developed the Cross-Layer Attention Network, where cross-layer attention is used to integrate shallow spatial details with deep semantic information, thereby mitigating the attenuation of tiny object features during downsampling [18]. Bu et al. proposed OD-YOLO, which embeds Dynamic Head into the feature pyramid of YOLOv8 and combines it with a detection refinement module to enhance geometric structures and ambiguous object representations, improving target responses under complex backgrounds [7]. In addition, Zhou et al. introduced SMA-YOLO, which constructs a bidirectional multi-branch auxiliary feature pyramid to strengthen the interaction between high-level semantic information and low-level spatial details, while using an adaptive detection mechanism to reduce multi-scale feature inconsistency [8]. Zhang et al. proposed FFCA-YOLO, in which feature enhancement, feature fusion, and spatial context-aware modules are jointly designed to reinforce local detail modeling, cross-scale information interaction, and global contextual association [21]. Overall, these methods have effectively improved the multi-scale representation capability for remote sensing tiny objects. Nevertheless, they mainly focus on feature propagation, scale fusion, and cross-level semantic interaction, while explicit enhancement of direction-sensitive local structures in shallow high-resolution features remains insufficiently explored.

Beyond scale modeling, feature enhancement and contextual representation have also become important research directions for remote sensing tiny object detection. Classical attention mechanisms, such as SE, CBAM, and Coordinate Attention, enhance intermediate features from the perspectives of channel recalibration, joint channel–spatial modeling, and coordinate-sensitive representation, respectively [22–24]. At the feature-modeling level, self-attention architectures such as ViT and Swin Transformer have demonstrated the effectiveness of Query–Key–Value-based self-attention in modeling long-range dependencies and representing multi-scale contextual information [25–27]. However, directly applying global self-attention to high-resolution remote sensing features usually incurs considerable computational and memory overhead. Therefore, localized, hierarchical, or lightweight contextual modeling strategies are generally more suitable for remote sensing tiny object detection. To address complex backgrounds and weak target responses in remote sensing imagery, Zheng et al. proposed ESL-YOLO, which combines multi-scale local contextual enhancement with spatial-context-guided feature fusion and a local attention pyramid to preserve tiny object information and suppress background interference [28]. Wang et al.

proposed a remote sensing tiny object detection method based on multi-contextual information aggregation, which integrates contextual cues across channel, depth, and pixel dimensions and further combines sparse cross self-attention with multi-receptive-field weighted fusion to strengthen tiny object representations [6]. In addition to attention and contextual mechanisms, deformable convolutions and large-kernel selective modeling have been employed to improve the network's ability to capture geometric variations, salient regions, and adaptive receptive fields [29–31]. Asymmetric convolution provides another approach to directional feature modeling. ACNet [32] employs asymmetric convolution branches to strengthen square convolution kernels during training and equivalently converts the multi-branch structure into the original convolutional form for inference, thereby avoiding additional inference-time computation. From another perspective, NWD models bounding boxes as Gaussian distributions and uses the normalized Wasserstein distance to alleviate the excessive sensitivity of tiny objects to pixel-level localization deviations [5]. Overall, existing methods have advanced remote sensing tiny object detection through attention enhancement, QKV/self-attention, dynamic sampling, adaptive receptive-field expansion, and improved matching criteria. Nevertheless, most of them still emphasize general feature enhancement or global context modeling, while a dedicated collaborative design for stable representation of direction-sensitive local structures and suppression of complex background interference in shallow high-resolution features remains insufficiently explored.

Meanwhile, the design of detection heads has gradually shifted from shared prediction toward decoupled modeling that accounts for task-specific differences. Ge et al. proposed YOLOX, in which a decoupled head and SimOTA are introduced into a one-stage detector, demonstrating the effectiveness of separating classification and regression branches for improving detection performance [33]. Feng et al. proposed TOOD, which explicitly coordinates the inconsistency between classification and localization through a Task-aligned Head and Task Alignment Learning [34]. Dai et al. introduced Dynamic Head, which enhances detection-head representations by unifying scale-aware, spatial-aware, and task-aware attention, providing a more general perspective for task interaction in detection heads [35]. Furthermore, Zhuang et al. proposed TSCODE, pointing out that classification favors semantic contextual information, whereas localization relies more on high-resolution boundary cues and spatial details. By decoupling task-specific contexts, TSCODE further enlarges the representational distinction between classification and regression [36]. In remote sensing and UAV tiny object detection, task-specific prediction has also received increasing attention. Xie et al. proposed MTD-YOLO, whose shared-decoupled detection head integrates task decoupling, dynamic alignment, and adaptive scale modulation to improve tiny object prediction under complex scenes [37]. In addition, YOLOv10 advances the end-to-end one-stage detection paradigm through consistent dual assignments and has become an important reference for comparing recent dual-head or task-separated detection structures [38]. These studies indicate that classification and regression are not fully homogeneous prediction tasks, and that detection heads should be specifically designed according to their task differences. Nevertheless, existing methods mainly focus on generic decoupling, lightweight enhancement, or contextual expansion, while fine-grained heterogeneous modeling for the semantic selection required by classification and the coordinate sensitivity required by regression remains insufficiently explored.

Based on the above analysis, we argue that dense tiny object detection in complex remote sensing scenarios involves two closely coupled representation challenges rather than two independent problems. First, the limited spatial extent and weak visual responses of tiny objects make their fine-grained structural information particularly vulnerable to attenuation during successive downsampling and cross-level feature propagation [5,18,20].

Once these spatial cues are substantially weakened, subsequent prediction stages have limited ability to recover the information required for reliable recognition and localization. Second, even when informative shallow features are preserved, classification and regression do not exploit them in the same manner: classification relies more on discriminative semantic information, whereas regression is more sensitive to coordinate-related spatial and structural cues [33,34,36]. Therefore, shallow-feature enhancement alone cannot fully resolve task-specific representation inconsistency, while task decoupling alone cannot compensate for fine-grained information that has already been degraded during feature extraction.

Motivated by this dependency, we design A^2 -Det as a progressive dual-asymmetric framework that jointly considers feature preservation and task-specific utilization. At the representation stage, the detection pyramid is shifted toward the high-resolution P2–P4 levels, and Q-ASFE further strengthens direction-sensitive local structures and suppresses background interference in shallow features. At the prediction stage, ACS-Head performs differentiated modeling according to the distinct semantic and coordinate-sensitive requirements of classification and regression. In this way, Q-ASFE introduces structural asymmetry during feature enhancement, while ACS-Head introduces task asymmetry during prediction. The two designs operate at complementary stages of the same detection pipeline, thereby forming a unified dual-asymmetric architecture rather than two independent architectural modifications. The main contributions are summarized as follows:

1. A dual asymmetric framework, termed A^2 -Det, is proposed for remote sensing tiny object detection. It establishes a progressive representation pathway from high-resolution feature preservation to task-specific prediction, in which structural asymmetry in feature enhancement and task asymmetry in prediction jointly address shallow-feature degradation and classification–regression inconsistency.
2. A QKV-guided Asymmetric Spatial Feature Enhancement module (Q-ASFE) is proposed. By combining direction-sensitive structural modeling with lightweight context modulation, the module strengthens target-related responses in shallow features while suppressing interference from complex backgrounds.
3. An Asymmetric Coordinate–Semantic Decoupled Head (ACS-Head) is designed to model, in a differentiated manner, the semantic selection requirement of the classification branch and the coordinate-sensitive requirement of the regression branch, thereby alleviating the representational conflict between classification and regression.
4. Experiments on the VisDrone, USOD, and RSOD datasets demonstrate that the proposed method achieves consistent advantages in detection performance, module effectiveness, and cross-scene generalization.

2. Methodology

2.1. Overview

For the baseline model, YOLO11n [39], which offers relatively comprehensive operator support and stable training configurations, is adopted as the fundamental detection framework in this study. In terms of naming convention, unless otherwise specified, A^2 -Det in this paper refers to the default n -scale version built upon YOLO11n, namely, the n -scale variant corresponding to YOLO11n. When larger model scales are involved, they are denoted as A^2 -Det-m and A^2 -Det-x, respectively. In the subsequent experimental tables and figures, if no additional suffix is explicitly marked, A^2 -Det denotes this default n -scale version.

The overall architecture of the proposed A^2 -Det is illustrated in Figure 1. To address the loss of spatial details caused by successive downsampling in deep layers for tiny objects [20,21,40], we remove the deep P5 branch from the macro-level topology and introduce a high-resolution P2 detection head with a downsampling ratio of 1/4, thereby

increasing the contribution of shallow fine-grained information to detection. Furthermore, Q-ASFE is embedded at the P2 and P3 stages of the backbone, while the original detection head is replaced at the prediction stage with an ACS-Head. The detailed structures of Q-ASFE and ACS-Head are presented in Section 2.2 and Section 2.3, respectively.

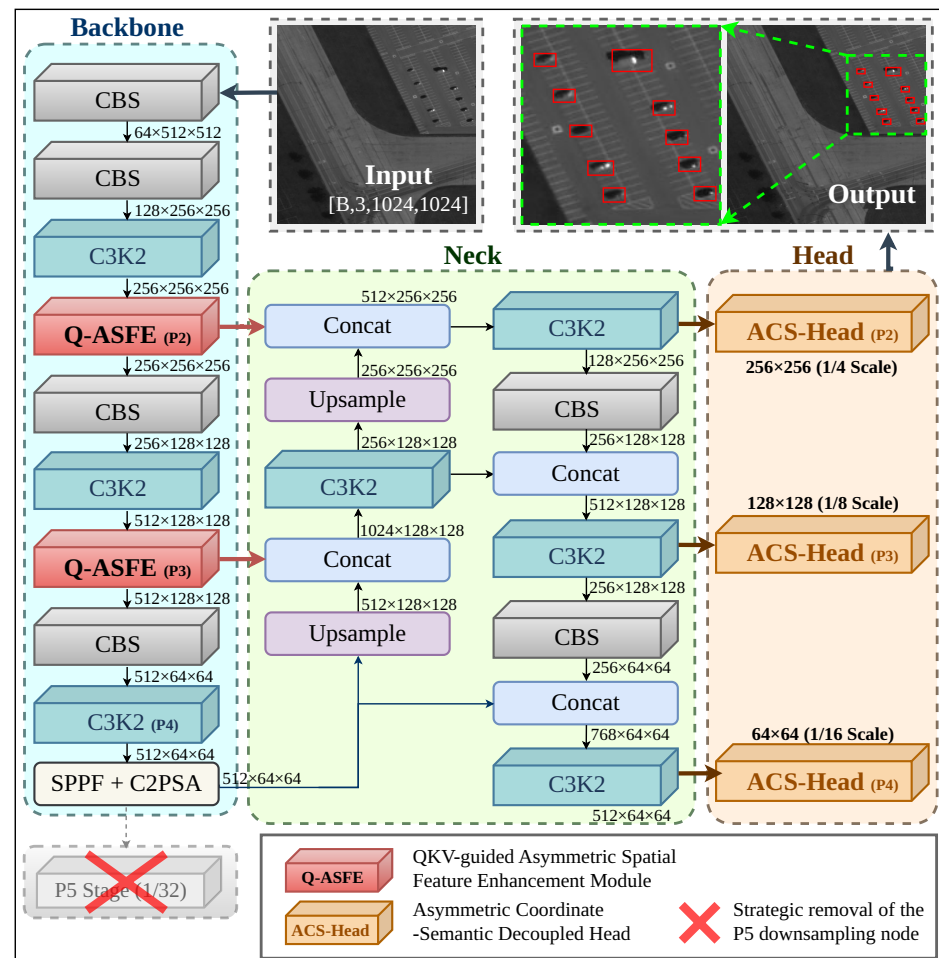


Figure 1. Overall network architecture of the proposed A²-Det for tiny object detection.

To further clarify the sequential operations and task-specific branches of the proposed framework, Algorithm 1 summarizes the complete training and inference procedures of A²-Det. The overall process includes high-resolution feature extraction, Q-ASFE-based shallow-feature enhancement, multi-scale feature fusion, ACS-Head prediction, and the subsequent training or inference procedure. The notation used in Algorithm 1 is consistent with the definitions of Q-ASFE and ACS-Head given in the following sections.

2.2. QKV-Guided Asymmetric Spatial Feature Enhancement Module (Q-ASFE)

Tiny objects in remote sensing imagery are typically characterized by limited scale and weak texture, and their local structures often exhibit pronounced spatial anisotropy. For such targets, standard convolution remains inherently limited in modeling direction-sensitive structures and contextual information. Although the P2–P4 pyramid preserves more fine-grained information, shallow high-resolution features also retain substantial background textures and redundant local responses. Therefore, simply preserving shallow features is insufficient; their target-related structures need to be selectively strengthened while background interference is suppressed. To address this issue, we propose a Q-ASFE, which couples an Asymmetric Spatial Feature Extractor (ASFE) with an efficient QKV-based context modulation branch. The overall structure of the module is shown in Figure 2.

Algorithm 1 Training and Inference Procedure of A^2 -Det

Require: Input image I ; annotation Y (training only); model parameters θ ; running mode mode

Ensure: Updated model parameters θ (training) or detection results $\mathcal{D}$ (inference)

```

1:  $\tilde{I} \leftarrow \text{Preprocess}(I)$ 
2:  $\mathbf{X}_{P2} \leftarrow \text{Backbone}_{P2}(\tilde{I})$ 
3:  $\mathbf{F}_{\text{out}}^{P2} \leftarrow \text{Q-ASFE}(\mathbf{X}_{P2})$  ▷ direction-sensitive feature enhancement
4:  $\mathbf{X}_{P3} \leftarrow \text{Backbone}_{P3}(\mathbf{F}_{\text{out}}^{P2})$ 
5:  $\mathbf{F}_{\text{out}}^{P3} \leftarrow \text{Q-ASFE}(\mathbf{X}_{P3})$  ▷ direction-sensitive feature enhancement
6:  $\mathbf{X}_{P4} \leftarrow \text{Backbone}_{P4}(\mathbf{F}_{\text{out}}^{P3})$ 
7:  $(\mathbf{F}_{\text{in}}^{P2}, \mathbf{F}_{\text{in}}^{P3}, \mathbf{F}_{\text{in}}^{P4}) \leftarrow \text{MultiScaleFeatureFusion}(\mathbf{F}_{\text{out}}^{P2}, \mathbf{F}_{\text{out}}^{P3}, \mathbf{X}_{P4})$  ▷  $P4 \rightarrow P3 \rightarrow P2$ ;  $P2 \rightarrow P3 \rightarrow P4$ 
8: for  $i \in \{P2, P3, P4\}$  do
9:    $\mathbf{P}_{\text{cls}}^i \leftarrow \text{ACS-Head}_{\text{cls}}(\mathbf{F}_{\text{in}}^i)$  ▷ semantic selection
10:   $\mathbf{P}_{\text{reg}}^i \leftarrow \text{ACS-Head}_{\text{reg}}(\mathbf{F}_{\text{in}}^i)$  ▷ coordinate-sensitive modeling
11: end for
12: if mode = training then
13:    $\mathcal{L}_{\text{det}} \leftarrow \text{DetectionLoss}(\{\mathbf{P}_{\text{cls}}^i, \mathbf{P}_{\text{reg}}^i\}_{i=P2}^{P4}, Y)$ 
14:    $\theta \leftarrow \text{Update}(\theta, \nabla_{\theta} \mathcal{L}_{\text{det}})$ 
15:   return  $\theta$ 
16: else
17:    $\mathcal{D} \leftarrow \text{PostProcess}(\{\mathbf{P}_{\text{cls}}^i, \mathbf{P}_{\text{reg}}^i\}_{i=P2}^{P4})$ 
18:   return  $\mathcal{D}$ 
19: end if

```

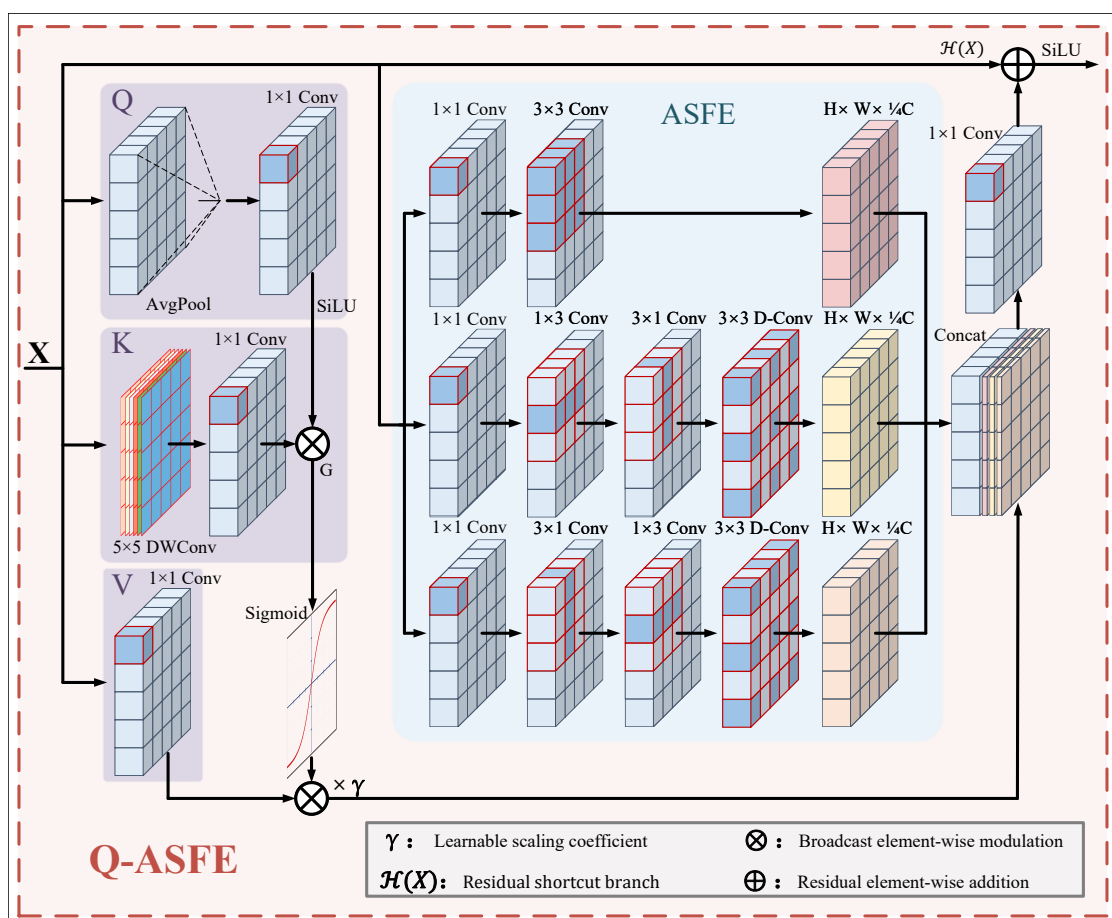


Figure 2. Structure of Q-ASFE. In the convolution schematics, the dark-blue cells highlighted by the red outlines indicate the local feature regions covered by the corresponding convolution kernels, whereas the light-blue cells denote the remaining feature-map elements.

To accommodate diverse target morphologies, ASFE employs three parallel branches to extract complementary local features. Let the input feature tensor be denoted as $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$. In each local branch, a 1×1 convolution is first applied to reduce the channel dimension to $C_r = C/4$, thereby controlling the computational cost. Branch 0 consists of a local perception path formed by a 1×1 convolution followed by a 3×3 convolution, which is used to preserve basic local structural information. The other two branches adopt asymmetric convolution combinations with different operation orders to capture complementary direction-sensitive structural features [32]. At the end of each asymmetric branch, a 3×3 dilated convolution with dilation rate $d = 2$ is introduced to enlarge the receptive field and supplement local structural information over a broader spatial range [41]. The asymmetric 1×3 and 3×1 kernels are used to emphasize complementary horizontal and vertical structures, while $d = 2$ expands the local receptive field without introducing a larger dense kernel. Together with the channel reduction $C_r = C/4$, these settings balance directional sensitivity, receptive-field coverage, and computational cost. The computation process can be formulated as

$$\mathbf{F}_l = \text{CBS}_{3 \times 3}(\text{CBS}_{1 \times 1}^{(l)}(\mathbf{X})) \quad (1)$$

$$\mathbf{F}_x = \text{CBS}_{3 \times 3, d=2}(\text{CBS}_{3 \times 1}(\text{CBS}_{1 \times 3}(\text{CBS}_{1 \times 1}^{(x)}(\mathbf{X})))) \quad (2)$$

$$\mathbf{F}_y = \text{CBS}_{3 \times 3, d=2}(\text{CBS}_{1 \times 3}(\text{CBS}_{3 \times 1}(\text{CBS}_{1 \times 1}^{(y)}(\mathbf{X})))) \quad (3)$$

where $\text{CBS}_{k_1 \times k_2, d}(\cdot)$ denotes a composite convolution block consisting of convolution, BatchNorm, and SiLU activation, with a kernel size of $k_1 \times k_2$ and dilation rate d . $\text{CBS}_{1 \times 1}^{(l)}$, $\text{CBS}_{1 \times 1}^{(x)}$, and $\text{CBS}_{1 \times 1}^{(y)}$ denote the branch-specific channel compression mappings for the three local branches, respectively. The outputs of the three branches are concatenated along the channel dimension to form the local structural features

$$\mathbf{F}_{\text{ASFE}} = \text{Concat}(\mathbf{F}_l, \mathbf{F}_x, \mathbf{F}_y) \quad (4)$$

where $\mathbf{F}_{\text{ASFE}} \in \mathbb{R}^{\frac{3C}{4} \times H \times W}$.

Different from reparameterizable asymmetric structures such as ACNet [32], ASFE retains its directional paths during inference to preserve complementary anisotropic responses.

ASFE primarily addresses local structural preservation through direction-sensitive modeling; however, local enhancement alone does not explicitly distinguish target-related structures from surrounding background responses. Therefore, an efficient QKV branch is further introduced as a complementary contextual modulation mechanism to enhance cross-region contextual discrimination. This branch follows the three-path Query/Key/Value projection paradigm, but does not explicitly construct a pairwise correlation matrix among spatial positions, thereby avoiding the quadratic growth in correlation computation with respect to the number of spatial positions N in standard self-attention [25]. Specifically, the input feature is projected into a query tensor $\mathbf{Q}$, a key tensor $\mathbf{K}$, and a value tensor $\mathbf{V}$:

$$\mathbf{Q} = \delta(\text{Conv}_{1 \times 1}(\text{GAP}(\mathbf{X}))) \quad (5)$$

$$\mathbf{K} = \text{CBS}_{1 \times 1}(\text{CBS}_{5 \times 5}^{dw}(\mathbf{X})) \quad (6)$$

$$\mathbf{V} = \text{Conv}_{1 \times 1}(\mathbf{X}) \quad (7)$$

where $\mathbf{Q} \in \mathbb{R}^{C_r \times 1 \times 1}$ denotes the global prior vector obtained by global average pooling, $\mathbf{K} \in \mathbb{R}^{C_r \times H \times W}$ denotes the local contextual feature extracted by depthwise convolution and pointwise projection, and $\mathbf{V} \in \mathbb{R}^{C \times H \times W}$ denotes the linearly projected feature carrier. The 5×5 depthwise convolution is adopted to enlarge the local contextual coverage while

keeping the additional parameter and computational overhead limited. Here, $\delta(\cdot)$ denotes the SiLU activation function, $\text{Conv}_{k_1 \times k_2}(\cdot)$ denotes a linear convolutional mapping without explicit activation, and $\text{CBS}_{5 \times 5}^{dw}(\cdot)$ denotes a 5×5 depthwise convolutional mapping.

Subsequently, $\mathbf{Q}$ is used to perform cross-dimensional gating on $\mathbf{K}$, yielding a gated tensor

$$\mathbf{G} = \sigma(\mathbf{Q} \otimes \mathbf{K}) \quad (8)$$

where $\sigma(\cdot)$ denotes the Sigmoid function, and $\otimes$ denotes broadcast element-wise multiplication. Since $\mathbf{G}$ and $\mathbf{V}$ have different channel dimensions, $\mathbf{G}$ is replicated along the channel dimension, denoted as

$$\tilde{\mathbf{G}} = \text{Repeat}_r(\mathbf{G}) \quad r = \frac{C}{C_r} = 4 \quad (9)$$

On this basis, the features from the value branch are modulated by the gating mechanism to obtain the context-enhanced features

$$\mathbf{F}_{\text{qkv}} = \gamma \cdot (\mathbf{V} \otimes \tilde{\mathbf{G}}) \quad (10)$$

where γ is a learnable scaling coefficient, initialized to 0 in practice to mitigate perturbation to the backbone feature distribution at the early stage of training.

Finally, the local structural feature $\mathbf{F}_{\text{ASFE}}$ and the context-enhanced feature $\mathbf{F}_{\text{qkv}}$ are concatenated along the channel dimension, and the module output is obtained through a 1×1 fusion projection and a residual connection [42]:

$$\mathbf{F}_{\text{out}} = \delta\left(\text{Conv}_{1 \times 1}\left(\text{Concat}(\mathbf{F}_{\text{ASFE}}, \mathbf{F}_{\text{qkv}})\right) + \mathcal{H}(\mathbf{X})\right) \quad (11)$$

where $\mathcal{H}(\mathbf{X})$ denotes the residual shortcut branch. When the input and output dimensions are identical, $\mathcal{H}(\mathbf{X})$ is an identity mapping; otherwise, a 1×1 convolution is adopted for dimension alignment.

In summary, through the coordinated design of local directional modeling and lightweight context modulation, Q-ASFE enhances the structural responses related to tiny objects in shallow high-resolution features without introducing an explicit global attention matrix.

2.3. Asymmetric Coordinate–Semantic Decoupled Head (ACS-Head)

Although Q-ASFE enhances shallow features for tiny object representation, using a shared enhanced representation for both classification and regression does not explicitly account for their task-specific requirements. The representation requirements of classification and regression in the detection head are inherently different: the classification branch places greater emphasis on semantic selectivity for category discrimination, whereas the regression branch relies more heavily on coordinate-related information associated with target localization [34,36]. Motivated by this observation, we design an ACS-Head, whose overall architecture is illustrated in Figure 3. Given an input feature $\mathbf{F}_{in} \in \mathbb{R}^{C \times H \times W}$, ACS-Head adopts different attention modeling strategies in the classification and regression branches to better accommodate the distinct requirements of category recognition and geometric localization, respectively.

Accordingly, SE is adopted in the classification branch because channel recalibration emphasizes category-discriminative semantic responses, whereas CA is used in the regression branch because its coordinate-aware directional encoding better preserves the position-sensitive information required for localization. In the classification branch, the input feature is first preprocessed by a branch-specific 3×3 convolution to obtain

$$\mathbf{F}_{cls} = \text{CBS}_{3 \times 3}(\mathbf{F}_{in}) \quad (12)$$

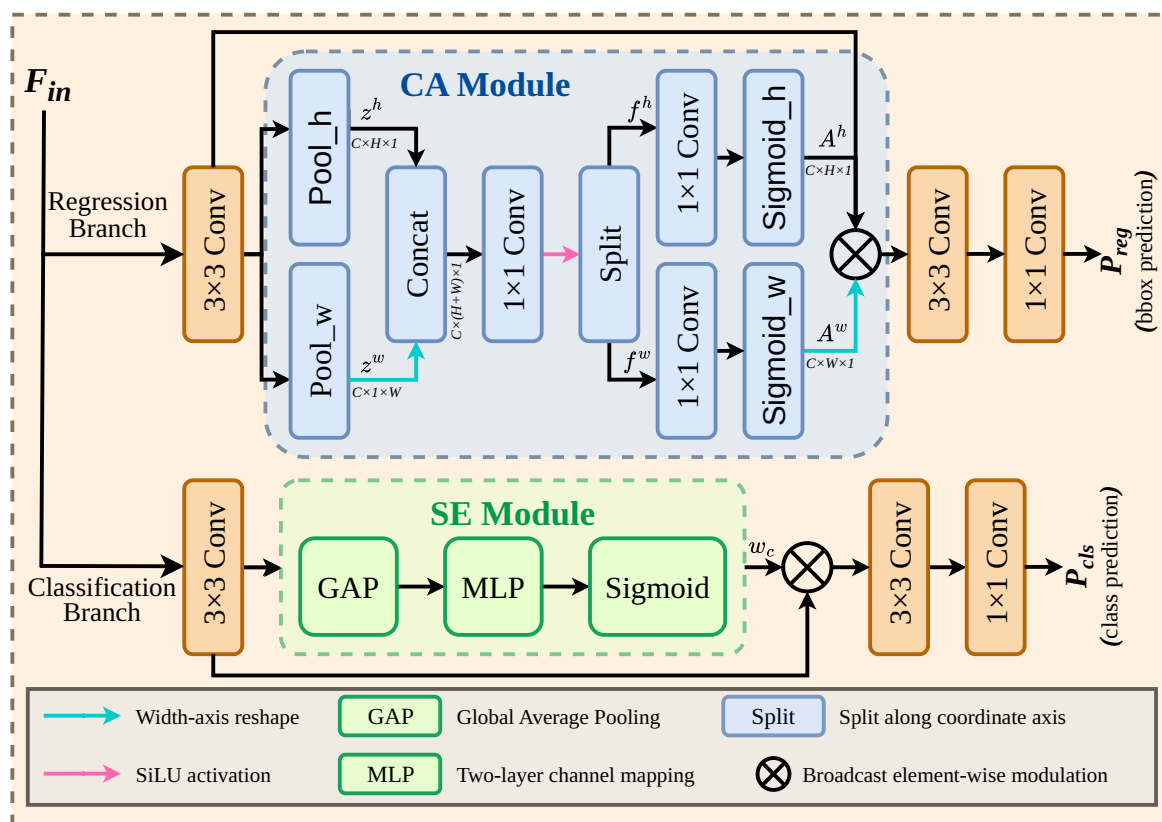


Figure 3. Structure of ACS-Head.

Subsequently, an SE module [22] is introduced to recalibrate the channel responses. Specifically, global average pooling is first applied to aggregate channel-wise semantic descriptors, after which a channel mapping is used to generate the weight vector $\mathbf{w}_c$, which is then applied to the original feature:

$$\mathbf{w}_c = \sigma(\text{MLP}(\text{GAP}(\mathbf{F}_{cls}))) \quad (13)$$

$$\tilde{\mathbf{F}}_{cls} = \mathbf{F}_{cls} \otimes \mathbf{w}_c \quad (14)$$

where $\text{MLP}(\cdot)$ denotes the two-layer channel mapping in the SE module, which is used to perform channel reduction and expansion. The channel-recalibrated classification feature is then passed through a 3×3 convolution and a 1×1 prediction projection to produce the classification output:

$$\mathbf{P}_{cls} = \text{Conv}_{1 \times 1}(\text{CBS}_{3 \times 3}(\tilde{\mathbf{F}}_{cls})) \quad (15)$$

In the regression branch, a branch-specific 3×3 convolution is likewise first applied to obtain the initial regression feature:

$$\mathbf{F}_{reg} = \text{CBS}_{3 \times 3}(\mathbf{F}_{in}) \quad (16)$$

Considering that bounding-box regression is more sensitive to coordinate information, this branch adopts the Coordinate Attention (CA) mechanism [24] to explicitly model

positional priors along the horizontal and vertical directions. For $\mathbf{F}_{reg} \in \mathbb{R}^{C \times H \times W}$, one-dimensional pooling is performed separately along the width and height dimensions

$$\mathbf{z}_c^h(h) = \frac{1}{W} \sum_{i=0}^{W-1} \mathbf{F}_{reg}(c, h, i) \quad (17)$$

$$\mathbf{z}_c^w(w) = \frac{1}{H} \sum_{j=0}^{H-1} \mathbf{F}_{reg}(c, j, w) \quad (18)$$

where H and W denote the height and width of the feature map, respectively; h and w represent the spatial coordinates along the vertical and horizontal directions, respectively; and c is the channel index of the feature tensor. Subsequently, the coordinate priors from the two directions are directionally aligned and concatenated along the spatial dimension, then encoded by a shared 1×1 transformation. The encoded feature is further split along the original dimensions to obtain two direction-aware intermediate features, $\mathbf{f}^h$ and $\mathbf{f}^w$:

$$\mathbf{f} = \delta\left(\text{Conv}_{1 \times 1}^{\text{share}}\left(\text{Concat}(\mathbf{z}^h, \mathbf{z}^w)\right)\right) \quad (19)$$

$$\mathbf{f}^h, \mathbf{f}^w = \text{Split}(\mathbf{f}) \quad (20)$$

where $\delta(\cdot)$ denotes the SiLU activation function. Thereafter, two independent 1×1 projections are employed to generate direction-aware attention weights:

$$\mathbf{A}^h = \sigma\left(\text{Conv}_{1 \times 1}^h(\mathbf{f}^h)\right) \quad (21)$$

$$\mathbf{A}^w = \sigma\left(\text{Conv}_{1 \times 1}^w(\mathbf{f}^w)\right) \quad (22)$$

Finally, the directional attention weights are used to modulate the regression features:

$$\tilde{\mathbf{F}}_{reg} = \mathbf{F}_{reg} \otimes \mathbf{A}^h \otimes \mathbf{A}^w \quad (23)$$

The regression output is then produced through a subsequent 3×3 convolution and a 1×1 prediction projection:

$$\mathbf{P}_{reg} = \text{Conv}_{1 \times 1}\left(\text{CBS}_{3 \times 3}(\tilde{\mathbf{F}}_{reg})\right) \quad (24)$$

It is worth noting that, to maintain consistency with the nonlinear forms used in the backbone and neck, the activation function inside CA is uniformly replaced with SiLU in this study, thereby reducing additional representational shifts introduced by activation inconsistency. In summary, ACS-Head achieves heterogeneous feature modeling for the classification and regression subtasks by strengthening channel-wise semantic selection in the classification branch while preserving direction-related coordinate information in the regression branch, thereby alleviating the representational conflict between the two tasks in a shared detection head.

3. Experiments and Results

3.1. Dataset

This study selects three datasets with complementary scene characteristics, namely VisDrone 2019, USOD, and RSOD, to evaluate the detection performance of the proposed model in three representative scenarios: complex urban aerial scenes from the UAV perspective, extremely small vehicle targets in wide-area airborne imagery, and medium-to-small rigid objects such as aircraft and oil tanks. The statistical information for these datasets is summarized in Table 1.

Table 1. Statistical summary of the datasets used in this study.

Dataset	Classes	Images (Train/Val)	Instances (Total/Small)	Small Ratio (%)	Target Characteristics
VisDrone 2019	10	6471/548	342,391/~205,434	~60.0%	Pedestrians & Vehicles, Dense, Scale Variation
USOD	1	2100/900	43,378/43,334	99.9%	Small Vehicles, Low-Illumination, Occlusion
RSOD	4	781/195	6950/~1042	~15.0%	Rigid Targets, Top-Down View

Note: Small objects are defined according to the MS COCO criterion, i.e., Area < 32 × 32 pixels.

(1) VisDrone 2019: Collected by UAV platforms in urban environments, this dataset contains a large number of pedestrian and non-motor vehicle instances. It is characterized by dense target distributions, substantial scale variations, and frequent occlusions, making it well suited for evaluating tiny object detection performance in complex scenes [4].

(2) USOD (Unicorn Small Object Dataset): Reconstructed by Zhang et al. [21] based on UNICORN2008, this dataset is primarily designed for extremely small vehicle detection. Tiny objects account for 99.9% of the instances, and the dataset also includes challenging imaging conditions such as low-illumination and shadow occlusion. USOD corresponds to a wide-area airborne imaging scenario and differs from low-altitude UAV imagery in both imaging perspective and scene characteristics.

(3) RSOD (Remote Sensing Object Detection Dataset): Released by Wuhan University, RSOD is an object detection dataset for high-resolution optical remote sensing imagery [43,44]. Compared with VisDrone and USOD, RSOD mainly contains rigid targets such as aircraft and oil tanks and can therefore serve as a complementary benchmark for evaluating the cross-scene generalization capability of the model on targets with regular geometric structures.

3.2. Implementation Details and Evaluation Metrics

All experiments in this study are implemented in PyTorch 2.9.0 with CUDA 12.8 and conducted on a single NVIDIA GeForce RTX 5090 GPU (NVIDIA Corporation, Santa Clara, CA, USA) under the Ubuntu 22.04.5 LTS. To ensure consistent settings across different datasets and to preserve the spatial details of tiny objects as much as possible, all input images are uniformly resized to 1024 × 1024 pixels during both training and inference. Unless otherwise specified, all models are trained on the same experimental platform using identical data augmentation and optimization settings.

For reproducibility, the training and validation batch sizes are set to 8 and 16, respectively, and the random seed is fixed to 42 for all experiments. The n-, m-, and x-scale models are initialized using the official YOLO11 weights from Ultralytics (version 8.3.63), pre-trained on COCO, namely, yolo11n.pt, yolo11m.pt, and yolo11x.pt, respectively. The main data augmentation strategies include Mosaic (disabled during the last 10 epochs), HSV augmentation, random translation and scaling, and horizontal flipping. Specifically, the HSV gains are set to 0.015, 0.7, and 0.4 for hue, saturation, and value, respectively; the translation and scaling factors are set to 0.1 and 0.5, and the horizontal-flip probability is 0.5. Low-probability appearance perturbations, including blur, median blur, grayscale conversion, CLAHE, and JPEG compression, are also applied during training. During validation and inference, the NMS IoU threshold is set to 0.7.

During training, the total number of epochs is set to 300. The optimizer is SGD with momentum, where the initial learning rate lr_0 is 0.01, the momentum is 0.937, and the

weight decay is 0.0005. A linear warmup strategy is adopted for the first 3 epochs, with the warmup momentum set to 0.8 to ensure stable convergence in the early stage of training.

Performance evaluation follows the standard MS COCO protocol [45], with mAP_{50} , $mAP_{50:95}$, and AP_S for small objects as the primary metrics. Specifically, mAP_{50} denotes the mean average precision at an IoU threshold of 0.5, while $mAP_{50:95}$ denotes the mean average precision averaged over IoU thresholds from 0.5 to 0.95 with a step size of 0.05. AP_S measures the detection performance on small-object instances. In this paper, the definition of small objects is consistent with that in MS COCO, namely, ground-truth bounding boxes with an area smaller than 32×32 pixels.

3.3. Comparison with State-of-the-Art Methods

To evaluate the overall performance and cross-scene generalization capability of A^2 -Det, this study compares it with a variety of representative detection methods on the VisDrone, USOD, and RSOD datasets. The compared models include n-scale YOLO variants, specialized detectors designed for tiny object detection, as well as typical two-stage and high-capacity detectors. The corresponding results are reported in Table 2, Table 3, and Table 4, respectively. It should be noted that, unless explicitly marked as the extended versions A^2 -Det-m and A^2 -Det-x, all references to A^2 -Det in this section denote the default n-scale variant built upon YOLO11n. In addition, to ensure a fair and comparable evaluation, all competing models involved in this section were retrained and evaluated by the authors under a unified training setting.

Table 2. Comparison with state-of-the-art methods on the VisDrone dataset.

Model	Precision (%) ↑	Recall (%) ↑	mAP_{50} (%) ↑	mAP_{75} (%) ↑	$mAP_{50:95}$ (%) ↑	AP_S (%) ↑	Params (M) ↓	GFLOPs ↓
YOLO11n	52.7	42.6	43.8	27.5	26.7	16.1	2.58	16.51
YOLOv8n	53.1	40.5	42.3	26.9	26.0	16.2	3.01	21.00
YOLOv9t	53.9	41.3	43.2	27.4	26.6	16.2	2.01	20.12
YOLOv10n	51.6	41.7	42.9	27.0	26.1	15.8	2.71	21.54
YOLOv12n	52.3	41.5	43.3	26.8	26.2	15.6	2.52	15.33
YOLOv13n	53.6	42.0	44.0	27.6	26.9	16.2	2.46	16.13
FFCA-YOLO	48.7	40.4	48.7	29.1	28.9	20.5	7.14	131.18
A^2-Det	57.8	47.4	49.6	32.5	31.0	20.7	1.60	34.58
TPH-YOLOv5	62.3	51.3	52.6	34.5	33.1	23.7	45.42	317.90
YOLOv13x	65.2	53.2	56.5	38.2	36.2	25.5	63.97	490.54
A^2-Det-m	64.6	55.9	56.3	39.2	37.0	26.9	25.23	530.08
A^2-Det-x	65.7	59.3	62.0	43.1	40.3	29.8	63.92	1326.01

Note: Bold values indicate the best performance in each column. ↑ denotes higher-is-better, and ↓ denotes lower-is-better.

Table 3. Comparison with state-of-the-art methods on the USOD dataset.

Model	AP_{50} (%) ↑	AP_{75} (%) ↑	$AP_{50:95}$ (%) ↑	AP_S (%) ↑	Params (M) ↓	GFLOPs ↓
YOLO11n	91.5	15.1	35.8	34.8	2.58	16.49
YOLOv13n	88.9	13.8	33.6	32.8	2.45	16.10
FFCA-YOLO	91.2	15.1	35.3	35.2	7.12	130.59
ATSS	89.7	14.1	34.4	34.4	32.11	206.00
GFL	88.6	14.8	34.5	34.5	32.26	209.00
Cascade R-CNN	84.1	14.5	33.0	33.0	69.15	239.00
DetectoRS + NWD	88.8	13.3	33.4	33.4	85.40	315.20
A^2-Det	92.6	16.2	36.4	35.5	1.60	34.51

Note: Bold values indicate the best performance in each column. ↑ denotes higher-is-better, and ↓ denotes lower-is-better.

(1) Evaluation on the VisDrone Scenario

VisDrone is primarily used to evaluate the capability of detection models in handling densely distributed tiny objects in complex urban environments. As shown in Table 2, among n-scale models, A^2 -Det achieves 49.6%, 31.0%, and 20.7% in mAP_{50} , $mAP_{50:95}$, and AP_S , respectively, outperforming the other compared models in this group. For high-capacity models, A^2 -Det-x reaches 62.0% in mAP_{50} . Compared with the classical method TPH-YOLOv5 [46], which performed strongly in the VisDrone2021 Challenge (52.6%), it achieves an improvement of nearly 10 percentage points, indicating that the proposed architecture has a strong performance potential in scenarios involving dense targets and substantial scale variation.

Table 4. Comparison with state-of-the-art methods on the RSOD dataset.

Model	mAP_{50} (%) ↑	mAP_{75} (%) ↑	$mAP_{50:95}$ (%) ↑	AP_S (%) ↑	Params (M) ↓	GFLOPs ↓
YOLO11n	97.4	76.9	65.2	40.0	2.58	16.49
YOLOv13n	97.0	78.2	65.1	39.2	2.45	16.10
FFCA-YOLO	98.0	73.9	62.8	43.9	7.13	130.79
ATSS	94.2	69.0	58.2	28.1	32.11	206.00
GFL	93.8	70.5	59.6	30.3	32.26	209.00
Cascade R-CNN	95.8	79.3	65.1	43.6	69.15	239.00
DetectoRS + NWD	95.5	78.8	65.6	42.6	85.40	315.20
A^2-Det	98.2	82.3	67.9	44.7	1.60	34.51

Note: Bold values indicate the best performance in each column. ↑ denotes higher-is-better, and ↓ denotes lower-is-better.

It should be noted that the improved detection accuracy of A^2 -Det is accompanied by increased computational complexity, particularly for the high-capacity A^2 -Det-x variant. The corresponding accuracy–efficiency trade-off and application scope are further discussed in Section 4.2.

For qualitative comparison, Figure 4 presents the detection results of several top-performing methods in terms of AP_S on the VisDrone dataset. Among the n-scale models, A^2 -Det yields more comprehensive detection results in regions with severe local occlusion and in areas containing distant tiny objects under high-altitude viewpoints. Among the high-capacity models, A^2 -Det-x also maintains high detection completeness in regions with heavily overlapping dense targets and pronounced scale variation. In particular, in the distant background region shown in the fourth column, the proposed method preserves more complete responses for remote tiny objects than the other compared models. This observation is consistent with the quantitative results in Table 2, further demonstrating the effectiveness of the proposed method for tiny object detection in complex urban UAV scenes.

(2) Generalization Evaluation on USOD and RSOD

USOD and RSOD provide two types of remote sensing imaging conditions that differ from those in VisDrone, namely, tiny vehicle targets under low-illumination and occlusion conditions, and rigid-object scenarios, respectively. Therefore, we further evaluate the cross-scene generalization capability of the proposed model on these two datasets. In this part, the comparison baselines include classical heavyweight two-stage detectors in the object detection literature, such as Cascade R-CNN [47]; one-stage models equipped with advanced label assignment strategies, such as ATSS [48] and GFL [49]; and DetectoRS [50], which incorporates recursive macro/micro feature modeling together with the tiny object-oriented metric NWD [5].

On USOD, A^2 -Det achieves the best results for AP_{50} , AP_{75} , $AP_{50:95}$, and AP_S in Table 3, indicating that the proposed method maintains strong detection performance even in scenarios involving extremely small targets and low-illumination. Notably, all models on

USOD exhibit a large gap between AP_{50} and AP_{75} . For example, YOLO11n attains an AP_{50} of 91.5%, whereas its AP_{75} is only 15.1%. This phenomenon is mainly attributable to the high IoU sensitivity of extremely tiny objects: when the target size is very limited, even a positional deviation of only 1–2 pixels in the predicted box can cause a substantial drop in IoU, making it difficult to satisfy the stricter AP_{75} criterion.

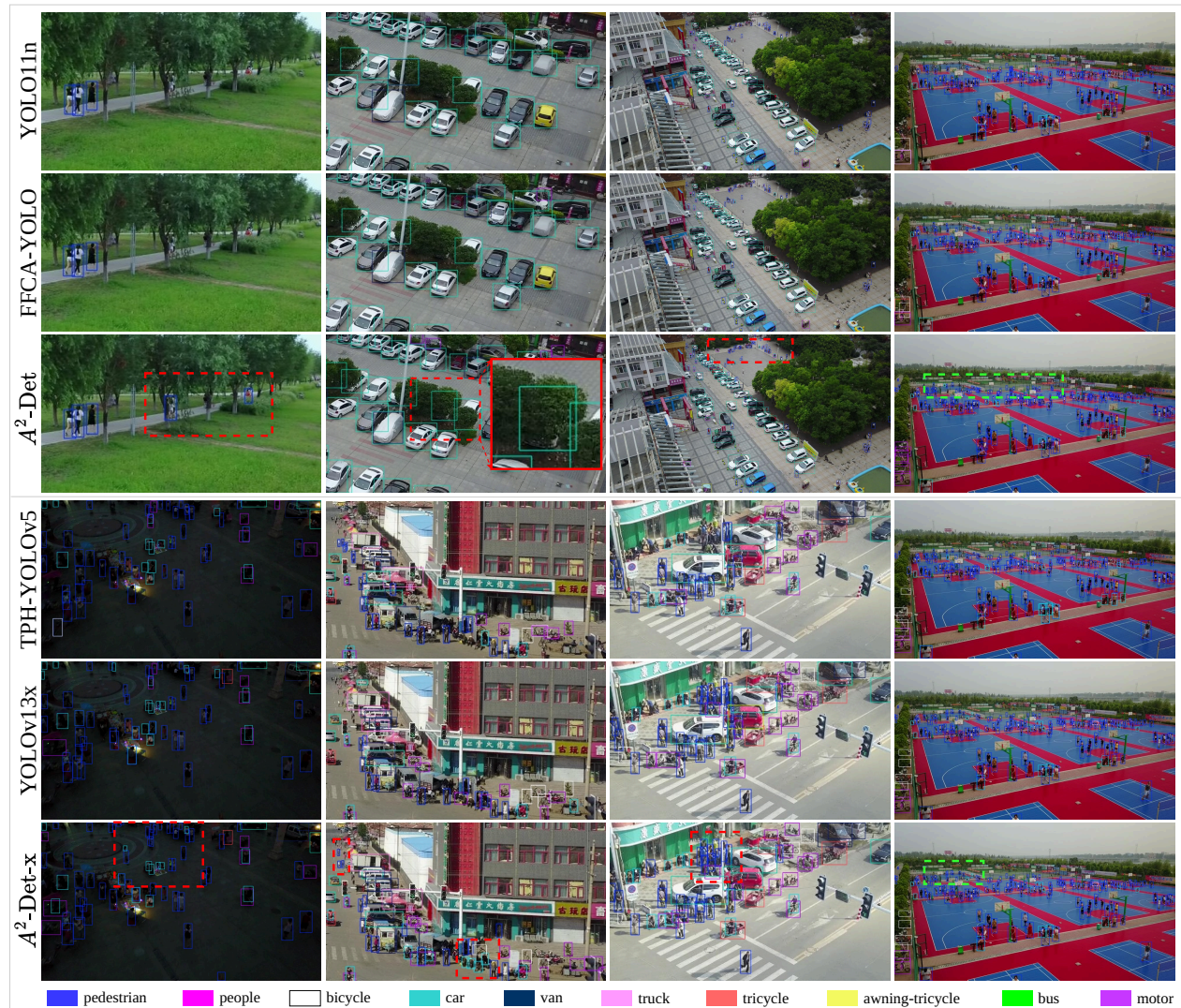


Figure 4. Qualitative comparison of detection results on the VisDrone dataset. Dashed boxes highlight representative regions for comparison.

To further provide a qualitative comparison, Figure 5 presents the detection results of the top-performing models in terms of AP_S on the USOD test set. The first two rows correspond to low-contrast examples, in which the compared methods generally exhibit weak responses to tiny and faint targets, leading to missed detections. In contrast, A^2 -Det preserves more complete target responses while maintaining relatively accurate localization. The third and fourth rows show dense-object scenarios. Although both FFCA-YOLO and A^2 -Det achieve high recall, A^2 -Det provides clearer separation between adjacent targets and more stable localization results. The last two rows compare false-positive cases. FFCA-YOLO produces more pronounced spurious responses in background regions, whereas A^2 -Det more effectively suppresses non-target activations. These observations indicate that A^2 -Det achieves a favorable balance between detection completeness and false-positive suppression in complex scenes, complementing the quantitative AP-based results in Table 3.

On RSOD, A^2 -Det likewise achieves the best overall performance. As shown in Table 4, it attains 98.2% in mAP_{50} and 44.7% in AP_S , demonstrating that the proposed method is not only effective for dense tiny and weak object scenarios, but also generalizes well to scenes dominated by rigid targets.

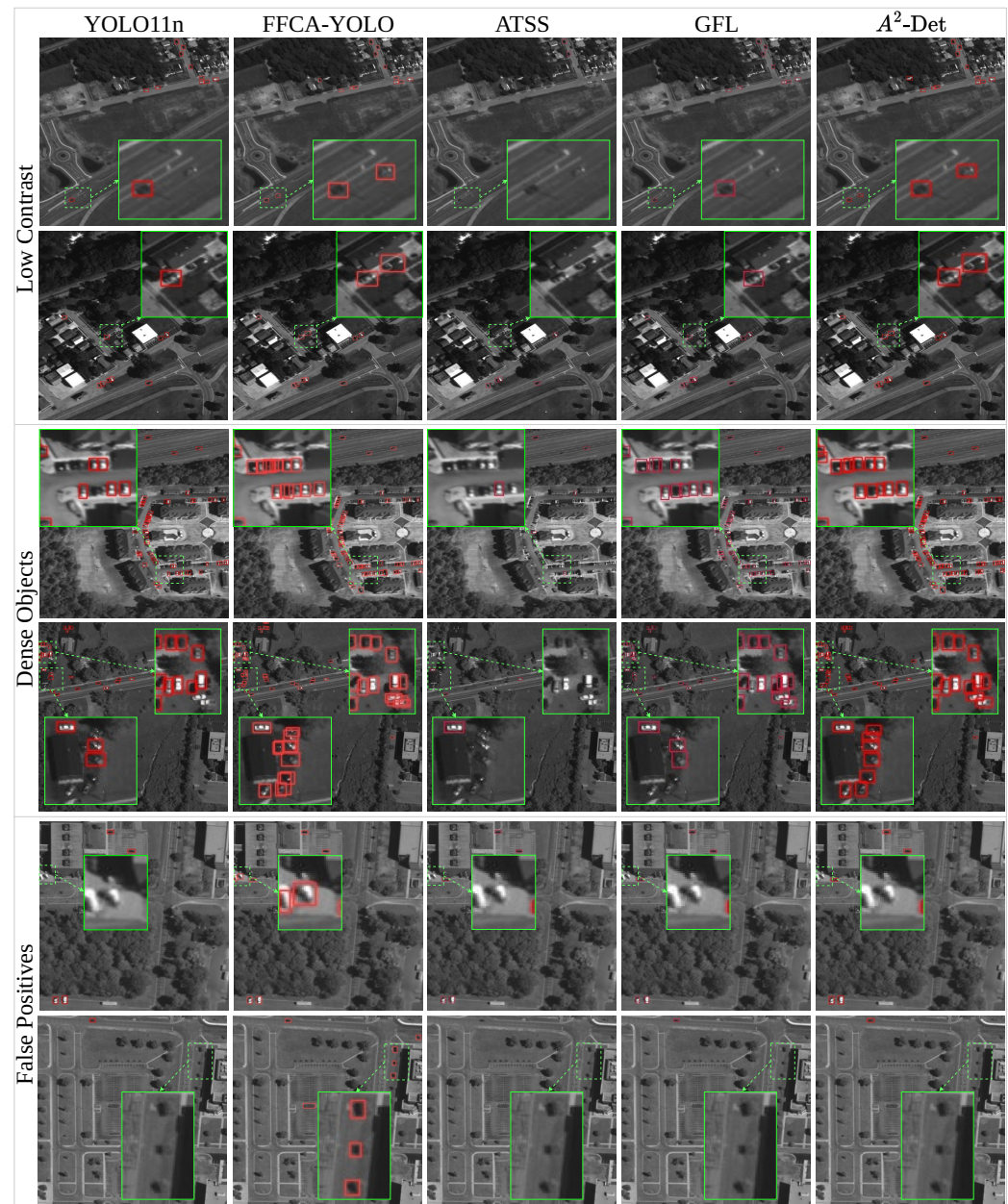


Figure 5. Qualitative comparison of the top-performing methods on the USOD dataset. The first two rows show low-contrast examples, the third and fourth rows show dense-object examples, and the fifth and sixth rows present typical false-positive cases. Red boxes denote detected objects, while green boxes indicate representative local regions highlighted for detailed comparison.

3.4. Module Comparison Results

(1) Comparative Evaluation of Feature Enhancement Modules

Under the unified YOLO11n (P2–P4) architecture, Q-ASFE is compared with several representative attention modules, including CBAM, CA, and EMA, as well as deformable convolution (DCNv3), to analyze its relative benefits [23,24,30,51]. The results are presented in Table 5. In addition to comparisons with external modules, the table also reports

isolated configurations of the two submodules, ASFE and QKV, for analyzing the individual contributions of each component within Q-ASFE.

Table 5 shows that Q-ASFE achieves the best results under the current experimental setting on both datasets. On VisDrone, Q-ASFE attains an AP_S of 20.5%, representing an improvement of 2.8 percentage points over the baseline; on USOD, its $mAP_{50:95}$ increases to 37.3%. Q-ASFE also consistently outperforms CBAM, CA, EMA, and DCNv3, indicating stable performance gains across different remote sensing tiny object scenarios. This observation is consistent with the analysis in Section 2, suggesting that regular directional feature modeling is more stable in such scenarios.

Table 5. Comparison of different feature enhancement modules on VisDrone and USOD under the P2–P4 architecture.

Module	VisDrone Dataset			USOD Dataset			Params * (M) ↓	GFLOPs * ↓
	mAP_{50} (%) ↑	$mAP_{50:95}$ (%) ↑	AP_S (%) ↑	AP_{50} (%) ↑	$AP_{50:95}$ (%) ↑	AP_S (%) ↑		
Baseline [†]	44.7	27.7	17.7	90.6	35.3	35.3	0.88	23.32
CBAM	45.9	28.4	18.6	92.0	36.1	35.0	0.89	23.34
CA	47.3	29.2	19.1	90.2	36.0	35.5	0.89	23.29
EMA	47.1	29.2	19.2	91.0	36.8	36.0	0.89	23.43
DCNv3	47.6	28.7	19.5	90.6	35.5	34.5	1.12	26.42
ASFE	48.1	30.1	19.9	92.1	37.0	36.2	0.98	28.20
QKV	46.4	29.0	18.2	92.2	37.0	36.0	0.94	26.22
Q-ASFE	48.9	30.7	20.5	92.4	37.3	36.4	1.02	30.58

* Note: Since all models use an input resolution of 1024×1024 , the differences in Params and GFLOPs caused by the detection output layers between VisDrone (10 classes) and USOD (1 class) are negligible; thus, the values are reported under the VisDrone setting. [†] Baseline denotes the task-oriented YOLO11n (P2–P4) architectural baseline used for module comparison; it differs from the original YOLO11n architecture, which employs the P3–P5 detection pyramid. Bold values indicate the best performance in each column. ↑ denotes higher-is-better, whereas ↓ denotes lower-is-better.

A further comparison between the isolated ASFE and QKV configurations shows that ASFE outperforms QKV across all evaluated metrics on VisDrone, whereas the two components exhibit comparable performance on USOD. Combining them into Q-ASFE yields further improvements on both datasets, indicating complementary performance contributions from the two submodules.

(2) Comparative Evaluation of Detection Heads

Under the same YOLO11n (P2–P4) architecture, we further compare ACS-Head with several representative detection heads to examine their performance differences. The compared methods can be grouped into four categories: (1) coupled heads and lightweight decoupled heads, including the native YOLO11n head and the dual-head structure of YOLOv10 [38]; (2) attention-enhanced heads, which introduce SE and CBAM modules into the detection head to strengthen feature representation [22,23]; (3) coordinate-aware heads, such as CoordAtt-Head, which explicitly models positional information based on the CoordAtt mechanism [24]; and (4) task-specialized heads, including custom comparison heads constructed according to representative existing design principles. To further analyze the effect of task-differentiated convolution design on detection head performance, we also implement a Task-Specific Large-Kernel Decoupled Head (TSK-Head) as an additional comparison target. This structure is not a standard published detection head, but is constructed by jointly drawing on the decoupled prediction framework of YOLOX [33], the analysis in TOOD regarding the different receptive-field requirements of classification and localization tasks [34], and a large-kernel depthwise convolution design. Specifically, the classification branch adopts convolutions with a larger receptive field to enhance contextual modeling, whereas the regression branch uses smaller kernels to preserve boundary

localization accuracy. Meanwhile, Rep-Head is introduced to examine the role of structural re-parameterization in detection head design [52]. The results are reported in Table 6.

As shown in Table 6, ACS-Head achieves the best or joint-best results on both datasets. On VisDrone, it reaches 29.6% in $mAP_{50:95}$ and 19.7% in AP_S , representing improvements of 1.9 and 2.0 percentage points over the baseline, respectively. On USOD, its $AP_{50:95}$ reaches 36.4%. Compared with SE-Head, CBAM-Head, CoordAtt-Head, and the other detection heads considered in this study, ACS-Head exhibits more consistent overall performance.

Table 6. Comparison of different detection heads on VisDrone and USOD under the P2–P4 architecture.

Head Type	Core Design	VisDrone Dataset			USOD Dataset			Params * (M) ↓	GFLOPs * ↓
		mAP_{50} (%) ↑	$mAP_{50:95}$ (%) ↑	AP_S (%) ↑	AP_{50} (%) ↑	$AP_{50:95}$ (%) ↑	AP_S (%) ↑		
Baseline [†]	-	44.7	27.7	17.7	90.6	35.3	35.3	0.88	23.32
YOLOv10-Head	Dual Head	41.3	25.7	16.5	90.0	35.5	34.5	1.21	35.36
SE-Head	Channel Attn	47.1	29.2	19.2	91.8	35.8	35.1	1.53	28.20
CBAM-Head	Mixed Attn	46.5	28.8	19.3	90.8	35.8	34.8	1.51	27.79
CoordAtt-Head	Coord Attn	46.1	28.6	18.7	91.9	36.3	35.3	0.97	26.47
TSLK-Head	Large Kernel	46.8	28.7	18.5	91.0	35.8	35.2	1.16	21.11
Rep-Head	Reparameter	46.6	28.8	18.6	91.6	35.3	35.4	1.17	32.32
ACS-Head	Heterogeneous	47.6	29.6	19.7	92.5	36.4	35.4	1.51	27.76

* Note: Params and GFLOPs are reported under the VisDrone setting. [†] Baseline denotes the task-oriented YOLO11n (P2–P4) architectural baseline used for module comparison; it differs from the original YOLO11n architecture, which employs the P3–P5 detection pyramid. TSLK-Head denotes a custom comparison head constructed in this work based on the decoupled-head paradigm and task-specific receptive-field design, rather than a previously published standard detector head. Bold values indicate the best performance in each column. ↑ denotes higher-is-better, whereas ↓ denotes lower-is-better.

A closer examination further shows that generic dual-head designs or convolutional enhancements are not necessarily well suited to the P2–P4 high-resolution remote sensing tiny object scenario considered in this study. YOLOv10-Head even performs worse than the baseline on VisDrone, suggesting that a general task-separation strategy without scene-specific adaptation cannot be directly translated into performance gains. Meanwhile, neither TSLK-Head nor Rep-Head yields superior results: although the former enlarges the receptive field, it may also introduce additional background interference; the latter mainly alters the convolution implementation form, but offers limited explicit adaptation to task-specific differences. Therefore, the advantage of ACS-Head does not lie merely in adding attention or modifying convolutional structures, but rather in its more effective adaptation to the distinct demands of classification and localization in dense remote sensing scenes with tiny objects.

3.5. Ablation Results

It should be noted that the P2–P4 configuration is not the original YOLO11n architecture, which adopts the P3–P5 detection pyramid. In this study, replacing the P5 branch with a higher-resolution P2 branch is treated as a task-oriented architectural modification specifically designed for tiny object detection, rather than as a parameter-tuning strategy. To quantify the contribution of the detection-pyramid redesign and verify the independent contributions of Q-ASFE and ACS-Head, as well as the effectiveness of their combination, ablation experiments were conducted on the VisDrone dataset, with the results reported in Table 7. Since the purpose of the ablation study is to quantitatively analyze the gains brought by the architectural and module-level designs, systematic quantitative comparisons are performed only on the representative VisDrone dataset. The USOD visualizations

introduced later are used solely as supplementary cross-scene evidence to show that the observed trends of enhanced feature responses and improved detection performance remain consistent in another type of remote sensing tiny object scenario. Specifically, considering that Q-ASFE mainly acts on fine-grained geometric information in high-resolution features, we first examine its deployment at different stages of the backbone under the P2–P4 architecture, in order to analyze the relationship between module effectiveness and feature hierarchy. Subsequently, with the backbone and neck configurations kept unchanged, we further evaluate the standalone gain of ACS-Head, as well as its overall effect when combined with the optimal Q-ASFE configuration.

Table 7. Ablation results of A^2 -Det on the VisDrone dataset.

No.	Model	Q-ASFE	ACS-Head	mAP ₅₀ (%) ↑	mAP _{50:95} (%) ↑	AP _S (%) ↑	Params (M) ↓	GFLOPs ↓
1	YOLO11n	–	–	43.8	26.7	16.1	2.58	16.51
2	YOLO11n (P2–P4)	–	–	44.7	27.7	17.7	0.88	23.32
3	YOLO11n (P2–P4)	✓ (P2)	–	46.5	28.9	19.3	0.91	27.05
4	YOLO11n (P2–P4)	✓ (P3)	–	45.9	28.5	18.5	0.99	26.84
5	YOLO11n (P2–P4)	✓ (P2, P3)	–	48.9	30.7	20.5	1.02	30.58
6	YOLO11n (P2–P4)	–	✓	47.6	29.6	19.7	1.51	27.76
7	A^2 -Det	✓ (P2, P3)	✓	49.6	31.0	20.7	1.60	34.58

Note: A check mark (✓) indicates that the corresponding module is included, whereas “–” indicates that it is not used. Bold values indicate the best performance in each column; ↑ denotes higher-is-better, whereas ↓ denotes lower-is-better.

First, from the perspective of the overall architecture, compared with the original YOLO11n baseline using the P3–P5 detection pyramid (No.1), the model adopting the P2–P4 detection pyramid without Q-ASFE or ACS-Head (No.2) achieves higher detection accuracy with fewer parameters. This direct comparison isolates the contribution of the detection-pyramid redesign itself, indicating that shifting the detection pyramid toward higher-resolution feature levels independently benefits tiny object detection. Experiments No.3–5 further show that the performance gain of Q-ASFE varies with its deployment level, with joint deployment at P2 and P3 achieving the best result and an AP_S of 20.5%.

To further verify the enhancement effect of Q-ASFE on shallow high-resolution features from the perspective of feature responses, Figure 6 presents visualizations at the Backbone P2 stage for representative examples from the VisDrone and USOD datasets. Overall, the original P2 features without enhancement exhibit relatively scattered responses, with evident non-target activations in complex background regions. After introducing Q-ASFE, the responses in target regions become more concentrated and continuous, while background interference is effectively suppressed. This trend is consistent across both datasets. Taking VisDrone as an example, for the distant tiny targets highlighted by the red dashed box in the first row of Figure 6a, the introduction of Q-ASFE produces clearer activations, whereas the baseline model shows relatively weak responses. In the second row, Q-ASFE more effectively suppresses non-target activations in background regions, making the target areas more prominent. For USOD, as shown in Figure 6b, the responses of background regions such as roads and buildings are likewise effectively suppressed, and the target-related structures become more distinct. These observations are consistent with the quantitative improvements brought by Q-ASFE in Table 7 on VisDrone. The similar response patterns observed on USOD serve as supplementary cross-scene visual evidence, further indicating that the enhancement effect of this module on shallow high-resolution features is not limited to a single dataset.

Under the fixed P2–P4 architecture, No.6 further verifies the independent contribution of ACS-Head. Compared with the standard detection head under the same backbone and neck configuration (No.2), replacing it with ACS-Head improves mAP_{50} , $mAP_{50:95}$, and AP_5 by 2.9, 1.9, and 2.0 percentage points, respectively. These results confirm the independent performance contribution of ACS-Head under the current P2–P4 architecture. Compared with No.5, further introducing ACS-Head yields an additional performance improvement. Relative to the original baseline (No.1), the final model improves mAP_{50} by 5.8 percentage points, validating the effectiveness of the joint Q-ASFE and ACS-Head configuration. Taken together, No.1, No.2, and No.7 provide a progressive comparison from the original P3–P5 detection pyramid, through the P2–P4 pyramid without the proposed modules, to the complete A^2 -Det, thereby distinguishing the independent contribution of the pyramid redesign from the additional gains brought by Q-ASFE and ACS-Head.

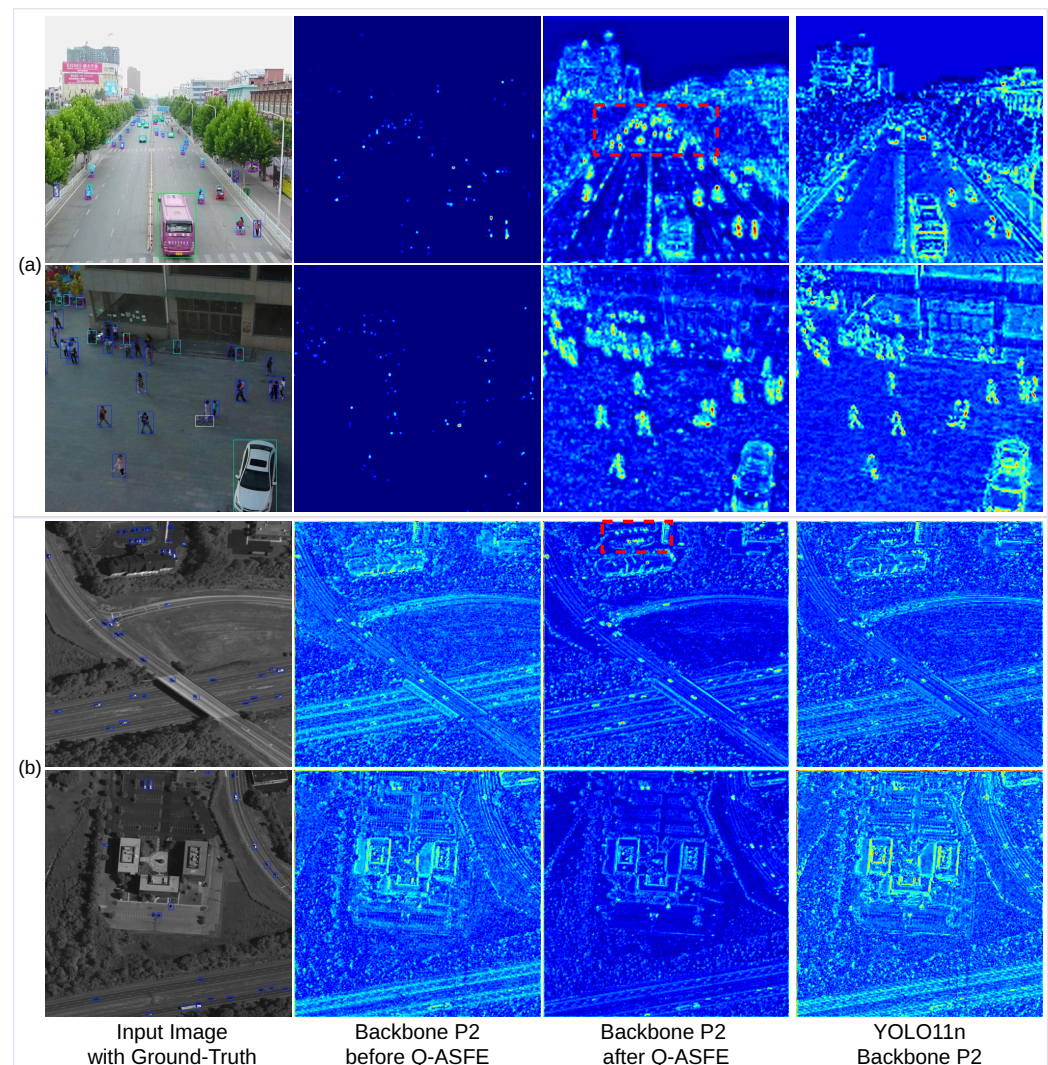


Figure 6. Feature visualization of backbone P2 responses with and without Q-ASFE on representative samples from (a) VisDrone and (b) USOD. Red dashed boxes highlight representative regions for comparison, and the heat-map colors qualitatively indicate relative feature-response intensity, with warmer colors corresponding to stronger responses.

To illustrate the practical effects of the progressive improvements at the detection-result level, Figure 7 presents qualitative comparisons of No.1, No.5, and No.7 on representative samples from VisDrone and USOD. Overall, progressing from the baseline to Q-ASFE and subsequently to A^2 -Det with ACS-Head gradually reduces missed detections while

improving the discrimination of densely distributed targets and bounding-box localization quality. These qualitative observations are consistent with the performance trends reported in Table 7.

3.6. Size-Stratified Tiny Object Evaluation

To further examine the detection capability of A^2 -Det across different tiny object scales, a size-stratified evaluation is conducted on VisDrone and USOD. Ground-truth instances within the small-object range are divided into two groups according to their bounding-box area A : extremely tiny objects with $A < 16^2$ and relatively larger tiny objects with $16^2 \leq A < 32^2$. YOLO11n is adopted as the baseline to evaluate the scale-specific performance gains of A^2 -Det-n.

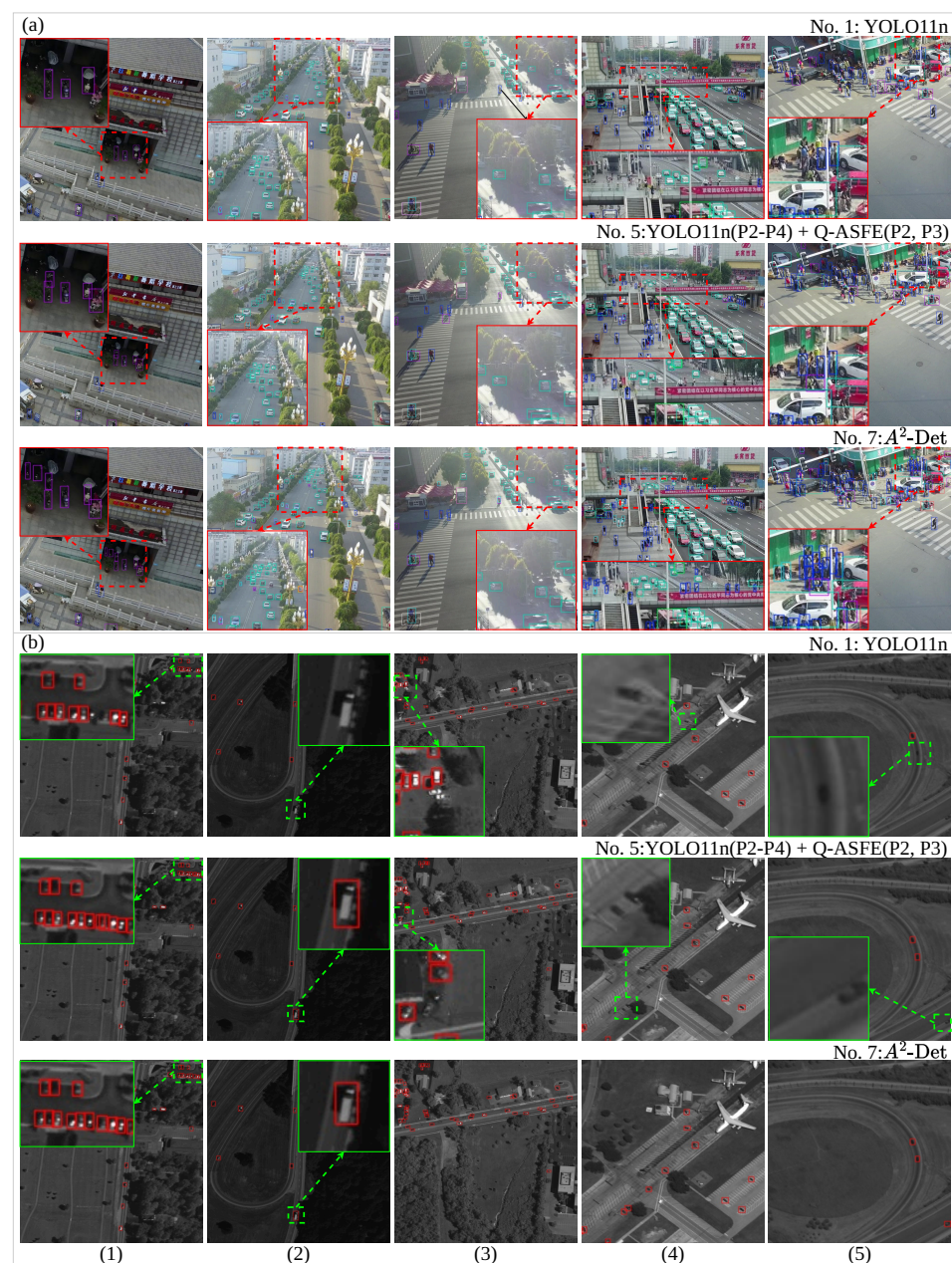


Figure 7. Progressive qualitative comparison of ablation results on representative samples from In Proceedings of the (a) VisDrone and (b) USOD. Red dashed boxes in VisDrone and green highlighted boxes in USOD indicate representative local regions selected for comparison. The labels No. 1, No. 5, and No. 7 correspond to the ablation configurations defined in Table 7, while (1)–(5) denote different representative samples.

As shown in Table 8, detection performance generally decreases as the target scale becomes smaller, particularly on VisDrone, reflecting the greater difficulty of detecting extremely tiny objects. Nevertheless, A^2 -Det-n consistently outperforms YOLO11n across both size ranges. For VisDrone objects with $A < 16^2$, A^2 -Det-n improves AP_{50} , AP_{75} , and $mAP_{50:95}$ by 8.7, 3.4, and 4.3 percentage points, respectively, with consistent gains also observed for the $16^2 \leq A < 32^2$ group. On USOD, where the baseline AP_{50} is already relatively high, the improvement in AP_{50} is comparatively modest, whereas AP_{75} shows clearer gains across both size ranges. Overall, these results suggest that A^2 -Det maintains favorable detection performance as the target scale decreases and are consistent with the intended design objective of preserving fine-grained information for tiny object detection.

Table 8. Size-stratified detection performance on the VisDrone and USOD datasets.

Dataset	Size Range	GT Instances	Model	AP_{50} (%) $\uparrow$	AP_{75} (%) $\uparrow$	$mAP_{50:95}$ (%) $\uparrow$
VisDrone	$A < 16^2$	8000	YOLO11n	17.4	4.1	7.0
VisDrone	$A < 16^2$	8000	A^2 -Det-n	26.1 (50.0%)	7.5 (82.9%)	11.3 (61.4%)
VisDrone	$16^2 \leq A < 32^2$	8000	YOLO11n	34.3	16.9	18.0
VisDrone	$16^2 \leq A < 32^2$	8000	A^2 -Det-n	43.4 (26.5%)	23.1 (36.7%)	23.5 (30.6%)
USOD	$A < 16^2$	6008	YOLO11n	90.0	12.9	34.0
USOD	$A < 16^2$	6008	A^2 -Det-n	92.8 (3.1%)	15.3 (18.6%)	35.7 (5.0%)
USOD	$16^2 \leq A < 32^2$	882	YOLO11n	94.4	19.8	39.3
USOD	$16^2 \leq A < 32^2$	882	A^2 -Det-n	95.0 (0.6%)	23.0 (16.2%)	40.5 (3.1%)

Note: A denotes the ground-truth bounding-box area. For each dataset and size range, a maximum of 8000 ground-truth instances is evaluated; when fewer than 8000 instances are available, all available instances are used. Values in parentheses indicate the relative improvement of A^2 -Det-n over YOLO11n. Bold values indicate the best performance in each column. $\uparrow$ denotes higher-is-better, whereas $\downarrow$ denotes lower-is-better.

3.7. Practical Inference Efficiency

Practical inference efficiency was evaluated under the same hardware platform described in Section 3.2, using a single NVIDIA GeForce RTX 5090 GPU. Single-image latency and frames per second (FPS) were measured under unified inference settings, with all models evaluated at an input resolution of 1024×1024 . As shown in Table 9, A^2 -Det-n was compared with YOLO11n on VisDrone, USOD, and RSOD, while A^2 -Det-x and YOLO11x were additionally evaluated on VisDrone to examine inference efficiency across different model scales.

Table 9. Practical inference efficiency of YOLO11 and A^2 -Det on different datasets.

Dataset	Model	Params (M) $\downarrow$	GFLOPs $\downarrow$	Latency (ms/Image) $\downarrow$	FPS $\uparrow$
VisDrone	YOLO11n	2.58	16.51	5.44	183.73
	A^2 -Det-n	1.60	34.58	8.12	123.13
	YOLO11x	56.89	500.50	27.81	35.96
	A^2 -Det-x	63.92	1326.01	56.55	17.68
USOD	YOLO11n	2.58	16.49	5.48	182.44
	A^2 -Det-n	1.60	34.51	8.12	123.13
RSOD	YOLO11n	2.58	16.49	5.66	176.57
	A^2 -Det-n	1.60	34.51	8.33	120.11

Note: $\uparrow$ denotes higher-is-better, whereas $\downarrow$ denotes lower-is-better.

The results show a consistent inference-efficiency trend for A^2 -Det-n across the three datasets. Compared with YOLO11n, A^2 -Det-n exhibits increased latency and reduced FPS, while the variations across datasets remain relatively small, indicating stable practical

inference behavior. At a larger model scale, the efficiency gap between A^2 -Det-x and YOLO11x becomes more pronounced, suggesting that the additional computational cost of A^2 -Det increases with model capacity.

3.8. Robustness Results

To evaluate the robustness of the proposed model under adverse imaging conditions, four common types of perturbations were introduced on the test set, including image blurring, Gaussian noise [53], stripe noise [54], and fog [55,56]. The degradation severity was quantified using the peak signal-to-noise ratio (PSNR). Figure 8 presents visual examples of degraded images under different perturbation conditions on the RSOD dataset together with the corresponding detection results, while Table 10 summarizes the robustness evaluation results of A^2 -Det and the baseline YOLO11n on RSOD and VisDrone. Considering that the targets in RSOD generally exhibit relatively clear rigid structures, making the degradation process easier to observe, the performance degradation curves on RSOD are further provided in Figure 9.

Table 10 and Figure 9 show that different perturbation types affect model performance in distinct ways. Under image blurring and fog, A^2 -Det maintains relatively high detection performance overall. Under Gaussian noise, both models degrade progressively as the noise intensity increases. At the strongest tested level of $\sigma^2 = 0.03$, A^2 -Det and YOLO11n achieve mAP₅₀ values of 76.0% and 81.3% on RSOD, corresponding to relative declines of 22.6% and 16.5%, respectively. On VisDrone, the corresponding mAP₅₀ values decrease to 13.5% and 11.5%, with comparable relative declines of 72.8% and 73.7%. These results indicate that the sensitivity to severe Gaussian noise is dataset dependent. Under stripe noise, performance also decreases with increasing perturbation strength. Although the asymmetric 1×3 and 3×1 convolutions may respond to directional stripe patterns, A^2 -Det does not exhibit greater degradation than YOLO11n in the present experiments. At $r = 0.2$, it retains both higher absolute mAP₅₀ and smaller relative declines on RSOD and VisDrone.

As shown in Figure 8, different perturbations produce distinct forms of degradation in target appearance. Image blurring and fog mainly weaken target edges and reduce contrast, Gaussian noise introduces abundant discrete perturbations in local regions, whereas stripe noise produces pronounced directional structural interference. These visual degradation patterns are broadly consistent with the performance trends in Table 10, indicating that A^2 -Det remains relatively stable under smooth degradation conditions such as image blurring and fog, but still exhibits a certain degree of sensitivity to strong random noise.

Table 10. Robustness comparison of A^2 -Det and YOLO11n on RSOD and VisDrone in terms of mAP₅₀.

Interference	Intensity	RSOD Dataset			VisDrone Dataset		
		PSNR (dB)	A^2 -Det	YOLO11n	PSNR (dB)	A^2 -Det	YOLO11n
Clean (Baseline)	—	∞	98.2	97.4	∞	49.6	43.8
Image Blurring	$w = 0.81$	34.45	98.2 (0.0%)	97.1 (0.3% ↓)	33.44	47.2 (4.8% ↓)	41.6 (5.0% ↓)
	$w = 0.64$	32.81	98.2 (0.0%)	96.8 (0.6% ↓)	31.71	46.6 (6.0% ↓)	40.9 (6.6% ↓)
	$w = 0.49$	31.09	97.9 (0.3% ↓)	96.2 (1.2% ↓)	29.83	45.1 (9.1% ↓)	39.5 (9.8% ↓)
	$w = 0.39$	29.66	97.3 (0.9% ↓)	96.3 (1.1% ↓)	28.28	42.0 (15.3% ↓)	37.6 (14.2% ↓)
	$w = 0.25$	27.22	92.5 (5.8% ↓)	92.2 (5.3% ↓)	25.89	32.2 (35.1% ↓)	29.8 (32.0% ↓)
Gaussian Noise	$\sigma^2 = 0.001$	30.00	98.2 (0.0%)	97.0 (0.4% ↓)	30.02	44.7 (9.9% ↓)	41.6 (5.0% ↓)
	$\sigma^2 = 0.005$	23.06	95.4 (2.9% ↓)	95.7 (1.7% ↓)	23.14	35.1 (29.2% ↓)	32.7 (25.3% ↓)
	$\sigma^2 = 0.01$	20.10	89.5 (8.9% ↓)	94.5 (3.0% ↓)	20.21	27.6 (44.4% ↓)	24.8 (43.4% ↓)
	$\sigma^2 = 0.02$	17.21	83.1 (15.4% ↓)	87.0 (10.7% ↓)	17.35	19.8 (60.1% ↓)	16.9 (61.4% ↓)
	$\sigma^2 = 0.03$	15.59	76.0 (22.6% ↓)	81.3 (16.5% ↓)	15.72	13.5 (72.8% ↓)	11.5 (73.7% ↓)

Table 10. Cont.

Interference	Intensity	RSOD Dataset			VisDrone Dataset		
		PSNR (dB)	A ² -Det	YOLO11n	PSNR (dB)	A ² -Det	YOLO11n
Stripe Noise	$r = 0.05$	21.89	90.2 (8.1% ↓)	91.5 (6.1% ↓)	21.94	38.4 (22.6% ↓)	32.9 (24.9% ↓)
	$r = 0.1$	18.84	87.5 (10.9% ↓)	84.7 (13.0% ↓)	18.93	28.9 (41.7% ↓)	23.4 (46.6% ↓)
	$r = 0.2$	15.80	68.7 (30.0% ↓)	62.9 (35.4% ↓)	15.90	12.7 (74.4% ↓)	8.63 (80.3% ↓)
Fog	$A = 0.2$	19.58	98.2 (0.0%)	97.0 (0.4% ↓)	18.91	46.2 (6.9% ↓)	40.8 (6.8% ↓)
	$A = 0.3$	16.02	98.3 (0.1% ↑)	96.8 (0.6% ↓)	15.36	45.6 (8.1% ↓)	39.6 (9.6% ↓)
	$A = 0.4$	13.49	98.2 (0.0%)	96.8 (0.6% ↓)	12.83	43.9 (11.5% ↓)	38.0 (13.2% ↓)
	$A = 0.5$	11.52	98.1 (0.1% ↓)	96.0 (1.4% ↓)	10.86	41.9 (15.5% ↓)	36.0 (17.8% ↓)

Note: Values in parentheses indicate the relative change in mAP₅₀ compared with the clean baseline. ↑ and ↓ denote relative increases and decreases, respectively.

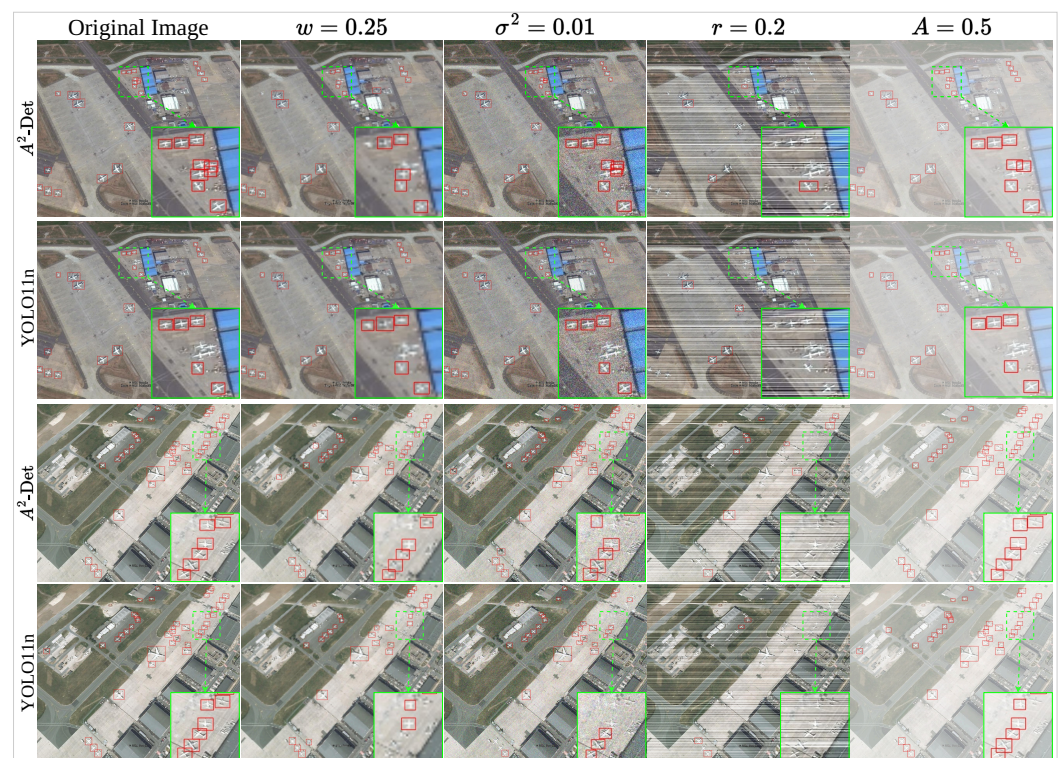


Figure 8. Visual examples of degraded RSOD images under different perturbations and the corresponding detection results. Red boxes denote detected objects, green boxes highlight representative local regions, and green arrows point to the corresponding magnified regions.

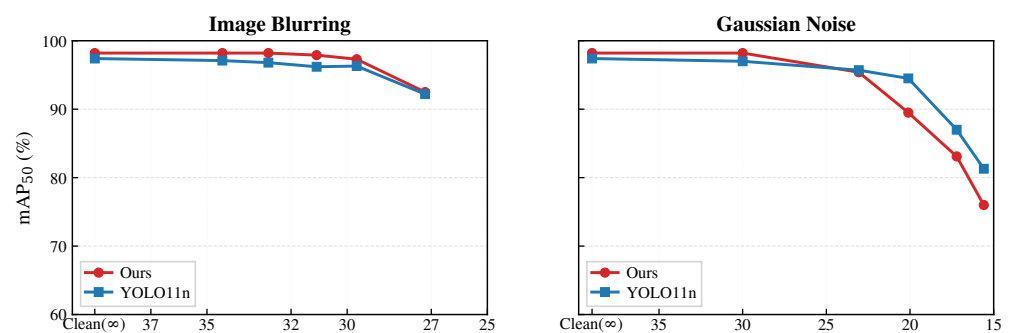


Figure 9. Cont.

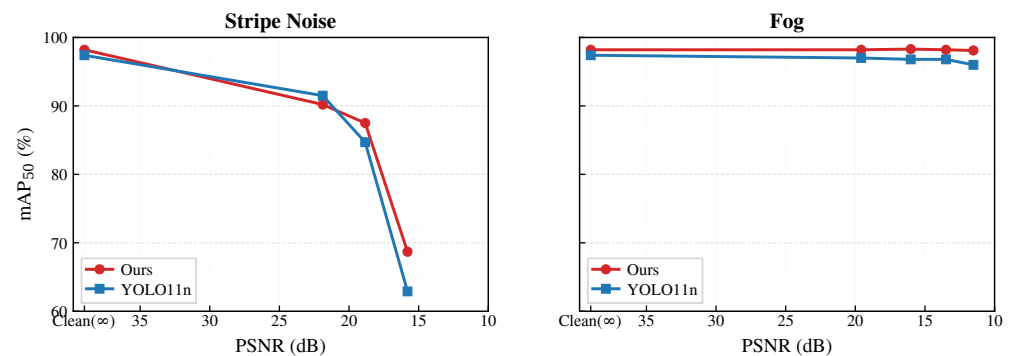


Figure 9. Performance degradation curves on the RSOD dataset under different perturbations.

4. Discussion

4.1. Mechanism of the Dual Asymmetric Design

The performance gains of A^2 -Det arise from two asymmetric mechanisms operating at the feature representation and task prediction stages, respectively, each tailored to distinct information requirements. Remote sensing tiny objects typically occupy only a few pixels, and their effective representation relies heavily on limited edge, contour, and local intensity variations. With successive downsampling, such fine-grained structures are readily diluted by spatial compression and background responses, while deep semantic features cannot recover spatial information once it has been lost. Accordingly, shifting the detection levels toward P2–P4 and enhancing shallow high-resolution features is not intended merely to increase the contribution of shallow features; rather, it strengthens target-related cues before fine-grained structures undergo irreversible attenuation, thereby providing more discriminative tiny object representations for subsequent cross-scale fusion.

Q-ASFE further improves shallow feature quality through direction-sensitive local structural modeling and context modulation. ASFE employs convolutional paths with different directional biases to reinforce horizontal, vertical, and local isotropic structures, thereby preserving the limited edge and contour cues of tiny objects during feature propagation. Unlike reparameterizable asymmetric structures that are folded into standard convolutions after training, ASFE retains its directional paths during inference to preserve complementary anisotropic spatial responses. However, strengthening local structures alone may also amplify background interference from road textures, building edges, and other clutter. Efficient QKV therefore conditions local spatial responses on global semantic information, enhancing target-related structures while suppressing background activations lacking semantic support. In this way, ASFE primarily preserves structural information, whereas efficient QKV performs context-guided response selection, making the two components complementary at the feature representation level.

The asymmetric design of ACS-Head is motivated by the distinct representation requirements of classification and regression. Classification relies more heavily on channel-wise semantic information with strong category discriminability and generally benefits from a certain degree of invariance to local positional perturbations, whereas regression depends directly on spatial cues such as object centers, boundaries, and scales, and is therefore more sensitive to positional variations. Applying an identical enhancement strategy to both branches may thus be insufficient to accommodate these task-specific requirements. Accordingly, ACS-Head strengthens channel-wise semantic selection in the classification branch to emphasize category-relevant information and suppress background redundancy, while preserving richer spatial and coordinate information in the regression branch to support accurate localization of tiny objects. This design exploits the distinct

inductive biases of different feature enhancement mechanisms, allowing each branch to acquire representations better matched to its task characteristics.

Overall, the two asymmetric mechanisms in A^2 -Det address information bottlenecks at different stages. Q-ASFE primarily mitigates the attenuation of fine-grained information and background interference in shallow high-resolution features, whereas ACS-Head performs task-specific modeling to alleviate the representational inconsistency between classification and regression. The former improves the quality of features delivered to the detection head, while the latter further optimizes how these features are utilized by the two prediction tasks, thereby forming a complementary processing chain from feature enhancement to task prediction. The module comparisons, progressive ablation studies, and feature visualizations presented in Section 3 are consistent with this mechanism-level interpretation and provide empirical support for the effectiveness and complementarity of the dual asymmetric design.

4.2. Accuracy–Efficiency Trade-Off and Application Scope

A^2 -Det improves tiny object detection performance at the cost of additional computation. This increase mainly stems from shifting the detection focus from deeper features toward the P2–P4 levels and introducing Q-ASFE on the high-resolution P2 and P3 feature maps. Because shallow feature maps have larger spatial dimensions, feature enhancement at these levels inevitably incurs greater computational cost. Meanwhile, removing the P5 branch and constraining the channel dimensions of the added modules limit parameter growth, resulting in a model with a relatively small parameter count but higher GFLOPs.

This computational profile reflects an explicit trade-off between detection accuracy and computational efficiency. Since the informative cues of remote sensing tiny objects are largely concentrated in limited spatial details, allocating more computation to high-resolution representations is well aligned with the task requirements. The experimental results show that this redistribution of computation consistently improves tiny object detection, indicating that the additional cost is primarily devoted to preserving and enhancing fine-grained features rather than simply increasing model capacity. A^2 -Det is therefore better characterized as a parameter-efficient, accuracy-oriented framework for tiny object detection, rather than an extremely lightweight model optimized for minimum computational cost.

From an application perspective, the default A^2 -Det-n is primarily intended for accuracy-oriented remote sensing and UAV scenarios where sufficient computational resources are available and accurate detection of distant and densely distributed tiny objects is prioritized. The practical efficiency results in Table 9 further show that A^2 -Det-n maintains an inference speed above 120 FPS on the RTX 5090 across the three evaluated datasets. On VisDrone, compared with YOLO11n, A^2 -Det-n improves mAP₅₀ and AP_S by 5.8 and 4.6 percentage points, respectively, while increasing the measured inference latency by approximately 49%. Meanwhile, although its GFLOPs are approximately 2.09 times those of YOLO11n, the increase in practical latency is substantially smaller, indicating that theoretical computational complexity does not translate linearly into actual inference time. As the model scale increases, the efficiency penalty becomes more pronounced; therefore, A^2 -Det-m and A^2 -Det-x are better regarded as high-capacity, performance-oriented configurations for computationally sufficient or offline processing scenarios. The deployment scenario considered in this study primarily involves off-board processing, where remotely sensed imagery acquired by satellite or aerial platforms is transmitted to ground stations or data centers for subsequent detection and analysis. Accordingly, the current practical efficiency evaluation focuses on desktop/server-class GPU platforms rather than real-time

onboard inference on resource-constrained devices. Deployment efficiency on dedicated embedded or edge-computing platforms remains to be further evaluated in future work.

4.3. Robustness, Limitations, and Future Directions

Different types of image degradation affect A^2 -Det in distinct ways because they disrupt target structural information and background characteristics through different mechanisms. Image blurring and fog primarily attenuate local high-frequency details and reduce contrast, while the overall spatial structure of targets and the contextual semantics of the scene remain relatively coherent. Under such conditions, the local structural modeling and context modulation in Q-ASFE can preserve target-related responses despite moderate detail degradation, resulting in relatively stable detection performance. The robustness results reported in the experimental section are generally consistent with this mechanism-level interpretation.

In contrast, random and structured noise interfere with the fine-grained representation of tiny objects through different mechanisms. As illustrated by the enlarged regions in Figure 8, Gaussian noise introduces dense pixel-level fluctuations over both targets and surrounding backgrounds, which obscure the already weak contours and local intensity differences of tiny objects. Since A^2 -Det shifts detection toward the high-resolution P2–P4 levels and enhances the P2 and P3 features before substantial spatial downsampling, this design preserves weak target details but also exposes the shallow representation to more high-frequency perturbations. Consequently, under severe Gaussian noise, random responses may become increasingly difficult to distinguish from genuine fine-grained target structures. The asymmetric directional branches rely on locally coherent structural patterns; random perturbations disrupt this local consistency and thereby reduce the discriminability of weak target edges. This provides a plausible explanation for the increased sensitivity observed under severe Gaussian noise, particularly on RSOD.

Stripe noise exhibits a different failure mechanism. As shown in Figure 8, strong stripe interference forms spatially continuous and highly directional patterns across both target and background regions. Because the 1×3 and 3×1 paths in ASFE are designed to capture complementary directional structures, such artificial directional patterns may compete with genuine target edges in the feature representation. As stripe intensity increases, the distinction between target-related directional responses and structured interference therefore becomes more difficult, which helps explain the substantial degradation observed under strong stripe noise.

These observations indicate that one robustness limitation of A^2 -Det lies in the trade-off between fine-grained structural enhancement and noise suppression. The current Q-ASFE is primarily designed to strengthen target-related structures and does not explicitly distinguish informative high-frequency details from random noise components. Future work may therefore incorporate frequency-domain awareness or adaptive denoising mechanisms to suppress noise responses while preserving the limited edge and contour cues of tiny objects. For practical remote sensing imaging systems, preprocessing strategies such as image denoising or non-uniformity correction may also be integrated to further improve detection stability under complex degradation conditions.

5. Conclusions

This study proposes A^2 -Det, a dual asymmetric framework for tiny object detection in remote sensing imagery. By restructuring the detection pyramid toward high-resolution levels and jointly incorporating Q-ASFE and ACS-Head, the proposed framework improves fine-grained feature representation and task-specific prediction for tiny objects. Experimental results show that A^2 -Det improves mAP₅₀ by 5.8 percentage points over YOLO11n

on the VisDrone dataset and achieves consistent performance gains on the USOD and RSOD datasets, demonstrating favorable cross-scene generalization. Module comparisons, ablation studies, and feature visualizations further verify the effectiveness and complementarity of the proposed components. Overall, A²-Det provides an effective accuracy-oriented solution for detecting dense tiny objects in complex remote sensing scenes.

Author Contributions: Conceptualization, S.-J.F. and F.-L.W.; methodology, S.-J.F. and Z.-X.H.; software, S.-J.F., X.G. and A.H.; validation, X.G. and A.H.; formal analysis, S.-J.F.; data curation, Z.-X.H.; writing—original draft preparation, S.-J.F.; writing—review and editing, X.-N.J. and F.-L.W.; visualization, S.-J.F.; supervision, X.-N.J. and F.-L.W.; funding acquisition, X.-N.J. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Key R&D Program of China under Grant 2022YFB3902300.

Data Availability Statement: The datasets analyzed in this study are publicly available from their original sources, including VisDrone2019, USOD, and RSOD. The corresponding data sources and references are provided in Section 3.1. No new datasets were generated in this study.

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

References

- Cheng, G.; Han, J. A Survey on Object Detection in Optical Remote Sensing Images. *ISPRS J. Photogramm. Remote Sens.* **2016**, *117*, 11–28. [\[CrossRef\]](#)
- Li, K.; Wan, G.; Cheng, G.; Meng, L.; Han, J. Object Detection in Optical Remote Sensing Images: A Survey and a New Benchmark. *ISPRS J. Photogramm. Remote Sens.* **2020**, *159*, 296–307. [\[CrossRef\]](#)
- 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), Salt Lake City, UT, USA, 18–23 June 2018*; IEEE: New York, NY, USA, 2018; pp. 3974–3983. [\[CrossRef\]](#)
- Zhu, P.; Wen, L.; Du, D.; Bian, X.; Fan, H.; Hu, Q.; Ling, H. Detection and Tracking Meet Drones Challenge. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 7380–7399. [\[CrossRef\]](#) [\[PubMed\]](#)
- Xu, C.; Wang, J.; Yang, W.; Yu, H.; Yu, L.; Xia, G.S. Detecting Tiny Objects in Aerial Images: A Normalized Wasserstein Distance and a New Benchmark. *ISPRS J. Photogramm. Remote Sens.* **2022**, *190*, 79–93. [\[CrossRef\]](#)
- Wang, J.; Ma, M.; Huang, P.; Mei, S.; Zhang, L.; Wang, H. Remote Sensing Small Object Detection Based on Multicontextual Information Aggregation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 8248–8260. [\[CrossRef\]](#)
- Bu, Y.; Ye, H.; Tie, Z.; Chen, Y.; Zhang, D. OD-YOLO: Robust Small Object Detection Model in Remote Sensing Image with a Novel Multi-Scale Feature Fusion. *Sensors* **2024**, *24*, 3596. [\[CrossRef\]](#) [\[PubMed\]](#)
- Zhou, S.; Zhou, H.; Qian, L. A Multi-Scale Small Object Detection Algorithm SMA-YOLO for UAV Remote Sensing Images. *Sci. Rep.* **2025**, *15*, 9255. [\[CrossRef\]](#) [\[PubMed\]](#)
- Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017*; IEEE: New York, NY, USA, 2017; pp. 2117–2125. [\[CrossRef\]](#)
- Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018*; IEEE: New York, NY, USA, 2018; pp. 8759–8768. [\[CrossRef\]](#)
- Chen, Y.; Liu, B.; Yuan, L. PR-Deformable DETR: DETR for Remote Sensing Object Detection. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 2506105. [\[CrossRef\]](#)
- Lin, S.; Zhang, M.; Cheng, X.; Shi, L.; Gamba, P.; Wang, H. Dynamic Low-Rank and Sparse Priors Constrained Deep Autoencoders for Hyperspectral Anomaly Detection. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 2500518. [\[CrossRef\]](#)
- Gu, Q.; Han, Z.; Kong, S.; Huang, H.; Li, Y.; Fan, Q.; Wu, R. DCYOLO: Dual Negative Weighting Label Assignment and Cross-Layer Decouple Head for YOLO in Remote Sensing Images. *Expert Syst. Appl.* **2025**, *281*, 127595. [\[CrossRef\]](#)
- Dong, C.; Jiang, X.; Hu, Y.; Du, Y.; Pan, L. EL-Net: An Efficient and Lightweight Optimized Network for Object Detection in Remote Sensing Images. *Expert Syst. Appl.* **2024**, *255*, 124661. [\[CrossRef\]](#)
- Pan, J.; Liu, Y.; Fu, Y.; Ma, M.; Li, J.; Paudel, D.P.; Van Gool, L.; Huang, X. Locate Anything on Earth: Advancing Open-Vocabulary Object Detection for Remote Sensing Community. In *Proceedings of the AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA, 25 February–4 March 2025*; Volume 39, pp. 6281–6289. [\[CrossRef\]](#)

16. Lin, S.; Zhang, M.; Cheng, X.; Zhou, K.; Zhao, S.; Wang, H. Hyperspectral Anomaly Detection via Sparse Representation and Collaborative Representation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 946–961. [[CrossRef](#)]
17. Lin, S.; Cai, E.; Zeng, Y.; Zhu, Y.; Zhao, W.; Zhang, L.; Cao, B. A Hyperspectral Target Detection Method by Combining Robust Dictionaries and Constrained Energy Minimization Regularized Low-Rank and Sparse Representation Model. *Adv. Space Res.* **2026**, *78*, 3523–3544. [[CrossRef](#)]
18. Li, Y.; Huang, Q.; Pei, X.; Chen, Y.; Jiao, L.; Shang, R. Cross-Layer Attention Network for Small Object Detection in Remote Sensing Imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 2148–2161. [[CrossRef](#)]
19. Li, Y.; Li, Q.; Pan, J.; Zhou, Y.; Zhu, H.; Wei, H.; Liu, C. SOD-YOLO: Small-Object-Detection Algorithm Based on Improved YOLOv8 for UAV Images. *Remote Sens.* **2024**, *16*, 3057. [[CrossRef](#)]
20. Tang, S.; Zhang, S.; Fang, Y. HIC-YOLOv5: Improved YOLOv5 for Small Object Detection. In *Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA), Yokohama, Japan, 13–17 May 2024*; IEEE: New York, NY, USA, 2024; pp. 6614–6619. [[CrossRef](#)]
21. Zhang, Y.; Ye, M.; Zhu, G.; Liu, Y.; Guo, P.; Yan, J. FFCA-YOLO for Small Object Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5611215. [[CrossRef](#)]
22. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018*; IEEE: New York, NY, USA, 2018; pp. 7132–7141. [[CrossRef](#)]
23. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In *Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018*; pp. 3–19. [[CrossRef](#)]
24. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 19–25 June 2021*; IEEE: New York, NY, USA, 2021; pp. 13713–13722.
25. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In *Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017*; Volume 30.
26. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. In *Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 3–7 May 2021*.
27. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Virtual, 11–17 October 2021*; IEEE: New York, NY, USA, 2021; pp. 10012–10022.
28. Zheng, X.; Qiu, Y.; Zhang, G.; Lei, T.; Jiang, P. ESL-YOLO: Small Object Detection with Effective Feature Enhancement and Spatial-Context-Guided Fusion Network for Remote Sensing. *Remote Sens.* **2024**, *16*, 4374. [[CrossRef](#)]
29. Zhu, X.; Hu, H.; Lin, S.; Dai, J. Deformable ConvNets v2: More Deformable, Better Results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019*; IEEE: New York, NY, USA, 2019; pp. 9308–9316. [[CrossRef](#)]
30. Wang, W.; Dai, J.; Chen, Z.; Huang, Z.; Li, Z.; Zhu, X.; Hu, X.; Lu, T.; Lu, L.; Li, H.; et al. InternImage: Exploring Large-Scale Vision Foundation Models with Deformable Convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023*; IEEE: New York, NY, USA, 2023; pp. 14408–14419.
31. Li, Y.; Hou, Q.; Zheng, Z.; Cheng, M.M.; Yang, J.; Li, X. Large Selective Kernel Network for Remote Sensing Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023*; IEEE: New York, NY, USA, 2023; pp. 16794–16805.
32. Ding, X.; Guo, Y.; Ding, G.; Han, J. ACNet: Strengthening the Kernel Skeletons for Powerful CNN via Asymmetric Convolution Blocks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019*; IEEE: New York, NY, USA, 2019; pp. 1911–1920. [[CrossRef](#)]
33. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. YOLOX: Exceeding YOLO Series in 2021. *arXiv* **2021**, arXiv:2107.08430.
34. Feng, C.; Zhong, Y.; Gao, Y.; Scott, M.R.; Huang, W. TOOD: Task-Aligned One-Stage Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Virtual, 11–17 October 2021*; IEEE: New York, NY, USA, 2021; pp. 3490–3499.
35. Dai, X.; Chen, Y.; Xiao, B.; Chen, D.; Liu, M.; Yuan, L.; Zhang, L. Dynamic Head: Unifying Object Detection Heads with Attentions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021*; IEEE: New York, NY, USA, 2021; pp. 7369–7378. [[CrossRef](#)]
36. Zhuang, J.; Qin, Z.; Yu, H.; Chen, X. Task-Specific Context Decoupling for Object Detection. *arXiv* **2023**, arXiv:2303.01047.
37. Xie, H.; Wang, M.; Cao, R.; Wang, J.; Jiang, Y.; Huang, Q.; Jiang, L. MTD-YOLO: A Multi-Scale Perception Framework with Task Decoupling and Dynamic Alignment for UAV Small Object Detection. *Remote Sens.* **2025**, *17*, 3823. [[CrossRef](#)]
38. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. YOLOv10: Real-Time End-to-End Object Detection. In *Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 10–15 December 2024*; Volume 37.
39. Jocher, G.; Qiu, J. Ultralytics YOLO11. Available online: <https://github.com/ultralytics/ultralytics> (accessed on 28 August 2026).

40. Ni, J.; Zhu, S.; Tang, G.; Ke, C.; Wang, T. A Small-Object Detection Model Based on Improved YOLOv8s for UAV Image Scenarios. *Remote Sens.* **2024**, *16*, 2465. [\[CrossRef\]](#)
41. Yu, F.; Koltun, V. Multi-Scale Context Aggregation by Dilated Convolutions. In Proceedings of the International Conference on Learning Representations (ICLR), San Juan, Puerto Rico, 2–4 May 2016.
42. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016*; IEEE: New York, NY, USA, 2016; pp. 770–778. [\[CrossRef\]](#)
43. Long, Y.; Gong, Y.; Xiao, Z.; Liu, Q. Accurate Object Localization in Remote Sensing Images Based on Convolutional Neural Networks. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 2486–2498. [\[CrossRef\]](#)
44. Xiao, Z.; Liu, Q.; Tang, G.; Zhai, X. Elliptic Fourier Transformation-Based Histograms of Oriented Gradients for Rotationally Invariant Object Detection in Remote-Sensing Images. *Int. J. Remote Sens.* **2015**, *36*, 618–644. [\[CrossRef\]](#)
45. Lin, T.Y.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In Proceedings of the European Conference on Computer Vision (ECCV), Zurich, Switzerland, 6–12 September 2014; pp. 740–755. [\[CrossRef\]](#)
46. Zhu, X.; Lyu, S.; Wang, X.; Zhao, Q. TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-Captured Scenarios. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Virtual, 11–17 October 2021*; IEEE: New York, NY, USA, 2021; pp. 2778–2788.
47. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving into High Quality Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018*; IEEE: New York, NY, USA, 2018; pp. 6154–6162. [\[CrossRef\]](#)
48. Zhang, S.; Chi, C.; Yao, Y.; Lei, Z.; Li, S.Z. Bridging the Gap Between Anchor-Based and Anchor-Free Detection via Adaptive Training Sample Selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 14–19 June 2020*; IEEE: New York, NY, USA, 2020; pp. 9759–9768.
49. Li, X.; Wang, W.; Wu, L.; Chen, S.; Hu, X.; Li, J.; Tang, J.; Yang, J. Generalized Focal Loss: Learning Qualified and Distributed Bounding Boxes for Dense Object Detection. In Proceedings of the Advances in Neural Information Processing Systems, Virtual, 6–12 December 2020; Volume 33, pp. 21002–21012.
50. Qiao, S.; Chen, L.C.; Yuille, A. DetectoRS: Detecting Objects With Recursive Feature Pyramid and Switchable Atrous Convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 19–25 June 2021*; IEEE: New York, NY, USA, 2021; pp. 10213–10224. [\[CrossRef\]](#)
51. Ouyang, D.; He, S.; Zhang, G.; Luo, M.; Guo, H.; Zhan, J.; Huang, Z. Efficient Multi-Scale Attention Module with Cross-Spatial Learning. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes, Greece, 4–10 June 2023*; IEEE: New York, NY, USA, 2023; pp. 1–5. [\[CrossRef\]](#)
52. 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 (CVPR), Virtual, 19–25 June 2021*; IEEE: New York, NY, USA, 2021; pp. 13733–13742.
53. Hendrycks, D.; Dietterich, T. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In Proceedings of the International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019.
54. Rakwatin, P.; Takeuchi, W.; Yasuoka, Y. Stripe Noise Reduction in MODIS Data by Combining Histogram Matching with Facet Filter. *IEEE Trans. Geosci. Remote Sens.* **2007**, *45*, 1844–1856. [\[CrossRef\]](#)
55. He, K.; Sun, J.; Tang, X. Single Image Haze Removal Using Dark Channel Prior. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *33*, 2341–2353. [\[CrossRef\]](#) [\[PubMed\]](#)
56. Dong, Q.; Han, T.; Wu, G.; Sun, L.; Lu, Y. Robust Object Detection for UAVs in Foggy Environments with Spatial-Edge Fusion and Dynamic Task Alignment. *Remote Sens.* **2026**, *18*, 169. [\[CrossRef\]](#)

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.